NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENTO DE ENGENHARIA ELETROTÉCNICA E DE COMPUTADORES

GUILHERME ELÓI DE CASTRO PACHECO

Licenciado em Ciências da Engenharia Eletrotécnica e de Computadores

ESTUDO DO IMPACTO DO FINANCIAMENTO DA INOVAÇÃO ATRAVÉS DE CONTRATOS PÚBLICOS EM PORTUGAL

DISSERTAÇÃO PARA A OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELETROTÉCNICA E DE COMPUTADORES

MESTRADO EM ENGENHARIA ELETROTÉCNICA E DE COMPUTADORES

Universidade NOVA de Lisboa
Setembro, 2023

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENTO DE ENGENHARIA ELETROTÉCNICA E DE COMPUTADORES

ESTUDO DO IMPACTO DO FINANCIAMENTO DA INOVAÇÃO ATRAVÉS DE CONTRATOS PÚBLICOS EM PORTUGAL

DISSERTAÇÃO PARA A OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELETROTÉCNICA E DE COMPUTADORES

GUILHERME ELÓI DE CASTRO PACHECO

Licenciado em Ciências da Engenharia
Eletrotécnica e de Computadores

Orientador: José Manuel Fonseca
Professor Associado com Agregação, NOVA School of Science and Technology

Coorientadora: Fernanda Llussá
Professora de Empreendedorismo, Gestão e Economia, NOVA School of Science and Technology

Júri:

Presidente: Anabela Monteiro Gonçalves Pronto
Professora Auxiliar, NOVA School of Science and Technology

Arguente: Filipe de Carvalho Moutinho
Professor Auxiliar, NOVA School of Science and Technology

Orientador: José Manuel Fonseca
Professor Associado com Agregação, NOVA School of Science and Technology

MESTRADO EM ENGENHARIA ELETROTÉCNICA E DE COMPUTADORES

Universidade NOVA de Lisboa
Setembro, 2023

Estudo do impacto do financiamento da inovação através de contratos públicos em Portugal

A todos aqueles que tornaram possível ter entregue esta dissertação.

Agradecimentos

Em primeiro lugar tenho de fazer um agradecimento especial ao professor José Manuel Fonseca e à minha família, porque se entrego esta dissertação é em grande parte devido a eles. Ao professor José porque nunca desistiu de mim, mesmo quando tinha razões para isso e o poderia ter feito, a ele o meu muito obrigado. À minha família porque sem eles nunca teria conseguido superar seis anos de mestrado integrado, obrigado pela paciência, por todo o suporte familiar, por me aturarem e por me darem força para continuar quando eu mais precisava, alento ao longo de todo o percurso e sobretudo pelas chamadas de atenção quando me desviava do caminho. Aos meus pais, Rui e Raquel, e à minha irmã Madalena, o meu muito obrigado.

Em segundo lugar, tenho naturalmente de agradecer à professora Fernanda Llussá, que entrou a meio de todo este processo e que se prontificou a ajudar na parte económica da dissertação com todo o entusiasmo desde o primeiro momento.

Não me posso também esquecer dos meus amigos de Sintra, que me ajudaram a navegar todos estes seis anos na vida, perguntando pelo estado da dissertação, mas, principalmente, sendo uma segunda família ao longo destes anos. A vocês Tomás, Santos, Faria e Mihail, o meu muito obrigado.

Tenho ainda de referir os meus amigos da faculdade, o Palma, o Gonçalo, a Sara, a Minal e o João, que espero levar para a vida e que me ajudaram a navegar nestes seis anos de faculdade até chegar a esta dissertação e, que mais que isso, foram verdadeiros amigos em todos os outros aspetos pessoais que influenciam bastante a motivação e a força necessária para levar até ao fim este percurso. O meu muito obrigado amigos.

Já os mencionei, mas vou ter de mencionar novamente, o Mihail e a Sara, por razões diferentes. O Mihail por ter sido a pessoa que mais esteve ao meu lado em todo este percurso. Se estou a entregar esta dissertação e a acabar este curso é devido a ele, ninguém foi mais porto de abrigo e suporte nesta caminhada que ele. Companheiro, amigo e praticamente irmão, um obrigado muito especial para ti Mihail. A Sara foi muito importante em dois momentos em que estive perto de desistir de entregar esta dissertação e foi em grande parte devido a ela não me ter deixado desistir nesses momentos e me ter dado estratégias e outros pontos de vista que decidi continuar, pelo que tinha de lhe fazer também este

agradecimento especial. Obrigado Sara.

Por último, não posso deixar de mencionar o professor Francesco, que foi quem me apresentou as temáticas desta dissertação pela primeira vez.

Resumo

São muitos os governos nacionais, locais e supranacionais que utilizam medidas do lado da procura para tentar influenciar positivamente a economia dos seus países. Uma medida muito utilizada neste âmbito é a contratação pública, tendo vários países e organizações supranacionais definido estratégias para otimizar a implementação destas contratações. Nesse sentido torna-se de extrema relevância o estudo do impacto efetivo desse investimento no país em que é utilizado esse recurso financeiro. Esta dissertação irá focar-se nos contratos públicos celebrados em Portugal, para estudar o seu impacto efetivo na inovação do país.

Nesta dissertação, apresentam-se algoritmos e metodologias técnicas para executar a extração, armazenamento, limpeza, análise e visualização de dados necessárias para a realização dos estudos económicos fundamentais para compreender o impacto efetivo dos contratos públicos de inovação em Portugal.

Para isso, primeiro foram extraídos os dados de várias fontes através de técnicas de *web scraping*, armazenaram-se e estruturaram-se esses dados em bases de dados relacionais e depois limparam-se e visualizaram-se os mesmos, para daí se extraírem análises, hipóteses, conclusões e questões relacionadas com o estudo do potencial para a inovação em Portugal deste mecanismo de investimento público. Tais como a relação entre financiamento e obtenção de contratos públicos de inovação, uma baixa correlação entre financiamento e dimensão das empresas e reflexões sobre possíveis otimizações na gestão de financiamento público tendo em vista a aumentar a inovação medida em termos de número de patentes registadas pelas empresas financiadas por este meio, que podem servir de suporte para estudos futuros neste âmbito.

Palavras-chave: web scraping, análise de dados, contratação pública, inovação

Abstract

There are many national, local, and supranational governments that use demand-side measures to positively influence their countries' economies. A commonly employed measure in this context is public procurement, with several countries and supranational organizations defining strategies to optimize the implementation of these contracts. In this regard, studying the effective impact of this investment in the country where this financial resource is utilized becomes extremely relevant. This dissertation will focus on public contracts established in Portugal, aiming to study their effective impact on the country's innovation.

This dissertation presents algorithms and technical methodologies for performing data extraction, storage, cleaning, analysis, and visualization required for conducting the necessary economic studies to comprehend the actual impact of innovation-focused public contracts within contracts established in Portugal.

To achieve this, data was first extracted from various sources using web scraping techniques, then stored and structured into relational databases. Subsequently, the data was cleaned and visualized to extract analyses, hypotheses, conclusions, and questions related to studying the potential for innovation in Portugal through this mechanism of public investment. These include the relationship between funding and obtaining public innovation contracts, a low correlation between funding and the size of companies, and reflections on possible optimizations in public funding management aimed at increasing innovation measured in terms of the number of patents registered by companies financed through this means, which can serve as a foundation for future studies in this field.

Keywords: data analysis, innovation, public procurement, web scraping

Índice

Índice de Figuras

Índice de Tabelas

Siglas

ACID Atomicidade, Consistência, Isolamento e Durabilidade 26, 27
API Application Program Interface 21

BCE Banco Central Europeu 2, 5

CAE Classificação das Atividades Económicas Portuguesas xi, xii, xiii, 32, 33, 35, 37, 39, 40, 45, 48, 50, 51, 52, 53, 54, 55, 57, 59, 60, 62, 63, 67, 68, 69
CSS Cascading Style Sheets 12, 16, 29
CSV Comma-separated values 22, 36, 37, 38, 39, 41

DBMS Database Management System 25
DOM Document Object Model 36

ER Entity-Relationship xi, 29

HTML Hyper Text Markup Language 12, 13, 14, 15, 16, 18, 19, 29, 34
HTML DOM Hyper Text Markup Language Document Object Model xi, 12, 15, 17, 18
HTTP Hypertext Transfer Protocol 12, 16, 18, 23

I&D Investigação e Desenvolvimento 8, 33
INE Instituto Nacional de Estatística 51
INPI Instituto Nacional da Propriedade Industrial 3, 32, 34, 36, 37

JS JavaScript 15, 16, 17, 29, 34

NIF Número de Identificação Fiscal 33, 35, 36, 37, 38, 40, 41

OCDE Organização para a Cooperação e Desenvolvimento Económico 1, 2

PIB Produto Interno Bruto 1, 9
PME Pequena e Média Empresa 41

SHDM Semantic Hypermedia Design Method 30
SQL Structured Query Language 22, 26, 37

UE União Europeia 1
UML Unified Modeling Language 30
URL Uniform Resource Locator 34, 35, 37, 41

VAB Valor Acrescentado Bruto 50, 51, 52, 67, 68

XML Extensible Markup Language 14, 22, 27

1

Introdução

1.1 Motivação e objetivos

Vários governos, sejam eles nacionais, locais ou supranacionais, adotam medidas de estímulo à procura como forma de influenciar positivamente a economia dos seus países. Estas medidas têm como objetivo impactar de maneira favorável a produção, o emprego e a inflação. Uma prática comum neste contexto é a contratação pública, sendo que diversos países e organizações supranacionais desenvolvem estratégias específicas para otimizar a implementação destas contratações.

Olhando para os valores monetários despendidos em contratação pública, por exemplo, pela União Europeia (UE) e pela Organização para a Cooperação e Desenvolvimento Económico (OCDE) podem extrair-se conclusões sobre a sua relevância. Segundo o site da Comissão Europeia [1], por ano, por volta de 250 000 autoridades públicas na UE gastam 14% do seu Produto Interno Bruto (PIB) em contratação pública, ou seja, por volta de 2 biliões de euros por ano. Segundo o site da OCDE [2], o valor gasto pelos países pertencentes a esta organização representa 12% do seu PIB.

No mesmo site da Comissão Europeia citado no parágrafo anterior [1], é referido que pequenas melhorias na forma como se efetua esta contratação pública podem trazer grandes poupanças financeiras. É dado um exemplo em que se estima que 1% de melhoramento na eficiência desses processos tragam poupanças na ordem dos 20 mil milhões de euros por ano. Para fazer estas melhorias, apresenta-se como um dos grandes pilares iniciais a análise dos dados já existentes para entender como está a ser gasto o dinheiro e que formas têm demonstrado melhores resultados e eficiência.

Nesta dissertação vamos focar-nos no panorama nacional e na temática da contratação pública de inovação. À partida temos alguns problemas com que qualquer análise tem de lidar. Em primeiro lugar, como definir o que é ou não contratação pública de inovação. Para esta dissertação será considerada a definição de contratação pública de inovação recomendada pela Comissão Europeia [3], que considera uma compra de inovação a compra do processo de inovação, em que se contratam os serviços de inovação e desenvolvimento em si, sem um produto final previamente estabelecido e com

resultados parciais, ou seja, apenas com provas teóricas da viabilidade do produto, mas sem provas dadas no mercado. Além desta, também é considerada compra de inovação a compra de um produto proveniente de um processo de inovação, ou seja, a compra dos resultados de inovação de agentes exteriores aos agentes que contratam, aqui podendo ser um produto que ainda não está no mercado ou que já esteja, mas há relativamente pouco tempo. A segunda problemática é, por vezes, a falta de dados para análise. Segundo a OCDE [4], 81% dos países pertencentes a esta organização desenvolveram estratégias e políticas de suporte à inovação de bens e serviços através da contratação pública. Contudo, apenas 39,4% estão a medir os resultados destas medidas. Isto quer dizer que há uma grande necessidade de pensar estas estratégias tendo já em mente a obtenção de dados ao longo do processo de implementação das mesmas para posterior análise e atuação.

É importante também discutir a importância da inovação para a economia de um país. O Banco Central Europeu (BCE) descreve a inovação como "um impulsionador do progresso económico, beneficiando consumidores, negócios e a economia no seu todo" [5]. No mesmo artigo a afirmação anterior é justificada tendo em conta a ligação entre inovação e aumento de produtividade, o que leva a que com os mesmos *inputs* se gere um maior número de *outputs*. Ainda no mesmo artigo, são referidos os benefícios deste aumento de produtividade para o aumento de salários e, consequentemente, para criação de novos empregos, dado que da primeira relação resulta um aumento de poder de compra dos consumidores, pelo que estes poderão adquirir mais bens e serviços, fazendo com que os negócios se tornem mais rentáveis e, assim sendo, possam investir e contratar mais colaboradores.

Abordadas as limitações, necessidades e potencial da contratação pública de inovação, é importante clarificar porque é que esta temática é relevante para esta dissertação e o que se pretende com a mesma. O objetivo desta dissertação será estudar se os contratos públicos como forma de financiamento da inovação resultam efetivamente num aumento objetivo de inovação. É relevante fazer este estudo, pois há evidências de que o aumento do mercado potencial conduza a um aumento da inovação num determinado setor, por exemplo, os estudos realizados em [6] concluem que, na indústria farmacêutica, um aumento de 1% do mercado potencial de um determinado medicamento possa levar a um aumento de 4 a 6% do número de medicamentos com o mesmo propósito do primeiro, ou seja, um aumento do mercado potencial desse medicamento, leva a que seja efetuada mais investigação e inovação para criar novos medicamentos para colmatar esse aumento do tamanho de mercado. Para que o objetivo principal desta dissertação seja cumprido é necessário estudar em detalhe o perfil das empresas que receberam financiamento público referente a inovação, nomeadamente, em termos de setor de atividade em que operam, receita bruta e número de colaboradores. Para completar este estudo, fez-se recurso do número de patentes registadas por estas empresas, apesar das limitações deste tipo de análise.

1.2 Introdução às metodologias utilizadas

Para cumprir os objetivos desta dissertação serão utilizadas técnicas de recolha, limpeza, análise, manipulação e visualização de dados.

Em Portugal, o site base.gov centraliza a informação sobre os contratos públicos celebrados em Portugal, pelo que essa é a principal fonte de informação utilizada nesta dissertação. Foram ainda recolhidos dados de outras fontes como o racius.com para obtenção de dados sobre o setor de atividade em que operam as empresas presentes nos contratos públicos identificados no base.gov. Recorreu-se ainda a outras duas fontes de dados. À base de dados da Orbis para recolher informações sobre a receita bruta das empresas e o número de colaboradores das mesmas e à base de dados de patentes registadas por entidades portuguesas do Instituto Nacional da Propriedade Industrial (INPI). Tudo isto a ser guardado em base de dados, para posterior limpeza, análise, manipulação e visualização de dados e usando como recurso várias bibliotecas de *Python* das quais se falarão mais adiante nesta dissertação. No final, para a tarefa de visualização é utilizado o *Google Colab*. De forma resumida, são apresentados em baixo os métodos utilizados nesta dissertação:

- Recolha de dados do site base.gov através de técnicas de *web scraping*, para criação de base de dados de contratos
- Recolha de dados referentes ao setor de atividade das empresas presentes nos contratos públicos recolhidos no racius.com, para criação de base de dados de empresas
- Recolha de dados de receita bruta e número de colaboradores, das empresas presentes na base de dados de empresas, na base de dados da Orbis, para complementação da informação presente na base de dados de empresas
- Recolha de dados relativos a patentes registadas pelas empresas presentes na base de dados de empresas, na base de dados do INPI, para complementação da informação presente na base de dados de empresas
- Tratamento dos dados e visualização dos mesmos no *Google Colab* tirando recurso de várias bibliotecas de *Python*

1.3 Organização do documento

Esta dissertação está organizada em quatro capítulos principais, divididos em sub-capítulos com as temáticas principais referentes a cada um. Os conteúdos presentes em cada capítulo são os seguintes:

- **Capítulo 1:** neste capítulo inaugural são introduzidos e definidos alguns conceitos importantes para entender a motivação desta dissertação e a sua relevância. Além

desse sub-capítulo, existem mais dois onde são sumarizados os objetivos e trabalhos desta dissertação e outro onde é feito o mesmo trabalho para o conteúdo de cada capítulo da dissertação.

- **Capítulo 2:** nesta secção é realizado o estado de arte, tendo dois grandes focos, um tecnológico e outro económico. A parte tecnológica debruça-se sobre os temas do *web scraping* e *base de dados*, de especial importância para esta dissertação e ainda uma pequena nota sobre desenvolvimento de aplicações para a *web*. Na parte económica, são abordados temas como o que é a inovação, os seus benefícios, como a medir e a importância de financiar a inovação através de contratos públicos. No final, existe um sub-capítulo sumário de todo o conteúdo presente neste capítulo.

- **Capítulo 3:** nesta terceira parte da dissertação são apresentadas as fontes de dados utilizadas, algoritmos desenvolvidos e as formas como os dados foram armazenados, estruturados e limpos. Terminando com a sumarização deste capítulo, tal como foi feito nos anteriores.

- **Capítulo 4:** no quarto capítulo da dissertação são apresentadas as análises empíricas realizadas. Começa-se por descrever os dados e metodologias utilizados e termina-se apresentando os resultados empíricos, com recurso a gráficos e coeficientes de correlação. Como o capítulo anterior, existe ainda um subcapítulo final que sumariza todo a informação do capítulo.

- **Capítulo 5:** para terminar, este capítulo, apresenta um resumo do trabalho da tese, as conclusões retiradas, uma reflexão sobre as mesmas e algumas sugestões para trabalho futuro, no sentido de continuar este estudo.

2

Estado da Arte

Neste capítulo irão ser abordadas todas as temáticas, tecnologias e conceitos económicos necessários para o desenvolvimento da dissertação. Serão apresentados vários conceitos referentes a três temas tecnológicos principais: *web scraping*, bases de dados e desenvolvimento de aplicações *web*. Serão também expostos vários temas relacionados com inovação, a sua importância económica, como fazer a sua medição e a importância do financiamento da inovação através de contratos públicos. Os secções começam sempre com uma breve definição e/ou contextualização do tema, para depois se debruçarem sobre eles em aspetos mais específicos e relevantes para esta dissertação.

2.1 O que é inovação?

Inovação refere-se a um novo processo ou produto que resulta num acréscimo de produtividade, ou seja, com os mesmos ou menos recursos produz-se mais, podendo envolver a criação de algo novo e/ou disruptivo com o mesmo propósito de aumento da produção ou da produtividade.

Vários académicos formularam definições de inovação. Schumpeter [7] define inovação como a formação de novos produtos ou serviços, novos processos, novas matérias-primas, novos mercados e novas organizações. Já Peter Drucker [8] refere-se a inovação como "a função específica do empreendedorismo", seja esta desenvolvida numa empresa já existente ou numa acabada de criar, diz ainda que esta é o "meio pelo qual o empreendedor ou cria novos recursos de produção de riqueza ou dota os recursos existentes de um maior potencial de criação de riqueza".

O BCE [5] descreve inovação como o "desenvolvimento e aplicação de ideias e tecnologias que melhoram bens e serviços ou tornam a sua produção mais eficiente".

Já a McKinsey [9], define inovação num contexto de negócio como "a habilidade de conceber, desenvolver, entregar e escalar novos produtos, serviços, processos e modelos de negócios para os consumidores".

É ainda importante referir a não dependência direta da inovação com a comercialização [10]. Ao contrário do conceito de invenção, que está sempre

diretamente ligado ao de comercialização, uma inovação pode apenas ser o melhoramento de um processo ou algum aspeto de uma empresa, não tem de estar diretamente ligado à criação de um produto ou de um serviço, pode, por exemplo, ser apenas uma inovação na forma do produto ser desenvolvido ou na forma de trabalhar da equipa que está encarregue de desenvolver certo produto. É importante que quando se fala em inovação, o foco esteja em gerar algum tipo de valor que aumente a produtividade da empresa e não apenas na criação de novos produtos e/ou serviços.

É também de referir a importância dos conceitos de novidade e de diferença significativa em várias definições de inovação [11]. Para estas, não existe a necessidade de a inovação ser completamente nova, pode ser só a adoção de uma ideia existente por parte de uma empresa que nunca a aplicou desde a sua fundação, por exemplo, havendo por isso uma diferença significativa, o facto de estar a ser utilizada numa empresa e/ou mercado totalmente diferente. Contudo, não quer isto dizer que se relega a novidade de um produto ou serviço da inovação, antes pelo contrário, tudo o que seja percecionado como totalmente novo, não só continua a estar no âmbito destas definições, como esta associação entre inovação e novidade é um pensamento que está mais presente na cabeça das pessoas.

É ainda importante referir algumas formas de inovação [11], sendo estas, por exemplo: processos, serviços, produtos, comunicação, políticas, sistemas, estratégias e conceitos. Como se pode observar nesta extensa e diversa lista, a inovação pode ser encontrada nas mais diversas formas, se mantiver o foco da disrupção de algum fator tendo em vista a melhoria para um aumento de produtividade.

Na próxima subsecção este estudo aborda os benefícios económicos da inovação.

2.2 Benefícios económicos da inovação

Os benefícios económicos da inovação do ponto de vista macroeconómico prendem-se com a ligação entre inovação, crescimento económico e aumento da produtividade e de um ponto de vista microeconómico envolvem o aumento dos salários, do lucro das empresas e criação de novos postos de trabalho. Num estudo conduzido pela McKinsey [9] concluiu-se que empresas com um bom nível de domínio da inovação podem gerar 2,4 vezes mais lucro económico do que os seus concorrentes com níveis de domínio da inovação inferiores.

Outro benefício bastante interessante da inovação, prende-se com o acréscimo de colaboração entre empresas [12]. Dado que a inovação acontece a partir de uma disrupção, ou seja, da ideação e aplicação de uma ideia nova ou de uma ideia antiga num contexto novo, esta ganha com a partilha de ideias e colaboração entre trabalhadores e empresas. Em tempos tão incertos e onde se perceciona cada vez menos segurança, estas colaborações entre empresas podem ser uma arma interessante para enfrentar estas dificuldades, diminuindo o risco para cada empresa, dado que este é partilhado por cada uma das mesmas. Logo, a partir da colaboração, se feita com qualidade, aumenta-se a

probabilidade de alcançar uma ideia disruptiva, dado que por norma a diversidade de ideias costuma enriquecer o resultado final, aumentando assim o valor acrescentado do mesmo, aumenta-se a robustez da equipa a novos desafios que surjam dado que, por norma, uma equipa multidisciplinar ou com várias valências está mais bem preparada para responder a desafios originados pelo clima de maior insegurança e incerteza em que vivemos atualmente e, além disto, também reduz o risco para cada empresa individualmente, o que em termos económicos se apresentam como fatores muito interessantes, geração de mais valor, logo mais ganhos e/ou lucro, com menos perdas e/ou despesas.

Outro benefício prende-se com a relação entre inovação e competitividade [12]. A competitividade entre empresas é um aspeto essencial para o funcionamento dos mercados. Um bom nível de competitividade entre empresas representa uma maior aproximação de um mercado saudável, onde surgem empresas novas, desaparecem empresas pouco competitivas e que não traziam tanto valor ao mercado como as restantes empresas presentes nesse mercado. Pelo contrário, um fraco nível de competitividade faz com que empresas pouco produtivas se consigam manter no mercado, estimulando monopólios e não estimulando novas empresas a entrar no mercado, fazendo com que este estagne e não continue a crescer. Assim sendo, a competitividade é uma componente essencial para o crescimento económico, para a saúde de um mercado e para o melhor funcionamento do mesmo (na perspetiva da criação de valor económico e de produtividade), pelo que a inovação se apresenta como um fator essencial para estimular esta competitividade entre empresas.

Para concluir, além de todos os benefícios abordados anteriormente neste capítulo, é também importante referir o impacto económico da inovação ao nível de uma empresa [10]. A inovação dentro de uma empresa prende-se com a adoção ou melhoramento de algum aspeto da mesma, como algum processo, norma interna ou até um investimento em novas tecnologias ou em formação dos colaboradores, ou também pode ser associado à criação de um novo produto ou serviço. Pelo que todos estes exemplos de inovação podem ter um impacto significativo na produtividade da empresa fazendo com que ela seja mais competitiva, consiga crescer e alcançar maiores margens de lucro e possa pagar melhor aos seus funcionários e até contratar novos e, neste cenário, voltamos à relação entre inovação, crescimento económico e criação de novos postos de trabalho, desta vez partindo de uma perspetiva mais micro, apenas de uma empresa, para olhar para a influência desta numa perspetiva mais macro, de um mercado ou da economia no seu todo, corroborando agora a segunda, que já tinha sido analisada neste, capítulo com a primeira.

A próxima subsecção aborda a questão de como quantificar a inovação.

2.3 Como medir inovação?

Existem várias formas de tentar medir inovação, mas não existe uma que seja completamente consensual ou que não tenha nenhuma limitação, o que poderá, de certa

forma, limitar a resposta ao objetivo proposto neste estudo que consiste no estudo do impacto dos financiamentos através de contratos públicos na inovação das empresas.

Shapiro [13] refere a abordagem de associar uma medida fixa com uma variável como medida de inovação, dando o exemplo das medidas receita proveniente de novos produtos e receita proveniente de novas plataformas, ou seja, a primeira foca-se apenas em produtos e a segunda em qualquer tipo de plataforma que origine inovação. Já Katila [14] conclui, num estudo a 100 empresas de ramo biofarmacêutico, que as patentes fornecem uma medida útil para a performance inovadora de uma empresa.

A McKinsey recomenda duas métricas para tentar fazer esta medição, a *R&D-to-product conversion* e a *New-products-to-margin conversion*. A primeira é uma métrica calculada através dos gastos em pesquisa e desenvolvimento em proporção das vendas, neste caso de vendas de novos produtos. A segunda considera o rácio da percentagem da margem bruta das vendas que representam as vendas de novos produtos. O problema destes indicadores é a escassez dos mesmos para empresas portuguesas, pelo que não foi possível utilizá-los para esta dissertação.

Harald Bathelt [12] refere os diferentes tipos de dados que podem ser utilizados para medir inovação. Sendo que destes temos dados de *input* e de *output* e ainda dentro destes são apresentadas subdivisões no que diz respeito à sua relação direta com práticas de Investigação e Desenvolvimento (I&D) e a relação direta com o conceito de inovação. No que diz respeito aos *inputs*, são apresentadas métricas relativas a práticas de I&D, sendo aconselhado a escolha de métricas originárias de práticas intencionais, sistemáticas e que mantenham certos elementos de incerteza, sendo assim são aconselhados indicadores como os gastos em I&D, o número de práticas de I&D dentro da empresa e as práticas de acumulação e organização do conhecimento gerado nestas atividades. Por outro lado (e ainda dentro dos *inputs*), temos ainda indicadores interessantes que não estão diretamente ligados a I&D, mas que indiretamente o afetam e devem ser considerados como o dinheiro gasto em ativos intangíveis, atividades inteligentes, compra de equipamento e/ou serviços de suporte à inovação dentro da empresa e gastos em formação para gerir a informação gerada em atividades de I&D. Do lado dos *outputs*, temos indicadores diretos de inovação, como novos produtos criados, número de inovações em produtos antigos, número de inovações em processos, organização ou marketing da empresa e temos indicadores indiretos como são os casos das patentes registadas, registo de *designs* industriais, marcas, *copyrights* e modelos de utilidade. Analisando os *outputs* diretos e indiretos, existe mais informação disponível sobre os indiretos, contudo existem também muito mais limitações neste tipo de dados, pelo que deve ser feita uma análise ainda mais cuidada a este tipo de dados.

Rogers [10] resume como um bom indicador de *output* para medir a inovação de uma empresa, um indicador que meça o seu sucesso, complementado com a utilização de algumas técnicas econométricas, como, por exemplo, os seguintes indicadores: lucro, aumento da receita, capitalização de mercado ou produtividade. É essencial complementar estes dados com técnicas econométricas, dado que não é direta a relação

entre sucesso de uma empresa e inovação, pelo que se tem de confirmar a causalidade, presente na correlação que normalmente existe. Além destes indicadores de *output* existem também indicadores como o número de novos ou melhorados produtos, serviços ou processos, para os quais também é aconselhado que se adicione alguns dados mais específicos a estes, como a percentagem de vendas gerada a partir das vendas destes novos ou melhorados produtos, serviços e/ou processos.

Rogers [10] menciona ainda as patentes como forma de medir inovação, sendo que apesar de todas as limitações derivadas da relação indireta entre registo de patentes e inovação e do que pode ser uma medida incompleta à relação, por vezes, não direta entre a formalização de uma patente e a sua utilização no mundo empresarial ou até quantificar o quanto esta foi utilizada. Apesar disto, é feito também um contra-argumento interessante, pois só a atividade em si de registar uma patente já demonstra um comportamento inovador por parte da empresa que a regista, pelo que mesmo não sendo um indicador muito sólido, por si só, em conjunto com mais dados e fazendo recurso de técnicas econométricas, pode revelar-se um indicador bastante interessante.

2.4 A importância do financiamento a partir de contratos públicos para a inovação

Os benefícios económicos da inovação, descritos anteriormente, reforçam a importância do financiamento de forma a estimular e acelerar a taxa de inovação das empresas. Em Portugal, o governo tem financiado a inovação das empresas por meio de contratos públicos.

Os contratos públicos de inovação são uma ferramenta bastante interessante para fomentar a inovação de duas maneiras distintas: fomentar a inovação no setor privado e fomentar a inovação dentro do setor público [11]. No caso do setor privado é bastante evidente que ao contratar os seus serviços para o desenvolvimento de alguma atividade inovadora, se fomenta, contribui e financia a prática da inovação dentro das empresas privadas, dando-lhes os recursos necessários, que, por vezes, escasseiam. No caso do setor público, a contratação do setor privado para modernizar o setor público e fazê-lo assim incorrer em mais e melhores práticas de inovação, torna-o efetivamente mais inovador, através da aprendizagem de boas práticas do setor privado. Como é referido em [11], é de extrema importância dar mais responsabilidade aos gestores do setor público, para que depois possam aplicar este conhecimento, sem grandes restrições hierárquicas ou burocráticas. Dado que o setor público representa em média 20% a 30% do PIB dos países desenvolvidos economicamente [11], torna-se ainda mais importante olhar para este tipo de investimento para potenciar a inovação e assim a produtividade do setor público de um país.

Dois dos fatores que se apresentam como maiores barreiras à inovação para as

empresas são a falta de conhecimento e de financiamento [15]. O investimento em contratos públicos de inovação apresenta-se como uma excelente ferramenta para contribuir para solucionar estas dificuldades, dado que este tipo de contrato pode, quando aplicado na ótica do conhecimento, contratar especialistas para providenciar o conhecimento em falta a quem necessita dele, que neste caso poderá ser o próprio setor público, mas, também, na ótica da falta de financiamento, estes contratos públicos podem ser um bom mecanismo de financiamento das entidades que necessitam deste para desenvolver uma investigação ou concretizar uma inovação já previamente investigada.

Por fim, é importante também referir que a colaboração entre o setor público e privado poderá ser bastante interessante para ambos [16], como estratégia de partilha de recursos, informação e/ou complementação de competências, tendo como objetivo um maior e melhor resultado no que diz respeito a tarefas relacionadas com inovação ou inovadoras em alguma medida.

Os próximos subcapítulos 2.5, 2.6 e 2.7 dizem respeito à revisão da literatura no que diz respeito à metodologia que será usada neste estudo e os dados disponíveis e o subcapítulo 2.8 apresenta um resumo dos principais pontos abordados.

2.5 *Web Scraping*

2.5.1 *Web scraping*: definição, objetivo, limitações, exemplos e relevância

Temos várias formas de definir o que é *web scraping*. A mais simples e consensual que se encontrou, na pesquisa para este capítulo, foi a que foi retirada de [17], onde é definido como "o processo de coleta de dados ou informação de sites da internet". Olhando para esta frase podemos afirmar que um utilizador comum pode realizar este processo "manualmente", através do tradicionalmente chamado "*copy and paste*", utilizando apenas o seu computador e uma ligação à internet. Embora não deixe de ser verdade, que este é um método de *web scraping*, este será o único método que irá ser excluído neste capítulo, por ser o único que não interessa explorar nesta dissertação, dado a sua ineficiência e não necessitar de qualquer tipo de programação adicional. Antes de se referir mais técnicas de *web scraping*, na próxima secção deste capítulo, é importante mencionar os objetivos, desafios, alguns exemplos e porque é que o *web scraping* é tão relevante para esta dissertação.

O principal objetivo deste processo de coleta de dados *web* está muito bem resumido em [18] onde é escrito que este é utilizado para "transformar dados não estruturados da *web* em dados estruturados que possam ser guardados e analisados numa base de dados central e local ou numa folha de cálculo". Esta frase introduz uma nova variável importante, que é a preocupação deste processo não só na recolha de dados, mas, também, na forma como esses dados são recolhidos, já a pensar na forma como serão gravados, mais precisamente, como serão guardados de uma forma estruturada, para ser

mais fácil uma futura análise dos mesmos. Este aspeto é muito relevante, pois é um ponto que deve ser pensado logo no início da programação deste tipo de sistema, para poupar trabalho e otimizar recursos, sejam eles o código em si, como mesmo o tempo do próprio programador mais adiante.

Além da preocupação expressa no parágrafo anterior, temos ainda outros aspetos que são importantes ter em mente na programação deste tipo de sistemas, muitos deles mencionados em [19] com maior detalhe.

O primeiro é a noção de que os *websites* ao longo do tempo são modificados, logo é importante equacionar esta hipótese (bastante provável) quando se pensa nestes sistemas, pois senão corre-se o risco de ter um sistema completamente obsoleto em pouco tempo, dado que, por exemplo, podemos estar a retirar informação de um determinado elemento de um site que por o seu nome ter sido alterado, o nosso código não conseguirá mais retirar os dados pretendidos. Para ultrapassar este problema existem duas boas práticas: tentar usar as expressões mais gerais possíveis, pois quanto mais específicos formos, maior a probabilidade de esse elemento e/ou nome ser alterado no futuro e acautelar inicialmente no nosso código, sistemas de deteção de erros, para que assim o programador seja notificado com a maior brevidade possível que o sistema está a falhar na obtenção de determinada informação e assim possa atualizar essa parte específica do código. Estas duas boas práticas são importantes para tornar estes sistemas mais robustos a médio, longo prazo.

A curto prazo, é importante ter em mente que como estamos a robotizar processos habitualmente efetuados por humanos, que existem sites que têm mecanismos para tentar detetar estes sistemas de *web scraping* e/ou de automação (o que é comummente chamado de *bots*) e que existem certas regras a cumprir, habitualmente, colocadas no ficheiro robots.txt do *website*. Dito isto, é importante não só ter este cuidado na programação, por exemplo, aumentando o tempo entre coletas de dados diferentes, pois há certos intervalos de tempo que não são humanos e é importante, novamente, pensar no futuro. Os sistemas de deteção de automações que mimetizam o comportamento de um utilizador, também, evoluem, pelo que é uma boa prática, pensar no nosso código as zonas onde existe possibilidade de ele falhar no futuro devido a esta problemática, para que caso ele falhe possamos notificar o utilizador e, posteriormente, o programador para atualização do código e resolução do problema.

Um dos exemplos mais conhecidos e usados como iniciação ao *web scraping*, é a recolha de dados meteorológicos. Através deste processo é possível automatizar a recolha de dados de um site com a meteorologia do local que se queira, num dia e hora específico, entre outros filtros e/ou informações que queiramos recolher e guardá-la de forma estruturada, já na forma com que se queira analisá-la ou utilizá-la. Este é apenas um exemplo de uma possível área de utilização, existem outros exemplos, como monitorizar valores financeiros da bolsa ou até monitorizar alterações num site específico.

Para terminar esta secção, é ainda pertinente especificar a importância deste processo para esta dissertação. Esta dissertação terá como base a recolha de dados para posterior

análise, pelo que é de extrema importância todo este processo de coleta de dados *web* e a sua otimização para que seja mais eficiente em termos de recursos e tempo, mas, também, que seja um sistema robusto à passagem do tempo, tanto na programação orientada para a obtenção de menos erros no futuro, como, igualmente, para a deteção rápida de erros, para rápida resolução.

2.5.2 Técnicas de *web scraping*

Existem variadas técnicas de *web scraping*, por exemplo: Expressão Regular (*Regular Expression*), Hyper Text Markup Language Document Object Model (HTML DOM), Xpath, Cascading Style Sheets (CSS) *selector*, *Vertical aggregation*, *Semantic Annotation Recognizing*, *Computer Vision web-page Analysis*.

Para esta dissertação, o foco inicial é a recolha de dados de um site a partir do código fonte que este têm inicialmente ou dos dados que adquire através de pedidos Hypertext Transfer Protocol (HTTP), para que este segundo o interprete e o apresente da forma como depois o utilizador o vê no seu aparelho eletrónico. Com este propósito, vai dar-se foco aos três primeiros exemplos do parágrafo anterior, dado que são as técnicas mais específicas para trabalhar com ficheiros Hyper Text Markup Language (HTML) e com o HTML DOM. Contudo, não se quis deixar de referir as especificidades das outras técnicas mencionadas, pois apresentam características diferentes para objetivos distintos, que após análise mais à frente, poderão ser interessantes para a dissertação.

Para serem visualmente apelativos e organizados, todos os *websites* necessitam de utilizar ficheiros CSS. Sem a utilização desta linguagem todos os sites são visualmente semelhantes a um ficheiro de texto sem qualquer personalização. Quando este ficheiro faz parte da estrutura do site (que faz muitas vezes), a técnica CSS *selector* pode ser interessante. Contudo, muitas vezes, utilizar técnicas mais específicas para HTML DOM apresenta-se como uma solução mais fácil.

Existem plataformas que "escondem" a programação por detrás desta obtenção de dados através de *web scraping*, facilitando o trabalho ao utilizador, que assim não tem de saber escrever código para obter os dados que pretende recolher através deste processo. A isto chamam-se plataformas de *Vertical aggregation*. Estas plataformas são interessantes, pois permitem ao utilizador modelar dentro desta ferramenta o sistema que quer ter, até chegar aos dados que quer recolher e depois, também, que dados quer recolher, de uma forma mais facilitada e sem programação. Contudo, tem a limitação de apenas ser possível escolher as opções que a plataforma nos fornece, que, por vezes, pode não estar ajustado às necessidades que se pretende satisfazer, pelo que nesse caso é melhor desenvolver mesmo algum código personalizado. Por último, deixar apenas a nota de que pode ser uma boa ferramenta para fazer pequenos testes antes de partir por completo para a programação [18].

Temos outra técnica interessante, chamada *Semantic Annotation Recognizing*. Esta técnica baseia-se na utilização dos elementos textuais do site, do qual queremos retirar

informação, sejam eles metadados ou textos que são apresentados visualmente no *website*, para gerar anotações sobre esse mesmo conteúdo, fazendo uma espécie de classificação das coisas relevantes do site, para depois reconhecer dentro da panóplia de diferentes sites que poderá analisar os que apresentam as anotações desejadas [20]. Esta técnica poderá ser interessante analisar para possíveis abordagens a ter na parte final desta dissertação, contudo não é uma das técnicas focadas neste capítulo, por não ser tão óbvia a sua utilização como a das 3 primeiras, que claramente podem ser utilizadas em todas as fases da dissertação.

Por último, antes de se entrar nas três técnicas de foco deste sub-capítulo, temos a *Computer Vision web-page Analysis*, que utiliza técnicas de aprendizagem de máquina (*machine learning*) e de visão computacional (*computer vision*), para tentar observar um *website* como o olho de um ser humano e assim tentar retirar as conclusões que uma pessoa conseguiria retirar a partir da observação de um site. [18] Tal como a técnica abordada no parágrafo anterior, esta da análise de uma página *web* a partir de processos de visão computacional, são técnicas mais sofisticadas e consequentemente mais complexas, pelo que para esta dissertação só devem ser utilizadas se tivermos um problema complexo que veja nestas ferramentas uma solução ideal, até lá as técnicas que serão abordar a seguir são mais simples e conseguem resolver muitos problemas comuns a esta dissertação.

Neste capítulo, o foco, a partir de agora, será então as três técnicas referidas no primeiro parágrafo. Para começar, é importante, dar uma definição de cada uma delas. A Expressão Regular ou Regex é uma fórmula que obedece a um padrão específico que tem de ser cumprido e que, na prática, define um conjunto de caracteres que queremos observar se existem num determinado conjunto de caracteres depois a comparar com a expressão pretendida. [18] Por exemplo, se quisermos fazer um sistema de validação de endereços de correio eletrónico e a expressão Regex utilizada fosse "@([a-zA-Z0-9_.+-]+)[a-zA-Z0-9_.+-]" e o endereço de correio eletrónico fosse o seguinte conjunto de caracteres "exemplo@gmail.com" (como está representado na figura 2.1), este seria validado, dado que a expressão pede um "@", seguido de vários caracteres, neste caso o "gmail", seguido de um "." e terminando com pelo menos mais um caracter depois do ".", logo este endereço seria validado. Contudo, o Regex além de facilitar esta confirmação de um certa estrutura num conjunto de caracteres, também, retorna uma parte específica da estrutura dentro do conjunto de caracteres fornecido. Olhando novamente para o exemplo dado em cima, o que seria retornado com a expressão e o conjunto de caracteres dados, seria o conjunto de caracteres "gmail". Esta segunda característica do Regex é aquela que nos interessa para esta dissertação, pois poderia ser utilizada para retirar a informação pretendida de um ficheiro HTML do *website* pretendido. Contudo, em termos de otimização para os dados a recolher nesta tese, das três técnicas em foco neste capítulo é a menos otimizada pelo que já se verá a seguir, mas resumidamente deve-se ao facto de o foco da dissertação não ser exatamente a forma como os dados estão estruturados no HTML, mas sim o que está dentro das suas *tags*, pelo que será mais otimizado, no sentido de ser mais direto, usar uma das técnicas que

serão abordadas em seguida.

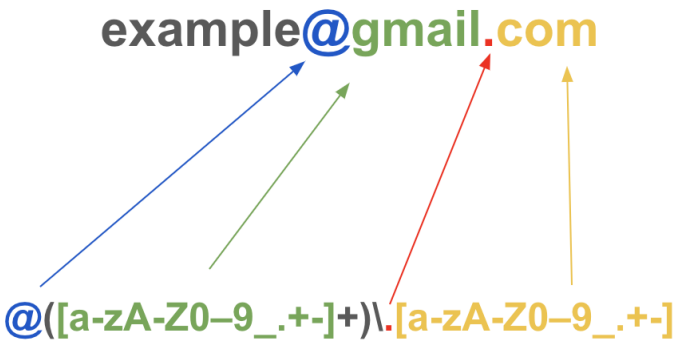


Figura 2.1: Exemplo de uma expressão Regex [21]

Por outro lado, o Xpath é uma técnica utilizada para ficheiros Extensible Markup Language (XML), mas que também pode ser utilizado para ficheiros HTML, que de forma muito simples utiliza a localização de um elemento no ficheiro em comparação com o elemento principal, a que se chama de Xpath [17]. No canto inferior esquerdo da figura 2.2, pode observar-se o Xpath do logótipo da NOVA School of Science and Technology na *homepage* do seu *website*, que não é mais que uma espécie de caminho que começa no caracter "/" e que depois vai referindo hierarquicamente os elementos aos quais pertence o elemento final pretendido, até chegar, neste caso, à imagem pretendida do logótipo da faculdade. Esta técnica é interessante pela sua simplicidade e pela sua fácil utilização num ficheiro HTML, logo consequentemente está mais otimizada para esta dissertação que todas as técnicas referidas até agora. Contudo, tem uma limitação que se prende com a sua característica principal de traçar um caminho desde o *root* até ao elemento pretendido. Isto é interessante por ser simples, mas ao mesmo tempo torna esta expressão pouco resiliente a modificações ao longo do tempo, dado que faz referências muito específicas e as boas práticas que foram abordadas no primeiro sub-capítulo deste capítulo dizem-nos que quanto mais geral melhor neste aspeto.

Figura 2.2: Exemplo de uma expressão Xpath, a partir de inspeção no Google Chrome do site da NOVA School of Science and Technology: [22]

Por último, mas possivelmente a técnica mais relevante para esta dissertação, temos o

HTML DOM. De forma muito simples o que esta técnica faz é tornar todos os elementos de um ficheiro HTML em objetos, o que faz com que seja mais simples "chamar" os elementos pretendidos através de JavaScript (JS) [17]. A grande diferença para a técnica abordada no parágrafo anterior é de que o elemento não depende dos seus "parentes", ou seja, não depende dessa relação hierarquia de estar "dentro" de outros elementos para conseguir alcançar o elemento pretendido, pode sim utilizar-se características específicas do elemento pretendido e chegar a ele de uma forma mais direta, mas não tão simples, contudo a complexidade acrescida não é assim tão limitadora, por isso, considerou-se a técnica mais adequada para esta dissertação. Um exemplo desta técnica pode ser observado na figura 2.3, onde com o objetivo de encontrar a mesma imagem do logótipo, na mesma página *web* da figura 2.2, foi executada a linha de código de JS que se encontra na consola do Google Chrome da primeira figura mencionada, onde se fez uma pesquisa por "id" e se recebeu como retorno um objeto 2.3 referente à imagem pretendida. Este método é mais resiliente à passagem do tempo que o anterior, dado que, caso houvesse alterações no site, as modificações que teriam de ser feitas ao código para encontrar o elemento pretendido, seriam menores que qualquer outro dos métodos mencionados, dado a pesquisa ser mais geral do que as anteriores, pois apenas se referem a uma ou duas características do elemento pretendido e não a várias delimitações sejam elas hierárquicas (no caso do Xpath) ou de estrutura de caracteres (no caso do Regex).

Figura 2.3: Exemplo da execução da técnica HTML DOM, a partir da utilização de JavaScript na consola do inspetor do Google Chrome no site da NOVA School of Science and Technology [22]

Em termos de performance, em [17] foram feitos testes para determinar qual destas três técnicas teria melhor desempenho para executar tarefas de *web scraping*, sendo que foi concluído que o Regex seria a melhor em termos de utilização de memória, mas o HTML DOM seria o mais rápido e é o melhor em termos de utilização de dados, sendo o que utiliza menos dados entre as três técnicas consideradas. Este é um dado interessante para esta dissertação, pois como já tinha referido o HTML DOM parece ser a técnica mais adequada a esta dissertação, sendo a melhor técnica em 2 de 3 fatores segundo a fonte referida no início do parágrafo, também, reforça esta adequação. Além disso, entre estas três técnicas que foram o foco deste capítulo, esta parece ser uma das técnicas que melhor

se adequa à problemática do próximo capítulo que é o *web scraping* em sites dinâmicos.

2.5.3 *Web Scraping* em *websites* dinâmicos

Antes de definir o conceito de *website* dinâmico, é importante em primeiro lugar definir, em traços gerais, o que é e no que consiste um *website*, que não é mais que uma aplicação baseada em tecnologias emphweb, composta, por exemplo, por arquivos HTML, CSS e JS, hospedados em servidores emphweb, e que utiliza protocolos de comunicação, como HTTP, para disponibilizar conteúdo de forma interativa e acessível aos seus utilizadores utilizando como recurso navegadores emphweb. O ficheiro HTML ou outro similar é o ficheiro com o qual o *browser* dará forma ao futuro *website* observado pelo utilizador. Este ficheiro HTML pode ainda conter imagens e ter chamadas a ficheiros CSS ou outros que estilizam o HTML e ficheiros JS ou outros que podem fazer apenas que o site tenha mais movimento ou execuções de ações, que sem este ficheiro não seriam possíveis ou pode mesmo tornar o site dinâmico fazendo chamadas de dados com novos pedidos HTTP ou apenas com a alteração dinâmica de elementos no *website* [23].

Como referido no parágrafo anterior de forma resumida, os sites dinâmicos trazem duas dificuldades acrescidas para o *web scraping*: alterações de elementos do *website* e acréscimo de dados e informação, após o carregamento inicial do *website*. A alteração de elementos é uma dificuldade importante de acautelar, pois pode fazer com que o código desenvolvido deixe de funcionar por completo, dado que quando o código é redigido é necessário sempre partir-se de certas assunções, como por exemplo, se queremos retirar informação de um certo elemento textual com a *tag* HTML <*p*>, contudo num site, o mais comum, é que existam muitas *tags* destas, pelo que é preciso especificar de alguma forma o elemento que queremos em específico, para isto uma das soluções, é retirar a sua classe ou id, contudo nem todos os elementos têm ids e as classes podem ser alteradas nos sites dinâmicos e assim gerar um erro no nosso código. Em relação às dificuldades criadas devido ao acréscimo dinâmico de dados ou informação no site ao longo do tempo prendem-se com os intervalos de tempo em que fazemos a leitura dos dados que pretendemos, por exemplo, se temos uma tabela que apenas é carregada para o *website*, após a inicialização do mesmo, se para o nosso processo de obtenção de dados, recolhermos o ficheiro HTML, proveniente do primeiro pedido HTTP GET, então a tabela não se encontrará lá, o que vai resultar num erro na procura do objetivo que se pretende obter.

Para esta dissertação, utilizamos como fonte principal o *website* base.gov [24]. Depois de observação do site e realização de alguns pequenos testes, podemos concluir que a primeira dificuldade exposta no parágrafo anterior, não se apresenta como dificuldade para a tarefa de *web scraping* desta dissertação. Contudo, a segunda dificuldade exposta é exatamente a dificuldade que se nos apresenta, pelo que quando se fazia a obtenção do ficheiro HTML inicial, a tabela que era possível observar no *browser*, não estava no ficheiro

recolhido, logo não era possível a partir dele retirar a informação recolhida. Isto deve-se ao facto de essa tabela ser gerada por pedidos executados no código em JS, logo não apresentados inicialmente. Para solucionar este problema será então necessário recorrer a técnicas mais sofisticadas de *web scraping*.

Tendo em conta estas duas dificuldades e, em particular, a dificuldade específica desta dissertação, expostas no parágrafo anterior, existem algumas possibilidades para solucionar estes problemas. Em relação ao primeiro de alterações dinâmicas de elementos do *website*, é importante seguir as boas práticas de usar expressões de pesquisa de elementos o mais gerais possíveis, pois quanto mais específicas, maior a possibilidade de alteração futura, ou seja, de tornar o código desenvolvido para esta pesquisa errático, mas, também, ter em atenção este aspeto logo na programação, tanto no sentido de quando se observa pela primeira vez o *website* verificar se apresenta esta característica e, por precaução, mesmo depois deste estudo inicial, colocar no nosso código mecanismos que verificam o correto funcionamento do mesmo, para que assim que exista um erro, se possa saber onde e se possa agir em relação a ele, pois não ter este erro de que falamos quando desenvolvemos o código, não quer dizer que não o possamos ter numa próxima atualização, pelo que é necessário acompanhar essa atualização, com uma alteração do código fonte de pesquisa desse elemento. Abordando agora a segunda limitação referida nos dois parágrafos anteriores, temos o problema de dados, informação ou elementos que apenas são "carregados" pelo *browser* após o "carregamento" inicial do *website*. Esta dificuldade pode ser superada com uma maior sofisticação dos algoritmos desenvolvidos. Na prática, isto quer dizer adicionando alguma automação que replique o comportamento humano. Por exemplo, em [23] são abordadas estratégias como a simulação de um *click* ou do preenchimento de um formulário ou até de esperar até que certo elemento seja carregado na página ou mesmo a esperar determinado intervalo de tempo entre a execução de determinadas tarefas. Todas estas estratégias são válidas e dependem das particularidades de cada problema. Dando o exemplo desta dissertação, já abordado no parágrafo anterior, apenas as estratégias de dar um certo intervalo de tempo, antes de retirar a informação pretendida da página e retirar essa informação tendo por base a técnica HTML DOM e não a recolha de dados a partir da página inicial endereçada ao *browser*, já faz com que a recolha de dados resulte como pretendido. Depois, é ainda necessário recolher os dados de várias páginas distintas, pelo que simular o comportamento humano de clicar num botão para ir para a página seguinte, também, poderá ser uma boa estratégia a seguir.

2.5.4 Tecnologias e a sua adequação

Atualmente, existem várias tecnologias para executar processos de auxílio ao *web scraping*. Existem plataformas e ferramentas que não necessitam de programação, mas que limitam o utilizador, pois apenas podem ser executadas as tarefas disponibilizadas pela plataforma. Por outro lado, temos a possibilidade de programar de raiz, muitas vezes,

utilizando-se para isso a linguagem C ou Java, para certos projetos pode ser interessante, pois atribui ao programador uma liberdade quase total, contudo é necessário despender mais tempo para desenvolver estes sistemas de raiz, pelo que para compensar tem de ser um projeto muito específico, possivelmente, onde exista uma maior complexidade, para esta liberdade acrescida valer a pena em comparação com o tempo extra despendido. Para projetos entre os dois abordados anteriormente, ou seja, projetos em que a complexidade e especificidade não são extremas nem são muito reduzidas, existem soluções em que parte da programação já foi feita à *priori* por outros programadores e assim podemos através dessas soluções chamá-las no código que estamos a desenvolver, mas mantém um certo grau de personalização ao qual o programador tem controlo, são estas soluções livrarias ou *frameworks*, existentes em várias linguagens de programação, sejam elas Java, PHP, Python ou JavaScript. Nos restantes parágrafos deste capítulo, será abordado não só como podemos comparar as tecnologias à nossa disposição tendo em conta o objetivo final, mas, também, quais são essas soluções e que benefícios e limitações apresentam.

Em termos de soluções, o foco deste capítulo serão as soluções apresentadas como intermédias no parágrafo anterior, ou seja, soluções não muito complexas, mas complexas o suficiente para permitir alguma liberdade ao programador para responder aos desafios desta dissertação. Pelo que já vimos no capítulo anterior, nesta dissertação necessitamos de desenvolver um sistema de *web scraping* que recolha dados de uma tabela gerada de forma dinâmica num *website* específico. Pelo que necessitamos de uma ferramenta tecnológica que vá além de recolher o ficheiro HTML inicial do pedido HTTP GET, utilizando mais as potencialidades de recolha de dados a partir do HTML DOM. Assim sendo, temos várias livrarias em Python, umas que satisfazem este desígnio, outras que nem tanto, em [25] são abordados dois tipos de módulos importantes para este processo de *web scraping*, um primeiro que recolhe os elementos onde a informação que é pretendida recolher no final se encontra e outro módulo em que esses elementos são filtrados para que apenas conste no final a informação pretendida. Por exemplo, no caso de o objetivo final de um processo destes ser o de recolher certas colunas de uma tabela, o primeiro módulo, seria o que nos permitiria aceder à tabela e o segundo módulo seria aquele que retirava a informação da tabela, apenas das colunas pretendidas. Para este primeiro módulo, temos livrarias em Python como a UrlLib2, Selenium ou Pyppeteer, para o segundo módulo temos as livrarias Beautiful Soup ou Pyquery.

O UrlLib2 fornece várias opções para lidar com pedidos HTTP, contudo tem algumas limitações para sites dinâmicos. O Selenium é mais completo, pois fornece não só opções para lidar com pedidos HTTP, como também com sites dinâmicos, pois fornece opções que permitem automatizar o processo de navegar num site tal como um ser humano o faria, o que permite recolher aqueles dados que são gerados após a inicialização do *website*, contudo o Selenium exige um pouco mais de configurações iniciais para instalar um *web browser* compatível com esta livraria. [25] Temos ainda uma livraria mais recente, que foi desenvolvida a partir de outra já existente em JavaScript, o Pyppeteer (Puppeteer em JavaScript), que é bastante semelhante ao Selenium em termos de funcionalidades,

contudo tem uma curva de aprendizagem menor e não necessita de tantas configurações iniciais, contudo como é mais recente, não tem tanta comunidade de suporte e não é tão utilizada como o Selenium.

Para o segundo módulo abordado anteriormente, a livraria Beautiful Soup é a mais utilizada pela comunidade de programadores. Esta livraria fornece funções para navegar, procurar e modificar a informação recebida do primeiro módulo, sendo que de forma automática deteta a codificação desses elementos (no caso desta dissertação serão elementos HTML) e modifica a sua codificação para algo legível e utilizável para o programador. No final, para extrair toda a informação, esta livraria, tem suporte para mais duas livrarias que conseguem tratar desta tarefa, são elas a lxml e a html5lib. A livraria Pyquery é semelhante, contudo as funções que fornece são menos semelhantes visualmente a funções Python como as funções do Beautiful Soup e mais aos pressupostos da biblioteca de JavaScript, JQuery. Em relação às livrarias de extração final a que dá suporte, apenas dá suporte a lxml [25].

Além das opções abordadas nos últimos parágrafos, temos o *framework* Scrapy, também muito utilizada pela comunidade de programadores nesta área do *web scraping*. É um *framework* de Python, pelas funcionalidades que oferece acelera o processo de desenvolvimento e o de escalar um sistema de *web scraping* [25], contudo, como é usual em qualquer *framework*, tem uma linguagem muito própria, que não é limitação quando o sistema a desenvolver é algo muito grande ou complexo, que vai demorar muito tempo a desenvolver e esse acréscimo temporal para aprender a linguagem é compensado mais à frente, contudo para o caso desta dissertação, como nenhum destes fatores se apresenta, esta solução apresenta-se com uma complexidade que apenas por alguma razão muito específica pode fazer sentido, senão as anteriormente abordadas parecem ser ferramentas com maior adequação aos problemas apresentados por esta dissertação.

2.5.5 Testes de performance

Tendo já sido abordado, no parágrafo anterior, algumas das tecnologias mais interessantes para o desenvolvimento desta dissertação, é importante agora abordar como na prática podem ser executados testes para sustentar a escolha com dados sobre a sua performance de cada tecnologia.

Em primeiro lugar, antes de se falar de alguns indicadores de performance que são interessantes de medir, é importante ter em consideração o processo de testagem e comparação em si e como ele é modelado e executado. Em [17] temos um exemplo interessante de como modelar este processo e que se colocou na figura 2.4, onde primeiro são mapeadas as páginas às quais queremos retirar informação através de processo de *web scraping*, em seguida é criado o código fonte, depois passa-se então à preparação dos testes propriamente ditos, depois são feitos testes de execução (testes temporais, ou por outras palavras, de tempo de execução), testes de utilização de memória e testes de utilização de dados, no final são comparados os resultados de cada

teste para cada uma das tecnologias consideradas e retiradas conclusões provenientes desses resultados. Em termos de metodologia utilizada em [17] são feitas 20 execuções de cada código a comparar, para que assim se possa fazer uma média dos valores resultantes da cada execução e assim ter um valor um pouco mais fidedigno e próximo do valor "real", dado que quanto menos execuções executadas, mais vulnerável o nosso teste está a uma má performance proveniente de um mau momento para executar o teste, por exemplo, o computador estava a executar mais tarefas que o normal e ficou mais lento, entre outros exemplos. Assim, quanto mais testes executados, menos se vão fazer notar esses possíveis azares, contudo, também, convém fazer um número de testes que não seja exagerado, em primeiro lugar, porque a partir de certo momento já não vai alterar assim tanto a qualidade dos resultados finais e sendo assim não vale a pena gastar tempo e recursos com essas execuções extra. Assim sendo, nos próximos parágrafos serão então explorados os 3 tipos de testes de performance abordados neste modelo e como podem ser executados.

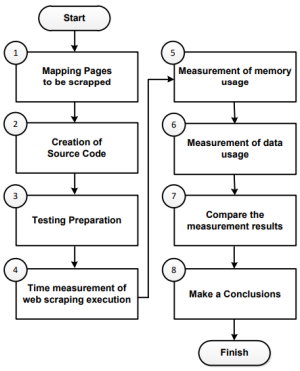


Figura 2.4: Exemplo de modelação do processo de testagem com vista à comparação de soluções tecnológicas [17]

O teste mais fácil de executar é o teste temporal, onde apenas é necessário registar o tempo inicial e final da execução de cada pedaço de código, utilizando cada tecnologia e mantendo as funcionalidades muito semelhantes para não haver discrepâncias temporais devido a um acréscimo de código substancial entre diferentes tecnologias, mas sim, discrepâncias temporais apenas devidas ao desempenho de cada tecnologia diferente à qual pretendemos testar a sua performance. Um exemplo de um teste deste género pode ser encontrado em [26], onde é feito exatamente o processo abordado anteriormente para chegar à conclusão de que processo, entre dois a analisar, teria melhor performance em termos de tempo de execução, um era um processo onde é utilizada uma livraria para executar processos em paralelo e outra que executa tudo sequencialmente. Como seria de esperar, o processo em paralelo é substancialmente mais rápido, mas com a testagem a escolha da livraria para execuções em paralelo está sustentada com dados. O mesmo objetivo terá esta dissertação. Testar as várias livrarias existentes, para escolher aquela que apresente melhor performance, contudo não é só de tempo de execução que se mede a performance e nos próximos parágrafos vão ser descritos mais dois indicadores

apresentados em [17].

De forma similar aos testes de tempo de execução, os testes de utilização de memória, também, são executados retirando os dados pretendidos no início e no fim da execução do programa adaptado para cada livraria a comparar. Pelo menos este é a ideia principal quando falamos de linguagens como o C e o Java, contudo em Python torna-se uma tarefa um pouco mais desafiante, contudo existem livrarias como o guppy e o mprof que auxiliam o programador em todo este processo, pelo que também não se torna uma tarefa assim tão complicada. O principal neste teste é entender a utilização da memória da máquina utilizada, ou seja, o acréscimo de utilização da memória após a execução do código que pretendemos testar.

Por fim, temos o teste de utilização de dados. Neste parágrafo é importante definir a diferença entre utilização de dados e memória. A utilização de memória refere-se à utilização das memórias primária e secundária e da *cache*, ou seja, memórias temporárias para guardar informação com o objetivo de serem armazenadas por um período de curta duração. Enquanto que a utilização de dados, se refere a tipos de memória HDD e SDD, tipos de memória para armazenar dados por um maior intervalo de tempo. Dito isto, o teste explicado no parágrafo anterior é feito a *bytes* gerados em localizações de memória temporária e o teste abordado neste parágrafo é feito a *bytes* gerados na execução do código pretendido em localizações de memória de mais longo prazo, mais precisamente, nas localizações onde são normalmente guardados ficheiros, aplicações e sistemas operativos, por exemplo. Este teste é o mais complexo de ser feito, em [17] foi executado em Java com o auxílio de um *thread*, contudo em Python a tarefa é um pouco mais complexa, pois utilizando as livrarias descritas a cima todo o tipo de memória é fornecido, sem fazer a diferenciação entre utilização dados e memória, pelo que é necessário fazer alguma filtragem adicional a esses resultados e colocá-los nos locais devidos para uma correta obtenção de resultados finais dos testes executados.

2.5.6 Modelo de processos de *web scraping*

Chegando a este capítulo, é importante pensar no processo de extração de dados, ou seja, no seu todo como se modela todo um sistema de *web scraping*, por que fases é que este sistema tem de passar e por que sub-fases tem de passar cada fase? Para isso, vai ser usado como base os casos de estudo apresentados em [27], onde é feito um sistema de onde se retiram dados meteorológicos da Sumatra, de um *website* com informações metrológicas de várias regiões do globo e, vai ainda ser utilizado, o caso de estudo apresentado em [28], onde são retirados dados da rede social Twitter, com o objetivo de comparar a performance deste sistema de *web scraping* em comparação com o Application Program Interface (API) disponibilizado pela própria rede social, em termos não só de performance de desempenho para tarefas semelhantes, mas, também, para a quantidade de tarefas que podem ser executadas, dado que no API existe um limite e no sistema criado a partir de programação potencialmente esse limite é bastante maior

e apenas limitado a possíveis defesas criadas pela rede social no seu site, para que não sejam executados os chamados *bots* a toda a hora na sua plataforma.

Começando pelo *web scraping* de dados meteorológicos em [27], existe uma figura que é citada na figura 2.5 desse *paper*, que descreve o processo geral de um sistema de *web scraping* e que será, também, a base desta dissertação. Observando a figura, o processo é iniciado com a escolha da fonte de dados, que no caso da figura é um site de dados meteorológicos, mas no caso desta dissertação será, inicialmente, um site governamental onde se encontram informação sobre os contratos públicos portugueses. Depois, segue-se a segunda parte do processo onde a fonte de dados é manipulada pelo código desenvolvido pelo programador, que neste caso foi produzido um *script* em Python de recolha dos dados pretendidos, mas existe toda uma panóplia de soluções já abordadas em capítulos anteriores, contudo nesta dissertação irá ser dada preferência a ferramentas em Python, pelo que aqui, também, existem algumas semelhanças, as razões prendem-se pela facilidade de utilização e disponibilização de todas as funcionalidades necessárias. Os resultados deste *script* são, então, a terceira fase deste processo onde os dados são estruturados e armazenados, no caso do processo citado, é dado o exemplo da estrutura e armazenamento em ficheiros, nomeadamente XML e Comma-separated values (CSV), contudo para projetos um pouco maiores e que queiram depois ter ferramentas com mais potencialidades no que diz respeito à análise de dados, pode ser interessante equacionar sistemas ou linguagens de base de dados mais complexos, como por exemplo, linguagens relacionais como a linguagem Structured Query Language (SQL) ou não relacionais como o MongoDB. Para esta dissertação, vai ser relevante equacionar estas soluções de base de dados, dado o volume de dados elevado que pretendemos armazenar e toda a visualização e análise de dados que se pretende executar, mas a utilização de bases de dados e de que bases de dados terá o seu capítulo mais à frente. Terminando agora este parágrafo, apenas falta abordar o último passo deste processo, que se prende com a análise e estatística feitas a partir dos dados estruturados recolhidos na etapa anterior. Nesta dissertação, são utilizadas livrarias de Python que tratam a visualização ordenada em formato tabela dos dados pretendidos, mas, também, bibliotecas que executam algumas estatística sobre os dados, como por exemplo, a média, mas também tratam da apresentação dos dados graficamente, sejam estes dados percentagens ou médias ou outros dados estatísticos relevantes. Além dos métodos estatísticos mais simples, que foram aplicados neste caso de estudo, também, pode ser interessante, dependendo do problema apresentado, pensar em métodos estatísticos mais sofisticados, sendo que aqui me estou a referir aos métodos estatísticos utilizados para suportar processos relacionados com *data science* e inteligência artificial. Para esta dissertação, será relevante fazer algumas estatísticas mais gerais e simples, como a média, mas, também, será equacionada a utilização de métodos mais sofisticados, principalmente, para a parte final desta dissertação.

Olhando agora para o segundo caso de estudo, presente em [28], a abordagem geral é semelhante, como não podia deixar de ser, dado que em traços gerais a figura 2.5,

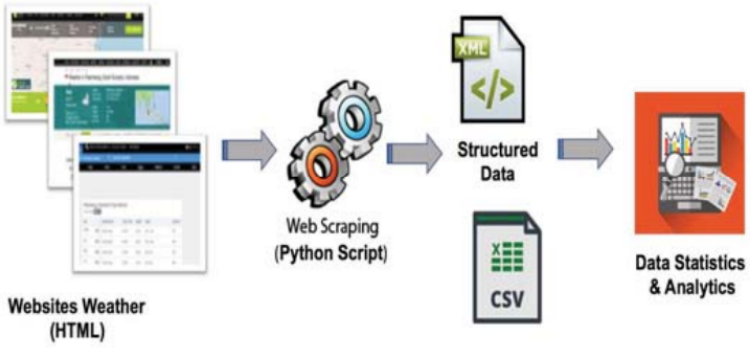


Figura 2.5: Exemplo de um processo de *web scraping* de dados meteorológicos [27]

representa a modelação de alto nível de grande parte dos processos de *web scraping*, por isso, é que é interessante olhar para a figura 2.6 presente no *paper* citado no início deste parágrafo, dado que nesta figura a modelação do processo já apresenta módulos específicos e funcionalidades específicas, pelo que é menos geral que a figura descrita no parágrafo anterior. Assim sendo, este processo tem um módulo com o nome de *downloader* que em traços gerais recebe pedidos HTTP e envia respostas a esses pedidos. Tem ainda, um módulo chamado de Spider, que é um módulo muito importante disponibilizado pelo *framework* Scrapy, que define como se deve extrair a informação com a estrutura pretendida no final, para que depois esta seja guardada no módulo chamado de Item Pipeline, que no caso deste exemplo, a informação é guardada localmente, mas muitos projetos, incluindo esta dissertação é relevante guardá-los em base de dados acessíveis não apenas localmente, para, por exemplo, disponibilizar os dados num *website*. Com algumas modificações referentes ao processo de extração de dados, que neste caso utiliza pedidos HTTP que como se referiu anteriormente pode ter complicações quando trabalhamos com sites dinâmicos e com as devidas adaptações tecnológicas e da fonte de dados, este processo pode ser interessante para ser executado de forma semelhante nesta dissertação, sendo que o Scrapy como já foi visto anteriormente, pode não ser a ferramenta ideal para esta dissertação, contudo é uma das que pode ser interessante testar e mesmo não utilizando esta ferramenta, os princípios gerais são semelhantes entre ferramentas de *web scraping*, logo pode ser interessante tirar ideias deste processo para a modelação do processo de *web scraping* desta dissertação.

2.5.7 Deteção de erros e falhas

Para esta dissertação é importante, como já foi abordado com menos detalhes em capítulos anteriores, ter em atenção à possibilidade de existência de possíveis erros futuros gerados, não totalmente pelo código desenvolvido pelo programador do processo de *web scraping*, mas, também, por alterações no site que serve como fonte de recolha de dados para este processo. Nesse sentido, é necessário acautelar isso no código, para que assim que seja feita uma modificação no código fonte do *website*, que serve como base de

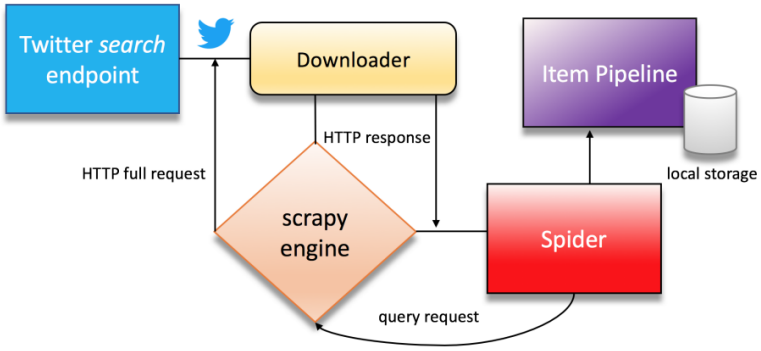


Figura 2.6: Exemplo de um processo de *web scraping* de dados da rede social Twitter [28]

recolha de informação para todo este processo, esta seja percecionada pelo código desenvolvido para o processo de *web scraping* e assim notificada essa falha ao programador, para que possa atualizar o seu código, com o objetivo de o corrigir no sentido do seu correto funcionamento.

Em [29] são apresentados exemplos de códigos de testes específicos na linguagem de programação C, que podem ser facilmente adaptados para outras linguagens de programação, tais como o Python. No Python existe a declaração *try and catch*, que consiste exatamente em dois blocos, o *try* onde se pode colocar o teste que se pretende fazer ao código para verificar o seu correto funcionamento, fazendo o paralelismo para o código em C apresentado no *paper* citado no início deste parágrafo, um *if* mais complexo e sofisticado, caso o código presente no *try* apresente erros, então entram em ação os blocos *catch*, que são os blocos que definem o que fazer com cada erro específico, ou seja, fazendo novamente um paralelismo com o código em C, os blocos *catch* acabam por ser um *else* ou vários *else if*, bastante mais sofisticados em termos de programação, mas de grosso modo com a mesma função.

No caso específico desta dissertação, os erros a ter em atenção são chamadas de elementos que não retornam os elementos pretendidos e paragens na execução do código devido a este estar "à espera" de receber elementos que não existam da maneira como foram chamados. Para isso, é necessário modelar o comportamento do sistema em si e as respostas e informações que se pretende retirar no final e nos passos intermédios do sistema, para depois então poder produzir estes testes através, por exemplo, dos sistemas de *try and catch*, para a partir da comparação com os resultados pretendidos seja possível averiguar o correto funcionamento do sistema em questão.

2.6 Bases de dados

No seu site oficial, a Oracle define uma base de dados como "uma coleção organizada de informação estruturada, ou dados, tipicamente armazenados eletronicamente num sistema computacional" [30]. Fazendo o paralelo com o que foi abordado nos outros capítulos, uma das fases finais de um processo de *web scraping* é a obtenção de dados de

forma estruturada, para depois os guardar, seja numa folha de cálculo ou numa base de dados. Uma base de dados, não é mais que uma forma de armazenamento de dados altamente otimizada para este objetivo da estruturação, acesso, alteração e edição dos dados de forma rápida e eficaz. Uma base de dados, normalmente, faz parte de um sistema, para que seja possível aceder-lhe e fazer operações com os seus dados, a essa ferramenta que permite controlar a base de dados chama-se de Database Management System (DBMS). O que se chama de sistema de base de dados é então todo o sistema formado pela base de dados em si, com todos os dados, o DBMS e a aplicação associada a estes dois, que será então o objetivo máximo deste sistema, que essa aplicação faça algo de relevante com as funcionalidades de acesso e manipulação da base de dados, que lhe são fornecidas pelo DBMS.

2.6.1 Base de dados relacionais vs não relacionais

Sendo ambas as bases de dados as mais utilizadas para aplicações *web* e um dos objetivos finais desta dissertação ser exatamente o armazenamento e estruturação dos dados obtidos em base de dados, então este capítulo vai focar-se na definição de cada tipo de base de dados, as suas características, vantagens e desvantagens e no fim abordar alguns resultados de testes feitos às suas performances individuais.

Uma base de dados relacional não é mais que um conjunto de tabelas, onde os seus dados estão dispostos nas linhas dessa tabela e cada coluna representa as categorias da informação da tabela. Por exemplo, na figura 2.7, temos o exemplo de uma base de dados que se criou na ferramenta de gerenciamento de bases de dados do tipo PostgreSQL, pgAdmin [31], com algumas informações sobre livros. Cada linha da tabela representa um livro. Cada coluna representa uma categoria atribuída ao livro, como por exemplo, o id (única categoria que tem de ser única, neste caso, pois é uma *primary key*), o nome da imagem associada ao livro (img_name), o título, o autor, a editora e o idioma. As bases de dados relacionais permitem ainda criar as chamadas constraints, que especificam regras para os dados da tabela, por exemplo, se os valores de certa coluna são únicos, ou não nulos, ou se satisfazem uma certa condição. O que define e dá nome às bases de dados relacionais é então a característica de relação entre duas tabelas através das suas chaves primárias e estrangeiras, por exemplo, no caso anteriormente abordado a tabela dos livros pode estar relacionada com uma tabela de clientes que são donos desses livros, ou seja, cada cliente tem um id (uma chave primária), que será a chave estrangeira do livro do qual é dono, totalmente distinta em termos de função da chave primária do livro, dado que a chave primária é um valor único e irrepetível na tabela, que identifica esse registo, ou linha, da tabela, enquanto que a chave estrangeira, pode ser repetida na mesma coluna de uma tabela, pois apenas faz a relação entre uma linha da coluna em que é mencionada com outra linha de outra tabela.

Para recolha, edição, remoção e atualização de dados, as bases de dados relacionais fornecem as chamadas *queries*. Uma das linguagens mais conhecidas para manipulação de

dados de bases de dados relacionais é a SQL. Nestas *queries*, temos então a possibilidade de editar os dados das tabelas através da *query Update*, adicionar novos dados através da *query Insert*, remover dados através da *query Delete*, recolher/visualizar dados através do *Select*, remover tabelas através do *Drop* e ainda fazer operações entre tabelas com os *Joins*. [32]

Data Output | Explain | Messages | Notifications

	id [PK] integer	img_name character varying (100)	titulo character varying (100)	autor character varying (100)	editora character varying (100)	idioma character vary
1	1	factfulness	Factfulness	Hans Rosling com Ola Rosling...	Temas e Debates	PT
2	2	ensaio-sobre-a-cegueira	Ensaio sobre a Cegueira	José Saramago	Caminho	PT
3	3	gerir-o-seu-dia-a-dia	Gerir o seu Dia a Dia	Jocelyn K. Glei	Dom Quixote	PT
4	4	the-portuguese	The Portuguese	Barry Hatton	Clube do Autor	EN
5	5	velocity	Velocity	Stefan Olander e Ajaz Ahmed	Ebury Publishing	EN
6	6	supercerebro	Supercérebro	Deepak Chopra e Rudolph E. T...	Self PT	PT
7	7	desobedecer	Desobedecer	Frédéric Gros	Antígona	PT
8	8	porque-falham-as-nacoes	Porque Falham as Nações	Daron Acemoglu e James A. ...	Temas e Debates	PT
9	9	palavra-que-falam-por-nos	Palavra que Falam por Nós	Pedro Braga Falcão	Clube do Autor	PT
10	10	o-livro-das-religioes	O Livro das Religiões	Jostein Gaarder	Editorial Presença	PT
11	11	manual-das-financas-pessoais	Manual das finanças pessoais	Ricardo Ferreira e João Pesso...	Arcádia	PT

Figura 2.7: Exemplo de uma tabela de uma base de dados de livros de um projeto pessoal exemplo feito no pgAdmin [31]

As bases de dados relacionais têm algumas limitações quando trabalhamos com um largo volume de dados. A partir de certo ponto, mesmo melhorando o *hardware*, é difícil de escalar. A sua estrutura em tabelas, para um grande volume de dados, pode tornar a estrutura dos dados muito complexa, se os dados não forem fáceis de encapsular no formato de tabela. Muitas das funcionalidades das bases de dados relacionais não são utilizadas pelo que apenas adicionam complexidade e custo aos sistemas desenvolvidos. A linguagem mais utilizada, SQL, é muito complexa quando é utilizada em dados não estruturados. Por último, quando o volume de dados é muito grande, a base de dados tem de ser distribuída por vários servidores, o que traz vários problemas dado que fazer *Joins* em servidores distribuídos é uma tarefa complexa [32].

Com o intuito de resolver alguns destes problemas, foram criadas as bases de dados não relacionais. O objetivo principal destas bases de dados, é resolver os problemas das bases de dados relacionais no que diz respeito ao armazenamento e acesso a grandes volumes de dados, tentando arranjar soluções mais eficientes e otimizadas para esta problemática. A característica mais importante das bases de dados não relacionais e que as diferencia estruturalmente das relacionais, é exatamente esse conceito de relação que não existe nas bases de dados não relacionais, ou seja, não existe nenhuma forma direta de declarar estas relações, como existe com as chaves primárias e estrangeiras nas bases de dados relacionais. Estas bases de dados já não utilizam a linguagem SQL, logo não permite operações de *Join* de forma direta, também, não garante as propriedades das transações das bases de dados relacionais Atomicidade, Consistência, Isolamento e Durabilidade (ACID) e pode ser escalável horizontalmente, sendo uma das principais razões porque é escolhida para a utilização em sistemas com um grande volume de dados, dado que esta característica de escalar horizontalmente, ou seja, na prática,

adicionar mais servidores e ter os tais servidores distribuídos como armazenamento, é muito mais fácil que em bases de dados relacionais. As bases de dados não relacionais têm várias maneiras de organizar os seus dados, sendo que deixam de se organizar em tabelas, pois não são bases de dados relacionais, estas bases de dados podem ser então organizadas por um sistema de armazenamento de chave valor, por um sistema de armazenamento baseado em documentos (por exemplo: XML), por um sistema baseado em grafos, colunas, objetos, entre outros. Contudo, uma das grandes vantagens das bases de dados não relacionais, pode ser, também, uma das suas grandes desvantagens, a flexibilidade, ou seja, este tipo de base de dados não cumpre o modelo transacional das bases de dados relacionais, o ACID, pelo que ao não forçar a consistência dos dados, logo quando estes são inseridos na base de dados, dá mais flexibilidade ao programador para tratar disto mais tarde, mas pode trazer alguns problemas adicionais de segurança e de confiabilidade dos dados [32].

Para terminar este sub-capítulo sobre bases de dados relacionais e não relacionais, é interessante olhar para os resultados de [33], onde foram executados testes de performance às principais operações de bases de dados, sendo elas *Insert*, *Select*, *Update* e *Remove*. Para estes testes foram utilizadas as bases de dados MSSQL (relacional) e MongoDB (não relacional). Em termos de formato, cada uma das operações especificadas anteriormente, foram executadas em cada uma das bases de dados para 1 registo, 100, 500, 1.000, 5.000, 10.000, 25.000 e 50.000, com o intuito de testar a performance destas operações quando expostas a quantidades de pedidos maiores, simulando assim, por exemplo, operações que podem acontecer quando temos uma aplicação com muitos utilizadores a utilizá-la em simultâneo. Na figura 2.7, temos o resultado destes testes. Olhando para os gráficos presentes na figura, podemos concluir que apenas na operação *Select*, a base de dados relacional ganha vantagem quanto mais operações executadas, em todas as outras a base de dados não relacional é mais rápida a executar as operações pretendidas. Estes resultados sustentam com dados as características abordadas nos parágrafos anteriores de cada tipo de base de dados. Como era de prever, a performance da base de dados não relacional é melhor, ou por outras palavras, demora menos tempo a executar operações que não dependam da forma como os dados estão estruturados, como é o caso das três operações (*Insert*, *Update* e *Delete*), pois aqui o seu fator de flexibilidade e não concordância à *priori* com o modelo de transação de base de dados ACID, faz com que seja mais rápido inserir, editar e remover elementos da base de dados, dado que são necessárias menos verificações para proceder a estas operações. Comprova-se a ligação entre quanto maior o volume de dados, maiores os benefícios das bases de dados não relacionais sobre as relacionais, exceto para a operação *Select*, o que pode fazer sentido, dado que para esta operação, a maior organização e os *standards* pré-estabelecidos, podem facilitar e acelerar a pesquisa por dados, que é, no fundo, o que faz esta operação, pelo que a desvantagem deste tipo de base de dados em todas as outras operações é uma vantagem para esta em específico.

Figura 2.8: Tempo de execução de várias operações sobre bases de dados relacionais e não relacionais por quantidade de registos executados [33]

2.6.2 Modelação de bases de dados

Existem vários modelos que podem ser utilizados para modelar a estrutura de uma bases de dados. Uma boa definição do que é modelação de dados pode ser encontrada no site da SAP sobre esta temática, em que este conceito é definido como: "o processo de esquematizar num diagrama os fluxos de dados", sendo que é a partir destes fluxos de dados que depois o engenheiro encarregue de desenhar a estrutura dos dados e da base de dados, vai modelar o formato dos dados, a estrutura da base de dados e as funções que lidam com as tarefas recebidas pela base de dados, para assim tentar otimizar estes processos no sentido de tornar eficiente todos estes processos de fluxo de dados. Quando pensamos neste processo de modelação de dados, existem três níveis de abstração de dados que devem ser pensados. O primeiro nível é conceptual, que é como se fosse apenas um rascunho geral de todo a base de dados que queremos modelar, apenas com a estrutura geral e os elementos principais, sendo, normalmente, o pontapé de saída no processo de modelação, o nível que prepara os seguintes. O segundo nível de abstração é o nível lógico, em que se adiciona aos elementos e estrutura gerais que se idealizou no nível anterior, é aqui neste nível que se especifica um pouco mais o conteúdo de cada tipo de dados da base de dados, as suas relações e os fluxos de dados. Por fim, o último nível é o físico em

que se pormenoriza como o nível lógico vai ser implementado, para facilitar o trabalho do engenheiro que o for executar, quanto melhor estiver feito este nível, mais fácil e direto será para o engenheiro passar desta modelação, para a base de dados a ser criada [34].

Dentro dos vários modelos existentes, o mais relevante para esta dissertação é o modelo ER, este modelo consegue modelar várias características importantes para o processo de modelação de bases de dados, como as relações entre entidades e caracterização do tipo de dados necessários para o nível lógico do processo de modelação. Sendo assim, consegue resolver algumas limitações de outros modelos, incorporando algumas características interessantes deles, mas tendo mais algumas extra que resolvem as suas limitações, como os modelos de rede e de entidades, onde a modelação do nível conceptual é fácil de ser executado, mas o lógico já é mais complicado. [35]

Na figura 2.9 pode ser observado um exemplo da modelação dos dados de uma base de dados através do modelo ER. Neste exemplo, podemos observar a modelação das relações entre entidades através das linhas que unem duas entidades e que têm terminações distintas consoante o tipo de relação, por exemplo, no caso da relação entre a entidade *user* e a entidade *lab*, temos uma relação de um para vários. Temos ainda em cada entidade o seu nome, o nome de cada um dos seus atributos e ainda o tipo de dados que representa cada atributo.

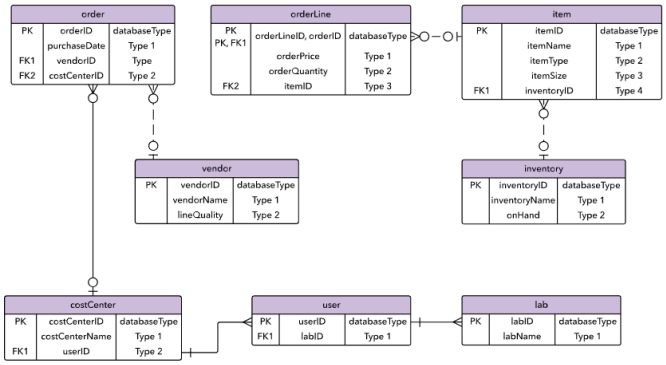


Figura 2.9: exemplo da modelação dos dados de uma base de dados através do modelo ER [36]

2.7 Desenvolvimento de aplicações para a *web*

Quando abordamos a temática do desenvolvimento *web*, o primeiro pensamento que ocorre são *websites*. Em certa medida não está incorreto, mas hoje em dia a tecnologia já evoluiu tanto que chamar *websites* às aplicações *web* muito mais complexas que os primeiros sites estáticos é bastante incompleto. Atualmente, é necessária toda uma arquitetura para desenvolver este tipo de aplicações *web*, que inclui servidores de hospedagem, bases de dados, servidores de domínio e própria programação já não é apenas de *front-end*, ou seja, o aspeto visual do *website* em que se utilizaria linguagens como o HTML, CSS e o JS, mas, também, linguagens de *back-end* cada vez mais

sofisticadas e otimizadas para *web* e para a interligação entre os vários componentes da aplicação e cada vez mais uma atualização quase que em tempo real da informação presente no *website*. Além disto, estas aplicações têm, cada vez mais, de lidar com maiores volumes de dados, pelo que para manterem performances temporais interessantes para o utilizador, as tecnologias e métodos utilizados tiveram de evoluir bastante para satisfazer esta necessidade [37].

Um aspeto muito importante do desenvolvimento de uma aplicação *web* é a sua modelação antes de partir para a fase de programação propriamente dita. Em [38] pode ser observado um exemplo de modelação utilizando uma abordagem de *design* de aplicações *web* focada na modelação, o Semantic Hypermedia Design Method (SHDM). Resumidamente, este método faz uma divisão clara entre modelação conceptual e modelação de navegação e divide essa navegação em 5 passos: o levantamento de requisitos, o *design* conceptual, o *design* da navegação, o *design* da interface abstrata e a implementação, sendo que as fases de *design* conceptual e de navegação, ainda se sub-dividem em mais algumas sub-fases para que ainda seja mais eficaz a modelação, sendo ainda mais específica e mais repartida.

Uma linguagem que é bastante usada e facilita bastante a modelação de aplicações *web*, como aquela que se pretende desenvolver nesta dissertação é a linguagem Unified Modeling Language (UML). Em [39], temos vários exemplos de como esta linguagem pode facilitar a modelação dos vários *frames* de uma aplicação *web*, ou seja, das várias janelas que o utilizador observa quando navega no *website*, pode, também, modelar o comportamento de um elemento em específico, como, por exemplo, um formulário, pode modelar ainda algumas classes principais da aplicação *web*, como uma classe que representa o comportamento da página inicial ao carregar, uma barra de pesquisa quando é utilizada, etc., por último pode ainda modelar os comportamentos do cliente e do servidor, modelando as suas classes, as relações entre elas e ainda os seus atributos e métodos.

Para terminar esta secção e capítulo, falta ainda referir as tecnologias utilizadas para o desenvolvimento de aplicações *web*. Três das mais utilizadas atualmente, são o PHP, o Python e o Node.js, sendo que as duas últimas começam a ser cada vez mais utilizadas em detrimento da primeira tecnologia referida. Em [40] são feitos testes de performance às 3 tecnologias, para fazer um estudo comparativo das vantagens e desvantagens de cada uma delas na performance de uma aplicação *web*. Este estudo conclui que em situações de alta concorrência o Node.js tem performances muito melhores nas suas execuções do que o PHP, dado que esta tecnologia comporta-se bem com pequenos pedidos, mas tem algumas dificuldades em dar resposta a vários pedidos ou pedidos de maiores dimensões. O Python tem grandes ferramentas, com já bastante maturidade para desenvolver plataformas de grandes dimensões, enquanto que o Node.js tem esse potencial, mas ainda é uma plataforma muito recente. O Node.js comporta-se bem em ambientes *IO-intensive*, ao contrário do Python, mas ambas as tecnologias não têm grandes desempenhos em ambientes de *compute-intensive*, sendo que ao Node.js ainda

acresce a dificuldade de depender bastante de programação assíncrona, pelo que pode ser uma dificuldade acrescida para alguns programadores.

2.8 Sumário do capítulo

A inovação refere-se a um novo processo ou produto que resulta num acréscimo de produtividade, trazendo diversos benefícios do ponto de vista macroeconómico, como crescimento do produto nacional bruto, aumento da produtividade e do nível de emprego, como, também, do ponto de vista microeconómico, como maior lucro e maior competitividade do tecido empresarial.

Em Portugal, o governos tem financiado a inovação através de contratos públicos, sendo estes ferramentas muito interessantes para fomentar tanto a inovação no setor privado, como no setor público.

Das várias maneiras de medir inovação, mesmo compreendendo as suas limitações, a que se achou mais adequada para utilização nesta dissertação foi o número de patentes registadas por cada empresa, dado que apenas o ato de registar uma patente já demonstra um comportamento inovador e existe bastante informação disponível sobre patentes registadas por empresas portuguesas.

Para se estudar o impacto do financiamento através de contratos públicos de estímulo à inovação das empresas será usada uma metodologia que compreende cada uma das seguintes fases: extração, armazenamento, limpeza, análise e visualização de dados. A revisão da literatura apresentada reforça a adequação das técnicas de *web scrapping*, armazenamento de dados em base de dados e algumas potencialidades do desenvolvimento de aplicações *web*, para a realização da metodologia apresentada, além das boas práticas mais adequadas em cada uma dessas técnicas, tais como o controlo de erros e os cuidados específicos a ter quando se lida com *websites* dinâmicos.

3

Extração, armazenamento, estruturação e limpeza de dados

Neste capítulo serão apresentadas as metodologias e algoritmos utilizados e desenvolvidos para extrair os dados necessários para o estudo que esta dissertação se propõe a desenvolver das diferentes fontes onde estes se encontram na *web*, através de técnicas de *web scraping*, irão apresentar-se os algoritmos desenvolvidos. São também referidas as formas como foram armazenados e estruturados os dados em base de dados. Para terminar, será também referido as diferentes maneiras com que se executou limpezas aos dados para estarem prontos para as análises necessárias de efetuar posteriormente aos mesmos.

3.1 Fontes de dados

Como fontes de dados foram escolhidas 4 fontes, cada uma com o seu propósito diferente, nomeadamente, uma primeira, o Portal BASE para recolher informações sobre contratos públicos portugueses, uma segunda, o Racius, para recolher o(s) CAE(s) de cada empresa desses contratos públicos recolhidos na primeira fonte referida, uma terceira, a *Orbis*, para recolher informações financeiras e de dimensão das empresas e uma última, o INPI para recolher informações sobre as patentes registadas por cada empresa. Nos próximos parágrafos irá falar-se com mais detalhe de cada fonte de dados.

O portal BASE [24], foi de onde se extraio a base de trabalho desta dissertação, os contratos públicos de inovação. Neste portal encontra-se informação centralizada sobre os contratos públicos celebrados pelo estado português. Aqui é possível obter dados relevantes sobre estes contratos, tais como as datas em que estes foram celebrados, as entidades presentes no mesmo (adjudicatárias e adjudicantes), o valor monetário pelo qual foi celebrado, duração do mesmo, local onde será executado e ainda (e talvez um dos mais importantes) o(s) CPV(s), um código usado pelo estado português para categorizar o âmbito do contrato público, no caso desta dissertação apenas foram utilizados os contratos com CPVs que se consideraram com maior ligação e mais direta com práticas de inovação, neste caso os CPVs iniciados pelo número 73, que podem ser observados na tabela 3.1.

CPVs	Descrição
73000000-2	Serviços de investigação e desenvolvimento e serviços de consultoria conexos
73100000-3	Serviços de desenvolvimento experimental e de investigação
73110000-6	Serviços de investigação
73111000-3	Serviços relacionados com laboratórios de investigação
73112000-0	Serviços de investigação marinha
73120000-9	Serviços de desenvolvimento experimental
73200000-4	Serviços de consultoria em matéria de investigação e desenvolvimento
73210000-7	Serviços de consultoria em matéria de investigação
73220000-0	Serviços de consultoria em matéria de desenvolvimento
73300000-5	Conceção e execução em matéria de investigação e desenvolvimento
73400000-6	Serviços de investigação e desenvolvimento sobre material de segurança e defesa
73410000-9	Investigação e tecnologia no domínio militar
73420000-2	Estudo de pré-viabilidade e demonstração tecnológica
73421000-9	Desenvolvimento de equipamentos de segurança
73422000-6	Desenvolvimento de armas de fogo e munições
73423000-3	Desenvolvimento de veículos militares
73424000-0	Desenvolvimento de vasos de guerra
73425000-7	Desenvolvimento de aeronaves, mísseis e naves espaciais militares
73426000-4	Desenvolvimento de sistemas eletrónicos militares
73430000-5	Ensaios e avaliações
73431000-2	Ensaio e avaliação de equipamentos de segurança
73432000-9	Ensaio e avaliação de armas de fogo e munições
73433000-6	Ensaio e avaliação de veículos militares
73434000-3	Ensaio e avaliação de vasos de guerra
73435000-0	Ensaio e avaliação de aeronaves, mísseis e naves espaciais militares
73436000-7	Ensaio e avaliação de sistemas eletrónicos militares

Tabela 3.1: Lista de CPVs considerados de inovação nesta dissertação

O *website* racius.com disponibiliza informações relevantes sobre as empresas presentes no seu banco de dados, que podem ser encontradas através de pesquisa por nome ou Número de Identificação Fiscal (NIF) no próprio *website*. Para esta dissertação, utilizou-se esta fonte apenas para a recolha de informação do CAE de cada empresa presente nos contratos recolhidos no portal BASE.

A base de dados *Orbis* [41] foi utilizada para recolher dados financeiros e de dimensão das empresas presentes nos contratos públicos de inovação. Verificou-se que dados que poderiam ser interessantes relativos aos gastos das empresas com I&D e outros indicadores financeiros também ligados a práticas de inovação, se encontram em número reduzido nesta fonte de dados, dado tratar-se de uma base de dados com pouco foco em empresas portuguesas, pelo que dados mais específicos como estes se encontram em menor número. Assim sendo, como a maioria das empresas presentes nos contratos públicos de inovação celebrados em Portugal, são empresas portuguesas, estes dados não foram recolhidos. Foram então recolhidos desta base de dados, informações referentes à dimensão de cada empresa, como o número de colaboradores da mesma e dados

financeiros, como a receita bruta da empresa.

Por último, a base de dados de patentes do INPI [42] foi utilizada para retirar informações sobre patentes registadas pelas empresas presentes nos contratos públicos de inovação. Em primeiro lugar, foi retirado o número de patentes registadas por cada empresa e, em segundo lugar, foram retirados os anos de registo de cada patente.

3.2 Algoritmos de extração de dados

Para extrair todos os dados necessários para esta dissertação foi necessário desenvolver três algoritmos de extração de dados das fontes de dados mencionadas anteriormente, sendo que apenas não se desenvolveu nenhum algoritmo para a extração de dados da base de dados *Orbis* dado que esta apresentava uma solução simples de o fazer de forma manual e em massa, que irá ser abordada brevemente nos próximos parágrafos. Os algoritmos foram desenvolvidos tendo como recurso chave a livraria Python Pyppeteer.

O critério que fez com que se optasse por obter de forma manual a informação necessária da fonte mencionada no parágrafo anterior, foi exatamente o que faltou nas restantes fontes, a não existência de uma forma rápida, fácil e em massa de obter a informação, pelo que se desenvolveram algoritmos que quando despoletados de forma automatizada executam as tarefas que um ser humano teria de executar manualmente, de forma repetitiva e durante um longo período de tempo. Estes algoritmos têm as vantagens de desocupar o programador dos mesmos destas tarefas e executá-las de forma mais rápida e precisa.

Para extrair os dados referentes aos contratos públicos de inovação do portal BASE, começou-se primeiro por desenvolver uma solução dividida em duas partes, mas que rapidamente foi abandonada por ser mais eficiente juntar as duas em apenas uma solução. Os principais desafios destes algoritmos foram lidar com o facto de o portal se tratar de um *website* dinâmico e o controlo de erros. O primeiro desafio prende-se com o facto de o conteúdo do *website* ser carregado por injeção na linguagem de programação JS no momento do carregamento do mesmo ou quando é selecionado algum botão para alterar a página de visualização, pelo que não é possível carregar a informação em massa pelo HTML inicial, que ainda não irá conter os dados pretendidos. Nesse sentido, foi necessário desenvolver um algoritmo que retirasse a informação do *website* por leitura do mesmo, após os dados serem carregados, pelo que se teve de programar que este esperasse pelo aparecimento de certos elementos no *website*, para que depois sim os pudesse obter, sem retornar um erro, por eles ainda não existirem. Além deste aspeto foi ainda necessário desenvolver um algoritmo que recolhesse os dados dos contratos um a um, pois nem toda a informação estava na página inicial, em que estes se encontravam numa tabela. Começou-se por dividir este processo em dois momentos, em que primeiro se recolhia o Uniform Resource Locator (URL) da página individual de cada contrato, onde se encontram todas as informações do mesmo e depois num segundo momento se ia a cada URL extrair os

dados de cada contrato. Contudo, rapidamente se percebeu, que se ganhava em juntar os dois apenas num, dado que a tarefa onde é despendido mais tempo para a sua execução, tipicamente é a tarefa de carregamento de uma nova página *web* ou de nova informação dentro da mesma página, pelo que na segunda solução se carregava apenas uma vez as páginas iniciais com as tabelas de contratos, ao contrário da segundo que se faria este trabalho em duplicado em cada uma das duas partes do algoritmo. No início do processo é necessário fornecer ao algoritmo um URL inicial por onde este começa a ser executado, para otimizar o processo foi dado um URL em que já se fazia a pesquisa no portal BASE, pelo número 73 nos CPVs, contudo foi necessário ainda fazer a verificação no algoritmo se esse número correspondia aos dois dígitos iniciais, como pretendido, ou não. Se sim, era iniciada a extração das informações desse contrato, senão continua-se para o contrato seguinte. Optou-se por extrair todos os dados como *strings*, devido ao facto de nem todos os dados estarem devidamente normalizados e assim esta não ser uma fonte de erros no algoritmo, sendo tratada mais adiante, assim que necessário, na limpeza dos dados. As tabelas de onde se retirou a informação de cada contrato, nem sempre se encontravam com a mesma estruturas e/ou com os mesmos cabeçalhos, pelo que foi necessário, sempre que se encontrava um erro deste género, entendê-lo e acrescentar uma exceção no algoritmo para tratar estes casos e recolher a informação corretamente ou sempre que isto não era possível, apenas fazer com que o algoritmo o entendesse como uma exceção e continuasse para a próxima iteração. Para terminar, foram desenvolvidas verificações mais gerais, para que se o algoritmo ficar preso em algum local devido à perda de internet, à lentidão da resposta do *website* ou à manutenção do mesmo, se termine o processo caso necessário ou se continue para a próxima iteração, sinalizando o erro na atual. Teve-se ainda o cuidado de adicionar algumas linhas de código com o efeito de o algoritmo começar de onde tinha terminado e não tudo do início, sempre que ocorresse um erro que interrompesse a execução do mesmo. Desta extração foram extraídos 8240 contratos públicos de inovação.

Após a extração mencionada no parágrafo anterior, são recolhidos o nome e respetivo NIF de cada empresa presente nos contratos. Este processo será explicado com mais detalhe no sub-capítulo da limpeza de dados, apenas é referido aqui por ser a partir destes dados que se fazem as extrações que serão abordadas nos próximos parágrafos. Deste processo foram recolhidas 3591 empresas.

O segundo algoritmo desenvolvido foi o que extrai do racius.com o(s) CAE(s) de cada empresa presente nos contratos públicos de inovação. Neste algoritmo, além da parte de extração de dados propriamente dita, foi também necessário desenvolver código que executasse a pesquisa por NIF na página inicial dado como *input* ao algoritmo. Novamente, foi necessário desenvolver estratégias para que não fosse colocada informação na barra de pesquisa antes desta ser carregada e para que não se tentasse selecionar resultados que ainda não tinham aparecido, tais como os resultados da pesquisa das empresas por NIF e os dados que se pretendem extrair após o redirecionamento originado pela seleção anterior da empresa resultante da pesquisa. Desta vez não foi necessário adicionar nenhuma verificação extra, dado que os dados se

encontravam normalizados e estruturados sempre da mesma maneira. Apenas se fez as verificações gerais, muito semelhantes às do primeiro algoritmo, mencionadas anteriormente. Apenas, se teve de ter o cuidado de recolher somente os dados das empresas que existiam no racius.com, bastando apenas para isso, ler o resultado da pesquisa e se este não resultasse em nenhuma empresa, então passar à próxima iteração do algoritmo.

O terceiro e último algoritmo desenvolvido, foi o de recolha de informações sobre patentes na base de dados para este efeito do INPI. Neste algoritmo, à semelhança do primeiro, começou-se por desenvolver em duas partes um algoritmo para extrair o número de patentes registadas por cada empresa e depois um algoritmo que recolhesse os anos de registo de cada patente. No final, optou-se por juntar os dois para que assim se obtivesse ganhos de desempenho, no sentido em que se elimina execuções duplicadas e, como será explicado com mais detalhe no próximo capítulo, apenas se chama uma vez a base de dados para os armazenar. Para este algoritmo, tal como o segundo, é necessário em primeiro lugar fazer a pesquisa por NIF, verificar se há resultados de patentes, se não houver será atribuído o valor 0 ao número de patentes registadas e continua-se para a próxima iteração, se houver extraísse do fundo da página o número de patentes registadas e, uma a uma, entra-se na página de informações de cada patente da empresa e recolhe-se o ano de registo para um vetor que será posteriormente guardado, devidamente ordenado, em base de dados, como será abordado no próximo capítulo com mais exatidão. Tal como os restantes algoritmos, é necessário esperar por algum elemento do Document Object Model (DOM) que comprove que os dados que queremos recolher já se encontram na página, para assim evitar erros neste âmbito. Foram extraídas informações de 2056 patentes.

Para terminar foram ainda extraídos manualmente dados referentes à receita bruta e número de colaboradores de algumas das empresas presentes nos contratos públicos de inovação da base de dados *Orbis*, leia-se algumas por nem todas existirem nesta base de dados, que não tem um foco nas empresas portuguesas tão grande como tem noutros países, mais relevantes economicamente no panorama mundial. Esta extração manual foi feita fazendo uso em primeira instância da ferramenta pgAdmin, para descarregar um ficheiro CSV com os NIFs das empresas dos contratos públicos de inovação. Este ficheiro era depois carregado na ferramenta de pesquisa da *Orbis* e após seleção dos dados a recolher de cada empresa é dada a opção de descarregamento de outro ficheiro CSV com toda esta informação. É a partir deste ficheiro que se fará algumas das análises presentes no próximo capítulo, mas antes de passar para este tema, irá abordar-se ainda o armazenamento e a limpeza de dados, matéria dos próximos subcapítulos.

3.3 Armazenamento de dados

Em cada algoritmo de extração de dados, no final da sua execução ou de uma iteração do mesmo existe tipicamente uma parte composta pelo armazenamento dos dados

	Data da publicação date	Tipos de contrato character varying[]	Nº do acordo quadro integer	Descrição do acordo quadro character varying	Medida especial character varying	Tipo de procedimento character varying	Descrição character varying
1	2022-12-15	{"Aquisição de serviços"}	[null]	[null]	[null]	Ajuste Direto Regime Geral	PARTICIPAÇÃO N
2	2022-12-15	{"Aquisição de bens móv...	5152738	CP 2021/23 - Acordo Quadro para f...	[null]	Ao abrigo de acordo-quadro ...	52005622
3	2022-12-15	{"Aquisição de bens móv...	[null]	[null]	[null]	Concurso público	CP 25-2022 - Aqu
4	2022-12-15	{"Aquisição de bens móv...	[null]	[null]	[null]	Concurso público	CP 25-2022 - Aqu
5	2022-12-15	{"Aquisição de serviços"}	[null]	[null]	[null]	Ajuste Direto Regime Geral	AM 14 0316/202
6	2022-12-15	{"Aquisição de serviços"}	[null]	[null]	[null]	Ajuste Direto Regime Geral	ADG/207/2022/I
7	2022-06-12	{"Aquisição de serviços"}	[null]	[null]	[null]	Ajuste Direto Regime Geral	Projeto Pedagóg
8	2022-12-15	{"Aquisição de bens móv...	[null]	[null]	[null]	Ajuste Direto Regime Geral	Aquisição de 2 e
9	2022-12-15	{"Aquisição de bens móv...	[null]	[null]	[null]	Ajuste Direto Regime Geral	DM CLINICO - CC

Tabela 3.2: Tabela contratos da base de dados PostgreSQL

extraídos, no caso desta dissertação em base de dados PostgreSQL, pois a sua base é SQL, mas apresenta alguns benefícios de desempenho, foi possível utilizá-la gratuitamente para os fins desta dissertação e tem boas livrarias e fáceis de utilizar de Python para integrar com os algoritmos também eles nesta linguagem de programação. Além disto, é também importante mostrar de forma geral como estão estruturados os dados. O armazenamento de dados foi desenvolvido tendo a livraria Python Psycopg2 como principal ferramenta.

Os dados extraídos pelos métodos referidos no subcapítulo anterior e utilizados nas fases seguintes desta dissertação, referentes a análise e visualização destes mesmos dados, estão armazenados em duas tabelas da mesma base de dados. Uma tabela com os contratos, com todos os contratos públicos de inovação recolhidos do portal BASE e uma outra tabela com as empresas presentes nesses contratos, sendo que os dados provenientes do portal BASE são apenas o nome, NIF da empresa, uma coluna para o tipo de empresa, para saber se é adjudicatária ou adjudicante nos contratos em que está presente, uma coluna com um vetor de URLs de cada contrato em que a empresa está presente e um vetor com os anos de cada um destes contratos para cada empresa. Apenas os dados referentes a CAEs, recolhidos do racius.com e a patentes, número de patentes e anos de registo das patentes foram recolhidos do INPI.

Como a análise final dos dados ia ser feita no Google Colab e para trabalhar com a biblioteca pandas existe uma função que permite ler ficheiros CSV, optou-se para esta etapa final da dissertação não ler diretamente os dados da base de dados PostgreSQL, mas sim aproveitar a funcionalidade de descarregamento destes dados em ficheiros CSV da aplicação pgAdmin, pelo que se optou para esta dissertação não complicar as bases de dados com chaves estrangeiras que ligam as bases de dados e conferem uma relação entre tabelas muito útil quando utilizadas as capacidades de leitura da linguagem SQL, através de *selects* filtrados com estas chaves, mas que nos ficheiros e ferramentas utilizados não teriam tanta leitura e assim optou-se por simplificar, não sendo a melhor solução do ponto de vista técnico, mas sendo a mais eficaz e simples para as ferramentas adotadas.

Nas figuras 3.2 e 3.3, estão exemplos do conteúdo das duas tabelas mencionadas, que serão explorados no próximo capítulo na parte de descrição dos dados, para já fica só uma visão geral dos mesmos e no próximo subcapítulo irá explicar-se melhor os processos de limpeza de dados executados e aí será explicado em mais detalhe as limitações dos

	NIF [PK] character varying	Nome character varying	Tipo Empresa character varying	CAEs character varying[]	Patentes integer	Contratos character varying[]	Anos Contratos integer[]	Anos Patentes integer[]
1	503767549	Instituto Politécnico de...	adjudicante	{""}	3	{https://www.base.gov.p...	1,2014,2019,2022,2023}	{2012,2013,2019}
2	506361438	Instituto Português de ...	adjudicante	{" 86100 - Estabelecime...	3	{https://www.base.gov.p...	9,2019,2020,2022,2023}	{2005,2013,2016}
3	506361616	Instituto Português de ...	adjudicante	{" 86100 - Estabelecime...	1	{https://www.base.gov.p...	{2022,2023}	{2007}
4	503494933	Instituto Politécnico do...	adjudicatárias	{""}	4	{https://www.base.gov.p...	0,2010,2011,2022,2023}	{2013,2015,2015,2021}
5	600006662	Polícia de Segurança P...	adjudicante	{""}	1	{https://www.base.gov.p...	5,2015,2021,2021,2022}	{2016}
6	507487648	ITeCons - Instituto de I...	adjudicatárias	{" 72190 - Outra investig...	1	{https://www.base.gov.p...	6,2016,2021,2022,2022}	{2018}
7	505141019	VORTAL, COMÉRCIO E...	adjudicatárias	{" 62090 - Outras activid...	2	{https://www.base.gov.p...	{2009,2020,2022}	{2015,2015}
8	600012662	Ministério da Defesa N...	adjudicante	{""}	1	{https://www.base.gov.p...	9,2019,2020,2020,2022}	{2015}
9	502286326	Faculdadde de Medicin...	adjudicante	{""}	1	{https://www.base.gov.p...	{2012,2022,2022}	{2019}

Tabela 3.3: Tabela empresas da base de dados PostgreSQL

ficheiros CSV e como foram superadas.

3.4 Limpeza e estruturação de dados

Os processos de limpeza de dados desta dissertação dividiram-se em dois grandes processos, que foram os automatizados antes da fase de análise de dados e os que foram realizados durante esta fase conforme as necessidades.

Nos processos que foram executados antes da análise de dados importa referir um principal e um secundário que acabou por não ter grande utilidade para os estudos executados nesta dissertação, mas para outras análises aos mesmos dados pode ter alguma expressão. Esse estudo secundário prende-se com uma pequena aplicação *web* que fazia dois processos de limpeza e no final é possível visualizar os dados obtidos em tabelas, posteriormente para os filtrar e editar manualmente, caso necessários, sendo que estas duas últimas funcionalidades não foram desenvolvidas dado que para o âmbito desta dissertação, a aplicação pgAdmin executa o mesmo trabalho, sem necessidade de desenvolvimentos extra. Dentro desses processos de limpeza temos um primeiro que limpa os dados, que por facilidade estão todos no formato de *string* para o formato que seja mais apropriado para a sua estruturação e outro que tem uma limpeza mais específica em que nas colunas onde os valores estão praticamente uniformizados em categorias, apenas difere a forma em que, por vezes, estão escritas e assim numa pequena página pode selecionar-se as categorias que são para unir dentro da base de dados. Estas soluções foram abandonadas porque eram referentes a dados que não foram utilizados nas análises finais abordadas no próximo capítulo, pelo que não eram ferramentas úteis no âmbito desta dissertação, mas para análises futuras a estes dados podem ser interessantes e, por isso, a menção neste parágrafo. Relativamente aos processos de limpeza de dados principais executados antes da fase de análise de dados importa ainda referir o processo principal que a partir das colunas com as empresas adjudicantes e adjudicatárias da tabela de contratos, localiza os nomes e os NIFs das empresas presentes em cada campo das colunas e os armazena na base de dados das empresas, caso o NIF ainda não exista na base de dados, senão continua para a próxima iteração. Para este algoritmo o grande segredo é a utilização de expressões regulares para localizar a informação pretendida. Este processo pegou nos 8240 contratos de inovação presentes

na tabela dos contratos e originou 3591 empresas.

Todos os restantes processos de limpeza de dados foram executados conforme as necessidades no Google Colab. Um processo muito importante foi o de "tradução" das *strings* que representavam vetores no ficheiro CSV das empresas para efetivos vetores utilizáveis no Colab. Isto porque o descarregamento deste ficheiro no pgAdmin originava esta limitação, mas esta foi facilmente superada com o auxílio de uma biblioteca de Python de nome "ast", que faculta um método exatamente para este efeito. Foram ainda executadas limpezas condicionais, tendo em conta os dados que queríamos obter para uma determinada análise, por exemplo, no caso de análises relacionadas com patentes era necessário condicionar os dados a dados que tenham pelo menos uma patente ou se queremos fazer análises por CAE, era necessário que esse valor não fosse nulo ou inexistente. Foram ainda efetuadas estruturações de dados para formar vetores, depois fornecidos às funções de formulação de gráficos, para assim se puder graficamente analisar os dados, de uma forma mais intuitiva. Após estes gráficos serem visualizados, por vezes, foi ainda necessário efetuar limpeza dos *outliers*, sendo apenas necessário para isto definir limites máximos e mínimos e voltar a formar gráficos com os novos dados obtidos.

3.5 Sumário do capítulo

Neste capítulo explorou-se todas as questões relacionadas com a extração, armazenamento, estruturação e limpeza de dados. Foi abordado em detalhe as várias fontes de dados utilizadas, os fins destas e o tipo de dados que foram recolhidos e como foi executada essa extração. Em geral, foram utilizadas técnicas de *webscraping* presentes na livraria Python Pyppeteer, tendo especial atenção às especificidades de cada extração no controlo de erros e na forma como se resolve a dificuldade inerente à questão dos *websites* dinâmicos. Em seguida abordou-se a questão do armazenamento de dados e deu-se uma visão geral de como eles foram estruturados em tabelas de PostgreSQL. Para terminar, abordaram-se as questões de limpeza e estruturação dos dados, não só os processos executados na fase de extração dos dados, mas também aqueles executados na fase de análise. Na prática, foram referidas todas as técnicas relacionadas com obtenção e manipulação dos dados, nas diferentes fases da dissertação, sem nunca chegar à parte de análise dos mesmos que será mencionada no próximo capítulo.

4

Análise Empírica

Neste capítulo está presente tudo o que diz respeito à análise empírica. Começa-se por descrever em detalhe os dados utilizados para esta análise, depois referem-se as metodologias utilizadas para os analisar e termina-se apresentando os resultados dessas análise e algumas reflexões, hipóteses e conclusões da mesma.

4.1 Descrição dos dados

Para descrever os dados é importante em primeiro lugar dar uma visão geral da forma como a base de dados se encontra estruturada e, em seguida, descrever um pouco cada campo de cada entidade, neste caso, das entidades contratos e empresas. Na tabela 4.1 encontra-se essa visão geral da base de dados, em termos técnicos, o que se chama diagrama de entidades da mesma, onde se encontram todos os campos de cada entidade. Dado que todos os campos da entidade contratos são do tipo varchar, optou-se por não colocar na tabela e apenas mencionar aqui por escrito, de notar que cada linha desta tabela representa um contrato público de inovação, sendo a sua chave primária o URL onde se pode encontrar o mesmo no portal BASE. Por outro lado, a entidade empresas, como vai ser substancialmente mais utilizada e não vindo do portal BASE, a sua extração provinha de fontes com os dados mais normalizados e estruturados, optou-se por aqui ter bastante mais cuidado na escolha dos tipos de dados atribuídos a cada campo desta entidade. Assim sendo, a chave primária da entidade empresas é o NIF, que se optou por ser do tipo varchar, por se tratar de um campo adquirido mediante extração e limpeza de dados provenientes da entidade contratos, ou seja, oriundos do portal BASE, pelo que, por vezes, por diversas questões, sendo uma das mais comuns o aparecimento de um NIF estrangeiro, estes dados não se encontravam normalizados, ou seja, continham caracteres diferentes dos caracteres numéricos e/ou tinham mais que 9 caracteres, portanto ao contrário do que seria de esperar de um NIF comum. O campo Tipo de Empresa também é varchar por o seu valor apenas poder ser uma de duas opções deste tipo de dados, "adjudicante" ou "adjudicatário". Temos ainda um vetor com dados do tipo varchar para os CAEs, devido a esta informação ter, não só caracteres numéricos, mas também não

numéricos. O campo Patentes representa o número de patentes registadas pela empresa pelo que lhe foi atribuído o tipo inteiro. O campo Contratos é um vetor varchar por se tratar de um vetor com todos os URLs dos contratos públicos de inovação da empresa. Por fim, os campos Anos Contratos e Anos Patentes são os dois vetores de inteiros, pois são os dois vetores de anos em que a empresa, respetivamente, obteve contratos públicos de inovação e registou patentes. Caso a empresa tenha obtido mais que um contrato ou registado mais que uma patente no mesmo ano, esse ano aparecerá no vetor repetido, as vezes respetivas ao número de vezes que a empresa obteve contratos e/ou registou patentes nesse ano.

Para terminar esta visão geral dos dados, temos ainda os dados provenientes da base de dados *Orbis*. Estes não se encontram em base de dados, apenas num ficheiro CSV, onde os dados relevantes para esta dissertação, são apenas o NIF, proveniente da entidade empresas da base de dados, o número de colaboradores das empresas e a receita bruta das mesmas. Estes dados encontram-se normalizados, mas como são oriundos deste ficheiro CSV, mencionado anteriormente, quando é feita a sua leitura são lidos números no formato de *string*, pelo que é necessário convertê-los para inteiro, mas esta ação é muito simples e direta em Python, pelo que apenas confere mais um passo, mas não muito mais dificuldade.

Após esta visão geral dos dados, foi executada uma descrição mais específica dos mesmos, usando alguma estatística descritiva básica, que se passará a apresentar.

Na tabela 4.2 apresenta-se o mínimo, máximo e média dos números de colaboradores de todas as empresas presentes no *dataset*. O valor mínimo faz sentido que seja 1, dado que é necessário que exista pelo menos uma pessoa para formar uma empresa e podem existir empresas de apenas um elemento. Contudo, pela grande diferença entre a média e o valor máximo, mais de 12.000, muito superior à diferença entre a média e o valor mínimo, 146,8, fica a sensação de que existem muito mais empresas com poucos colaboradores do que com muitos colaboradores. Esta hipótese confirma-se pela moda e pela média, sendo que são respetivamente 1 e 6 e pela análise dos dados que nos mostra que 20% das empresas do *dataset* têm apenas 1 colaborador, 58% tem menos de 10 colaboradores, 23% tem 10 ou mais colaboradores e menos de 50 (ou seja, 81% das empresas têm menos de 50 colaboradores), 11% tem 50 ou mais colaboradores e menos de 250 e apenas 8% tem mais de 250 colaboradores (ou seja, apenas 19% das empresas têm mais de 50 trabalhadores no *dataset*). Comparando estes dados com valores nacionais retirados do Pordata [43], esta discrepância faz sentido, dado que se observarmos, por exemplo, os últimos dados fornecidos por esta fonte, os de 2021, as Pequena e Média Empresa (PME)s representam 99,89% do total das empresas nacionais.

Relativamente à receita bruta (tabela 4.3), temos um valor mínimo negativo de uma empresa que apresentou prejuízo, mas analisando os dados pode concluir-se que é a única empresa que apresenta prejuízo no *dataset*, todas as outras apresentam valores positivos (facto que poderá ser importante mais adiante na identificação de *outliers*). Novamente, a diferença entre o máximo e a média é muito maior que a diferença entre o mínimo e a

Contratos
Data da publicação
Tipos de contrato
Nº do acordo quadro
Descrição do acordo quadro
Medida especial
Tipo de procedimento
Descrição
Fundamentação
Fundamentação para recurso ao ajuste direto (se aplicável)
Regime
Critérios materiais
Entidades adjudicantes
Entidades adjudicatárias
Objeto do contrato
Procedimento centralizado
CPVs
Data do contrato
Preço contratual
Prazo de execução
Local de execução
Entidades concorrentes
Anúncio
Peças do procedimento
Modificações contratuais
Documentos
Observações
Critérios ambientais
Justificação para não redução a escrito do contrato
Aviso
Causa da extinção do contrato
Data do fecho do contrato
Preço total efetivo
Causas das alterações ao prazo
Causas das alterações ao preço
URL (chave primária)

Empresas	
NIF (chave primária)	varchar
Nome	varchar
Tipo Empresa	varchar
CAEs	varchar[]
Patentes	integer
Contratos	varchar[]
Anos Contratos	integer[]
Anos Patentes	integer[]

Tabela 4.1: Diagrama de entidades da base de dados

Nº de colaboradores	
Mínimo	1
Média	147,8
Máximo	12.665
Moda	1
Mediana	6

Tabela 4.2: Estatística descritiva do número de colaboradores das empresas presentes no *dataset*

Receita bruta (€)	
Mínimo	-4.272.448
Média	14.276.769,7
Máximo	2.471.927.000
Mediana	339.382

Tabela 4.3: Estatística descritiva da receita bruta das empresas presentes no *dataset*

média, o que faz sentido tendo em conta o valor da mediana também ele muito inferior ao valor da média, pelo que podemos concluir que cerca de 50% das empresas apresentam lucros inferiores a 339.382€, pelo que a média é "inflacionada" pelos poucos valores mais altos que a média de receita bruta, apenas 11% das empresas.

Pela estatística descritiva dos anos dos contratos presentes no *dataset* (tabela 4.4), pode observar-se que são contratos que foram celebrados entre os anos de 2008 e 2023 e que a média e a mediana estão bastante próximas, o que pode indicar uma boa distribuição entre o número de contratos, o que pode ser, mais uma vez, comprovado pela moda que indica que o ano de 2021 é o ano com mais contratos e representa apenas 11% dos contratos do *dataset*, em conjunto com a observação do gráfico de barras da figura 4.1, em que se pode verificar também esta distribuição. Além disto, pode também verificar-se pela observação do gráfico que tanto o primeiro, como o último ano apresentado, 2023 e 2008, são anos com poucos contratos, o que se pode dever, no caso de 2023, por os dados terem sido recolhidos sensivelmente a meio deste ano, pelo que o registo de contratos celebrados em 2023 no portal BASE faz sentido ser de menor volume. Em 2008, por ser o primeiro ano, também faz sentido ter um volume baixo, dado que a transição para registar estes contratos digitalmente no portal BASE terá sido faseada e é lógico o primeiro ano ser aquele em que estes são menos registados. Além disto, pode ainda observar-se que o período entre 2022 e 2016 é aquele com mais contratos celebrados.

Anos dos contratos	
Mínimo	2008
Média	2017,1
Máximo	2023
Moda	2021
Mediana	2017

Tabela 4.4: Estatística descritiva dos anos dos contratos presentes no *dataset*

Por observação da tabela 4.5 constata-se que os valores de preço contratual dos contratos presentes no *dataset*, vão de 0 a mais de 9,7 milhões de euros, sendo que tanto a média e a mediana se encontram na casa dos milhares respetivamente, cerca de 31.000€ e 16.000€. Tratando-se de contratos públicos de inovação, valores como 0 ou inferiores a algumas centenas de euros, são valores a ter em atenção, pois podem não só ser *outliers*, como tratar-se de possíveis gralhas. Examinando estes valores pode-se inferir que apenas 20 contratos, cerca de 0,2% do total de contratos tem o valor mínimo zero e apenas 101

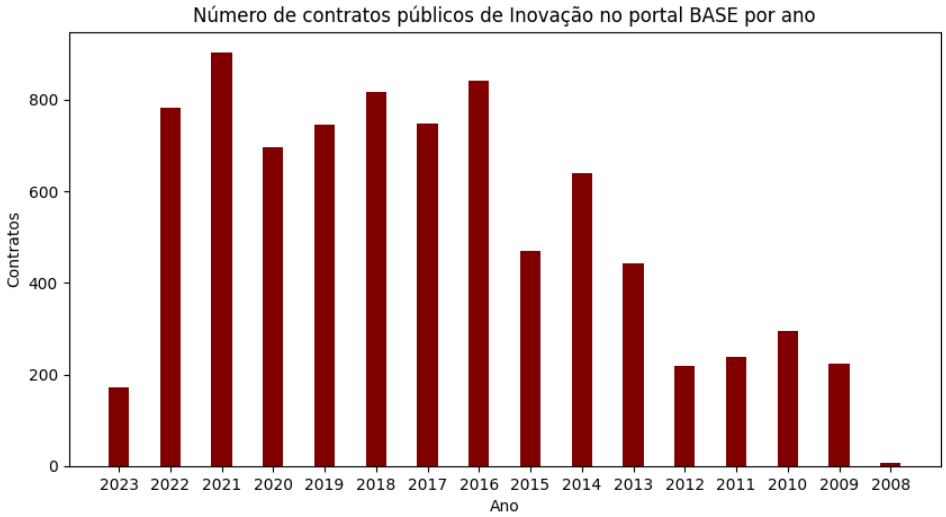


Figura 4.1: Gráfico de barras com o número de contratos por ano

Preço contratual (€)	
Mínimo	0
Média	31.094,1
Máximo	9.779.231,3
Mediana	16.888,34

Tabela 4.5: Estatística descritiva do preço contratual dos contratos presentes no *dataset*

contratos, cerca de 1% do *dataset* apresentam valores inferiores a 150€, pelo que na análise feita no próximo subcapítulo se terá bastante atenção a estes valores. Dado que a diferença entre a média e a mediana ainda é na ordem das dezenas de milhares (mais de 14 mil), mas os valores máximos são nas ordens dos milhões, fez-se um estudo para entender a quantidade de contratos que apresentavam valores superiores e inferiores à média, sendo que 73% se encontram a baixo da média e 27% acima, pelo que se conclui que a diferença entre a média e a mediana é bastante relevante e são realizados substancialmente mais contratos de valor abaixo da média, do que acima e que estes últimos fazem aumentar o valor da média.

No que diz respeito aos CPVs dos contratos, pode observar-se na figura 4.2 uma predominância dos contratos com o CPV 73000000-2, representando 50% do *dataset*, que olhando para a tabela 3.1 representa "Serviços de investigação e desenvolvimento e serviços de consultoria conexos". Os dois outros CPVs com alguma expressão são o 73220000-0 e o 73200000-4, representando respetivamente "Serviços de consultoria em matéria de desenvolvimento" e "Serviços de consultoria em matéria de investigação e desenvolvimento" e respetivamente constituíndo 15% e 12% do total de contratos. Todos os restantes CPVs representam individualmente menos de 5% do total de contratos e em conjunto apenas representam 21% do mesmo, pelo que se optou por ocultar do gráfico circular. Olhando para as três descrições dos CPVs com maior predominância, podemos observar a frequência da ligação com serviços de consultoria relacionados com

investigação e desenvolvimento.

Figura 4.2: Número de contratos por CPV

A distribuição de CAEs nas empresas do *dataset* é bastante repartida, não havendo uma grande predominância de nenhum CAE, como pode ser observado na figura 4.3, em que o CAE presente em mais empresas, o 70220, representa 11% do total de empresas e o segundo mais representado, o 74900, representa apenas 6%. Todos os restantes CAEs representam 83% do *dataset* e nenhum individualmente representa mais que 5%. Novamente, a palavra consultoria a aparecer nos dois CAEs mais predominantes, em linha daquilo que já tinha sido observado nos CPVs.

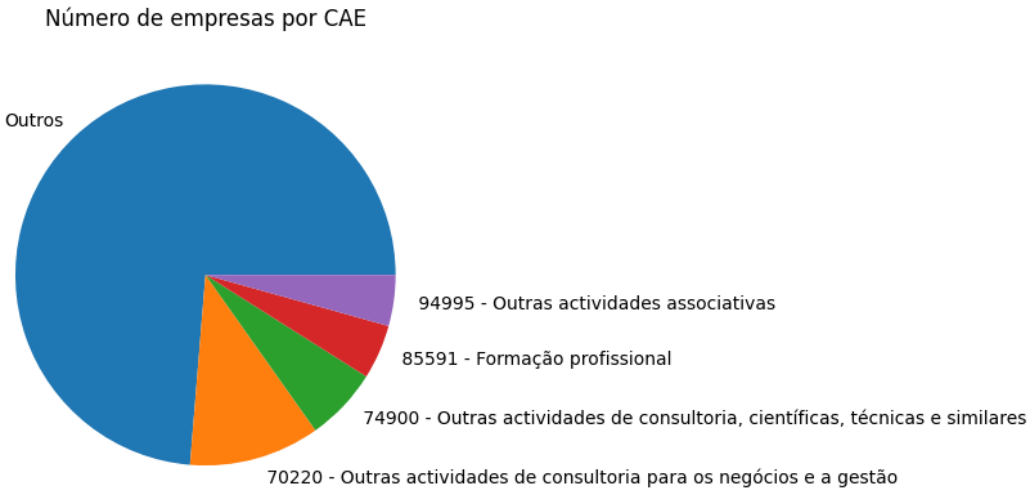


Figura 4.3: Número de empresas por CAE

O número de patentes registadas é um dado muito peculiar e diferente de todos os outros, dado que apenas 3% das empresas do *dataset* têm registos de patentes na fonte consultada, o que ainda são 111 empresas e 2056 patentes registadas, mas faz com

Número de patentes por empresa	
Mínimo	0
Média	0,6
Máximo	405
Moda	0
Mediana	0

Tabela 4.6: Estatística descritiva do número de patentes por empresa presente no *dataset*

Número de patentes por empresa com patentes registadas	
Mínimo	1
Média	18,5
Máximo	405
Moda	1
Mediana	3

Tabela 4.7: Estatística descritiva do número de patentes registadas por empresa com patentes registadas presentes no *dataset*

que os valores da primeira estatística descritiva presente na tabela 4.6, não sejam tão relevantes para esta dissertação, como aqueles que se encontram na tabela 4.7, contudo apresentam-se as duas tabelas, para que se possa fazer uma comparação entre as mesmas. Pode observar-se que os valores médios diferem bastante dado que se retira os zeros de uma para outra e assim sendo o valor mínimo também passa de 0 para 1, tal como a moda que também é transferida do valor 0 para 1 de uma tabela para outra, sendo que este último número de patentes, representa 39% do número de patentes registadas pelas empresas do *dataset*, o que faz bastante sentido, tendo em conta que o valor máximo de patentes registadas por uma só empresa é 405 e a mediana e a média, são respetivamente 3 e 18,5, valores bastantes mais baixos que o máximo e muito mais próximos do valor mínimo. Os valores inferiores à média apresentada representam 86% das empresas do *dataset* e apenas 14% têm um número de patentes registadas acima deste valor, dentro das empresas que registaram pelo menos uma patente.

Para terminar, falta ainda apresentar a estatística descritiva dos anos de registo de patentes pelas empresas presentes no *dataset* (tabela 4.8). Estes anos estão entre 1979 e 2021, sendo a moda igual a 2008, ou seja, o ano com mais patentes registadas pelas empresas presentes no *dataset* é este ano e a diferença entre a média, 2011,6, e a mediana 2012, é bastante pequena (0,4), pelo que indica que é possível que exista uma distribuição homogénea no número de patentes registadas em cada ano, hipótese que ganha ainda mais força tendo em conta que as patentes registadas de 2012 em diante representam 51% do *dataset* e todas as outras, os restante 49%, não sendo 50% em ambas é muito próximo, pelo que estes dois factos juntos parecem corroborar com a hipótese apresentada.

Anos de registo de patentes	
Mínimo	1979
Média	2011,6
Máximo	2021
Moda	2008
Mediana	2012

Tabela 4.8: Estatística descritiva dos anos de registo de patentes pelas empresas presentes no *dataset*

4.2 Metodologia

Em relação às metodologias utilizadas, começou-se por utilizar alguma estatística descritiva básica para descrever e entender os dados com que se iria trabalhar nesta dissertação, utilizando valores máximos, mínimos e médios e também medianas e modas (a moda foi usada apenas nos dados em que faria sentido empregá-la). Para esta análise inicial foi ainda feito recurso de gráficos circulares e de barras, para entender a distribuição dos dados, quando isso fosse relevante e um acréscimo de qualidade para a análise. Todas estas metodologias estão presentes no subcapítulo anterior onde se abordou com mais detalhe a descrição dos dados.

No subcapítulo seguinte irá ser abordado os resultados empíricos obtidos recorrendo a gráficos de dispersão aos quais se calcularam regressões lineares apropriadas aos seus valores e correlações entre os seus eixos. Foram ainda obtidos gráficos de barras para entender as frequências de valores em que esta análise era de extrema importância. Por fim, foram ainda calculados os coeficientes de correlação de Pearson e Spearman para aquelas variáveis onde se pretendia acrescentar alguma robustez aos coeficientes resultantes dos cálculos.

4.3 Resultados empíricos

A análise do *dataset* iniciou-se por duas variáveis muito importantes e que à partida logicamente estariam relacionadas, o número de contratos de inovação obtidos por cada empresa e o valor em euros da soma dos financiamentos que cada empresa arrecadou com os contratos que obteve. À primeira vista, faria sentido que quantos mais contratos a empresa tenha obtido mais financiamento. Para testar esta hipótese utilizou-se novamente estatística descritiva, primeiro olhando para os valores médios, máximos, mínimos, modas e medianas e depois utilizando gráficos de dispersão, regressões lineares e calculando-se o coeficiente de correlação entre estas duas variáveis para ter um valor que represente o que observando os gráficos se poderia presumir.

Na tabela 4.9 encontra-se alguma estatística descritiva relacionada com o financiamento obtido por cada empresa em contratos públicos de inovação. Nesta tabela pode observar-se que o valor mínimo de financiamento é zero, contudo, tal como foi

referido em relação à tabela 4.5, da estatística descritiva do preço contratual, é algo estranho este valor, o que poderá indicar que este é um *outlier*, contudo não se irá fazer essa classificação neste caso, dado que observando o gráfico da figura 4.4, estes valores não apresentam características para fazer esta classificação. Em relação ao valor máximo e à média encontram-se com uma diferença significativa entre ambos, sendo o máximo na ordem das dezenas de milhões e a média na centena de milhares. Além dessa diferença, a que existe entre a média e a mediana, também é significativa, o que pode significar que existem *outliers* nos valores mais altos de financiamento e, olhando para o gráfico da figura 4.4, esta hipótese parece confirmar-se, dado que os dois valores com o financiamento mais alto do gráfico têm características de *outliers*.

No que diz respeito à tabela 4.10, o valor mínimo apresentado para o número de contratos de inovação obtidos por cada empresa é 1, que é também o valor da moda e da mediana, o que dá sentido à média de 3,95, quando o valor máximo é de 237 contratos obtidos. Além disto é também interessante referir que as empresas com apenas 1 contrato obtido representam 54% do total de empresas presentes no *dataset*, contudo grande parte dos valores têm uma relação uniforme com o financiamento, apresentando valores mais baixo de financiamento, assim sendo não serão considerados *outliers*. Observando mais uma vez o gráfico da figura 4.4, a tese de que existem *outliers* próximos do valor máximo, como se observou para o financiamento, é novamente corroborada, neste caso, pelo valor máximo do número de contratos (237) que é claramente um *outlier* e o único ponto que será removido devido a esta variável.

Para pôr em prática a remoção de *outliers* mencionada nos dois parágrafos anteriores, serão removidos todos os pontos que tenham financiamento maior que 0,5 milhões de euros e número de contratos maior que 236, que na prática resultará na remoção dos três pontos referidos como *outliers* anteriormente, o que resultou no gráfico apresentado na figura 4.5. Removendo os *outliers*, o coeficiente de correlação entre as variáveis número de contratos obtidos e financiamento total obtido passa de 59% para 74%. Dois valores que indicam uma boa relação de correlação entre estas duas variáveis.

Nos gráficos mencionados nos parágrafos anteriores agrupou-se os dados por empresa, ou seja, cada ponto representa no eixo do x o número de contratos de inovação obtidos por uma determinada empresa e no eixo do y o valor total de financiamento obtido somando o valor de todos os contratos obtidos. Agrupando os dados por CAEs (gráfico da figura 4.6), ou seja, na prática, agrupando os valores das empresas do mesmo setor de atividade, mantêm-se valores altos de correlação entre as duas variáveis em estudo. Neste caso, coeficiente de correlação até aumenta para 97%, um ótimo indicador da relação entre estas variáveis. Pela observação dos três gráficos mencionados podemos concluir que, por norma, as empresas que recebem mais contratos públicos de inovação, são também aquelas que recebem maior financiamento através destes.

Financiamento (€)	
Mínimo	0
Média	136.216
Máximo	21.280.842
Mediana	35.529,15

Tabela 4.9: Estatística descritiva do financiamento obtido por empresa em contratos público de inovação

Número de contratos	
Mínimo	1
Média	3,95
Máximo	237
Moda	1
Mediana	1

Tabela 4.10: Estatística descritiva do número de contratos público de inovação obtidos por empresa

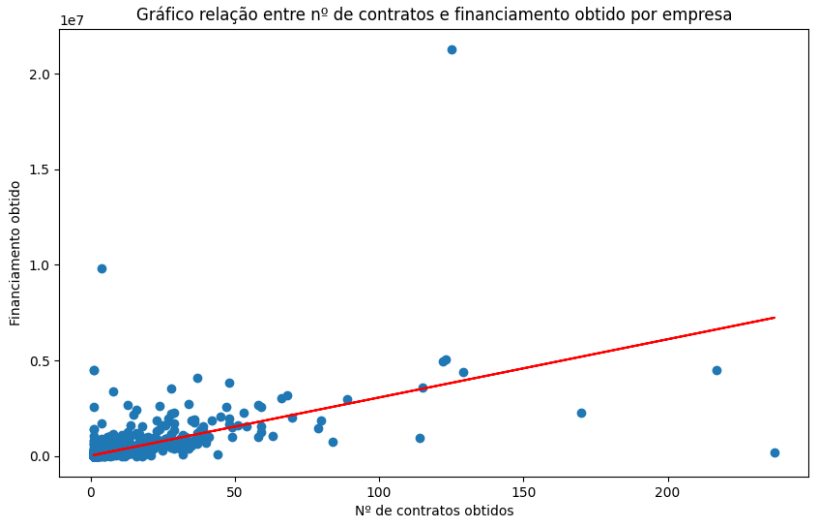


Figura 4.4: Gráfico de dispersão que relaciona o número de contratos públicos de inovação obtidos por cada empresa presente no *dataset* com a soma do financiamento obtido em todos esses contratos

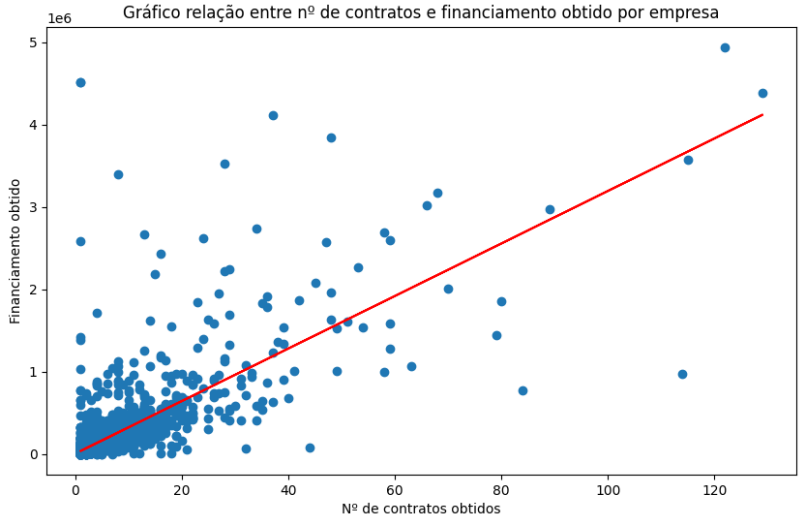


Figura 4.5: Gráfico de dispersão que relaciona o número de contratos públicos de inovação obtidos por cada empresa presente no *dataset* com a soma do financiamento obtido em todos esses contratos, removendo os *outliers*

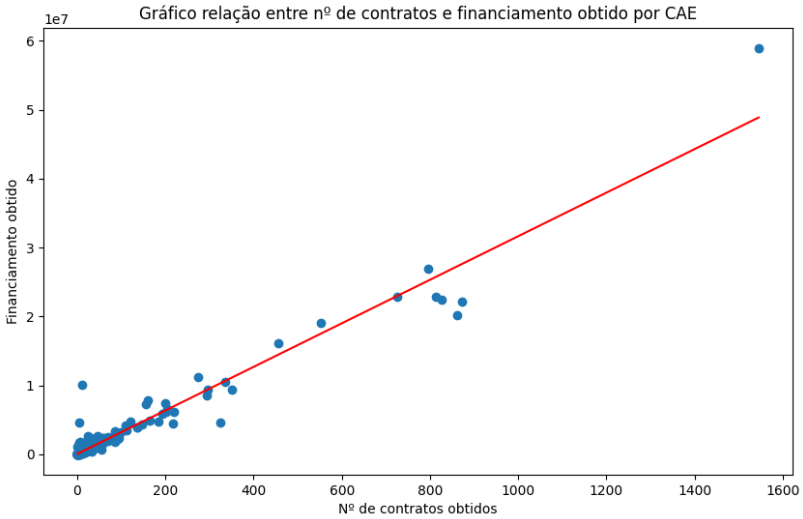
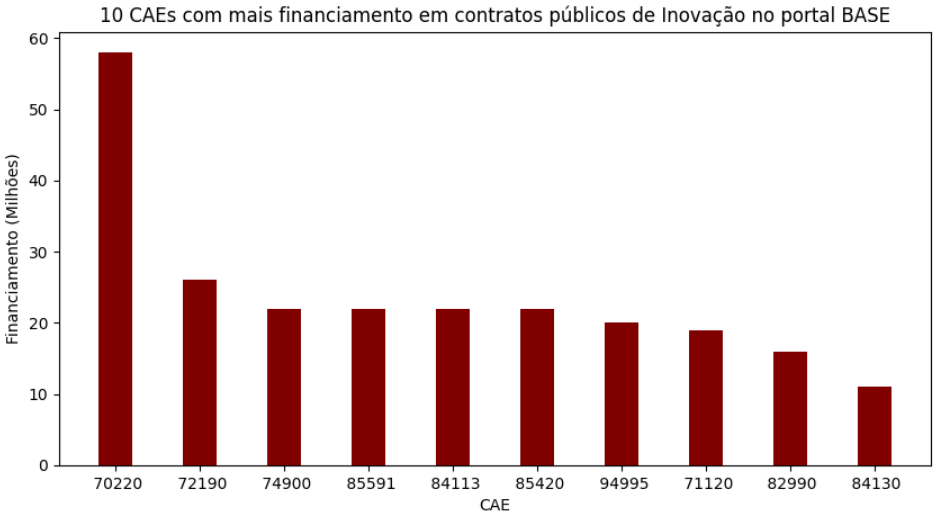


Figura 4.6: Gráfico de dispersão que relaciona o número de contratos públicos de inovação obtidos em cada CAE com a soma do financiamento obtido em todos esses contratos

Em seguida, estudou-se os CAEs que obtêm mais e menos financiamento e os resultados foram, respetivamente, os gráficos de barras das figuras 4.7 e 4.8. Os CAEs com menos financiamento obtido são, em sua grande maioria, relacionados com atividades de produção, excetuando apenas dois dos dez CAEs, o 80100 e o 82922. Olhando para as áreas de atividade subjacentes a estes CAEs também parece intuitivo investir-se menos nestas em contratos relacionados com inovação, tendo em conta que não são áreas que sejam associadas a inovação de forma direta, contudo é interessante encontrar nesta lista CAEs que pertencem a setores de atividades que até têm alguma representatividade no Valor Acrescentado Bruto (VAB) da economia portuguesa, como o

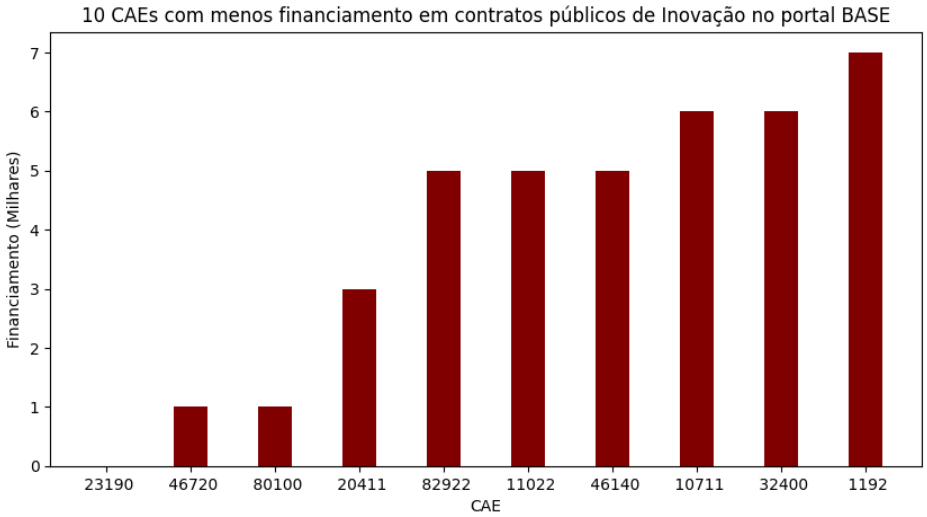
gls*cae 11022, que se inclui no setor das Indústrias alimentares, das bebidas e do tabaco, que em 2011 representou mais de 2% do gls*vab da economia portuguesa [44]. Analisando os CAEs com mais financiamento obtido em contratos públicos de inovação, podemos observar uma tendência para atividades relacionadas com consultoria, ciência, investigação e ensino, excetuando três CAEs, o 84113, 94995 e o 84130, que podem ter este tipo de atividades, mas não se associam a elas de forma direta. Além disto, é de assinalar a diferença de quase o dobro dos contratos atribuídos ao CAE com mais contratos públicos de inovação, o 70220 e o segundo CAE com mais contratos, o 72190 e olhando para o que eles representam é de estranhar que não seja o contrário, dado que o primeiro se relaciona com atividades de consultoria para negócios e gestão e o segundo para atividades de investigação e desenvolvimento das ciências físicas e naturais, pelo que se analisou os CPVs dos contratos que foram atribuídos a cada um destes para se analisar para que tipo de atividades mais em concreto as empresas deste ramo de atividade foram contratadas. Na figura 4.9 apresentam-se os resultados para o CAE 70220 e na figura 4.10 os resultados para o CAE 72190. Para o primeiro existe uma predominância de contratos com o CPV 73000000-2 (praticamente 50%), que representam "Serviços de investigação e desenvolvimento de consutoria conexos", seguido do CPV 73220000-0 (cerca de 26% dos contratos) que representa "Serviços de consultoria em matéria de desenvolvimento" e o último CPV significativo é o 73200000-4 (com cerca de 17%), que representa "Serviços de consultoria em matéria de investigação e desenvolvimento", o que faz sentido tendo em conta que o CAE se relaciona com atividades de consultoria e todos os três CPVs principais dos contratos deste também. Olhando para o CAE 72190, tal como o anterior também existe uma predominância de contratos com o CPV 73000000-2, desta feita menor (apenas 35% dos contratos), mas o que faz sentido, tendo em conta que o este é o CPV com maior predominância se olharmos para todos os contratos do *dataset*, como se pode comprovar pelos dados apresentados na figura **??**. O CPV seguinte é o 73111000-3 (com cerca de 18% dos contratos), que representa "Serviços relacionados com laboratórios de investigação" e o último CPV significativo é o 73200000-4 (com cerca de 17%), novamente relacionado com consultoria e o terceiro CPV mais representado nos contratos do *dataset* no seu todo, pelo que mais uma vez, faz sentido estar presente, não sendo um dado tão interessante como o CPV anterior. Além destes dados, podemos também observar a percentagem do VAB das atividades onde se inserem os dois CAEs, sendo que o 70220 se insere em "Atividades jurídicas, de contabilidade, gestão, arquitetura, engenharia e atividades de ensaios e análises técnicas", que nos últimos dados que se conseguiu obter do Instituto Nacional de Estatística (INE) de 2011, representava 3,2% do VAB nacional desse ano e a atividade "Investigação científica e desenvolvimento", onde se insere o CAE 72190, apenas 0,4%, pelo que se pode concluir, que neste caso se está a financiar mais a inovação através de contratos públicos de uma área de atividade que tem mais peso no VAB nacional, contudo, por outro lado, pode argumentar-se que se está a financiar menos o setor que em teoria traria mais inovação, mas menos retorno imediato e pode

argumentar-se ainda que existe um grande volume de investimento em inovação através do financiamento em atividades de consultoria, contudo, em teoria, deveria tentar investir-se mais em atividades relacionados de forma mais direta com investigação e desenvolvimento. Por outro lado e por fim, analisou-se a relação entre a percentagem do %VAB de cada setor e o financiamento obtido em contratos públicos de inovação, gráfico que se encontra representado na figura 4.10 e observou-se pouca relação entre estas duas variáveis, como se pode comprovar pelo baixo coeficiente de correlação obtido, coeficiente esse de apenas 15%, o que nos permite concluir que não se está a financiar a inovação através de contratos públicos, tendo em conta a percentagem do VAB do setor da empresa a quem este é entregue, o que se pode argumentar, por um lado, que é positivo para diversificar e não financiar apenas ou tendencialmente as áreas de atividade com mais peso na economia nacional, contudo também se pode argumentar, que se poderia utilizar um pouco mais esta estratégia, dado que a inovação, em teoria traz crescimento económico e faz crescer os mercados em que se insere, pelo que seria bastante positivo trazer inovação a indústrias com um maior peso na economia e que, tendencialmente, tendem a estagnar e não inovar tanto, por estarem mais estabelecidas no mercado.



70220 - Outras actividades de consultoria para os negócios e a gestão
72190 - Outra investigação e desenvolvimento das ciências físicas e naturais
74900 - Outras actividades de consultoria, científicas, técnicas e similares
85591 - Formação profissional
84113 - Administração local
85420 - Ensino superior
94995 - Outras actividades associativas
71120 - Engenharia e técnicas afins
82990 - Outras actividades de serviços de apoio prestados às empresas
84130 - Actividades económicas

Figura 4.7: Gráfico de barras com os 10 CAEs que obtiveram mais financiamento em contratos públicos de inovação



23190 - Fabricação e transformação de outro vidro
46720 - Minérios e metais
80100 - Actividades de segurança privada
20411 - Sabões, detergentes e glicerina
82922 - Outras actividades de embalagem
11022 - Produção de vinhos espumantes e espumosos
46140 - Agentes do comércio por grosso de máquinas, equipamento industrial, embarcações e aeronaves
10711 - Panificação
32400 - Jogos e de brinquedos
1192 - Outras culturas temporárias

Figura 4.8: Gráfico de barras com os 10 CAEs que obtiveram menos financiamento em contratos públicos de inovação

Figura 4.9: Número de contratos por CPV para o CAE 70220

Figura 4.10: Número de contratos por CPV para o CAE 72190

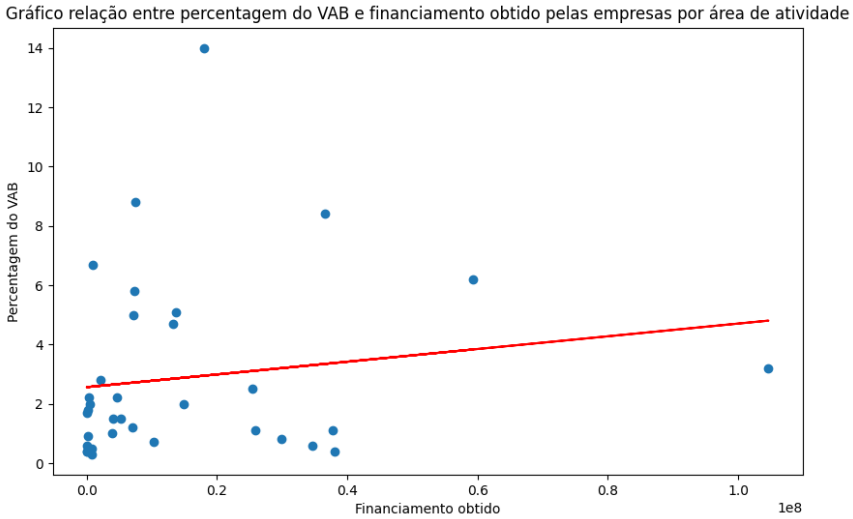


Figura 4.11: Gráfico de dispersão que relaciona a percentagem do CAE do setor de atividade com o financiamento obtido pelo mesmo em contratos públicos de inovação

Estudou-se também a relação entre a faturação da empresa e os financiamentos obtidos. Olhando, em primeira instância, para a estatística descritiva básica da faturação apresentada na tabela 4.11, podemos constatar que o valor mínimo é negativo em mais de 4 milhões, o máximo é positivo em mais de 2 mil milhões e a média ser de mais de 14 milhões e a mediana apenas mais de 339 milhares, indica que o valor da média parece estar "inflacionado" por valores mais próximos do valor máximo, que devem estar em minoria no *dataset* e realmente é o que se verifica, sendo que 89% das empresas apresentam faturação abaixo da média apresentada e apenas as restantes 11% acima da média, pelo que é um sinal de alarme para a possibilidade de alguns destes valores poderem vir a ser *outliers*. Além disso, também se verifica que o único valor negativo de faturação de uma empresa é o valor mínimo do *dataset*, pelo que também se terá de ter

Faturação (€)	
Mínimo	-4.272.448
Média	14.276.769,8
Máximo	2.471.927.000
Mediana	339.382

Tabela 4.11: Estatística descritiva da faturação de cada empresa presente no *dataset*

atenção para a possibilidade de este valor ser um *outlier*. Como já referido anteriormente, na análise que se fez à tabela 4.9, com alguma estatística descritiva básica do financiamento obtido pelas empresas, é também necessário ter em atenção os possíveis *outliers* nos valores mais próximos do valor máximo desta variável.

Após visualização do gráfico que relaciona a faturação das empresas com o financiamento obtido, apresentado na figura 4.12, consegue retirar-se deste que a correlação entre estas duas variáveis, neste caso, é baixa, apenas de cerca de 14%, valor que apenas sobe para 22% quando retirados os *outliers*, gráfico que pode ser observado na figura 4.13. Agrupando as empresas por CAE observa-se um aumento no coeficiente de correlação entre o financiamento e a faturação, como se pode denotar olhando para os gráficos das figuras 4.14 e 4.15, respetivamente com e sem *outliers*, o primeiro com um coeficiente de correlação de cerca de 24% e o segundo com um de cerca de 41%. A questão que se impõe, tendo por base os factos anteriormente referidos, é se a baixa correlação entre estas duas variáveis se mantém como nos primeiros gráficos ou se continua a aumentar se se dividir o *dataset* em empresas com faturação acima da média e empresas abaixo desse valor. Para responder a esta pergunta, dividiu-se o *dataset* como descrito anteriormente e formaram-se os gráficos das figuras 4.16 e 4.17, respetivamente para as empresas com faturação abaixo e acima da média de faturação das empresas presente no *dataset*, sendo que os coeficientes de correlação continuaram baixos, sendo respetivamente de 28% e 25%, já após a remoção dos *outliers*, seguindo assim a tendência dos coeficientes de correlação baixos anteriormente mencionados. Com tudo isto, chega-se à conclusão de que não existe grande correlação entre a faturação das empresas e a maior ou menor obtenção de financiamento através de contratos públicos de inovação, o que até pode ser bom sinal, dado que esta homogeneidade pode querer denotar uma igual distribuição deste tipo de recursos financeiros por parte do estado português sem favorecimento de empresas com mais ou menos faturação na hora de atribuição de contratos públicos e até uma boa decisão económico tendo como principal objetivo a inovação, dado que as maiores disrupções acontecem, por norma, nas empresas mais pequenas, logo, frequentemente, com menos faturação, mas, por vezes, podem também ser interessante atribuir estes fundos tendo como principal foco a inovação, a empresas de maior dimensão, logo, geralmente, com mais faturação, pois estas têm mais recursos e talvez até maior necessidade de se inovar, que empresas de menor dimensão que têm essa urgência de forma mais frequente no seu dia-a-dia por motivos de sobrevivência.

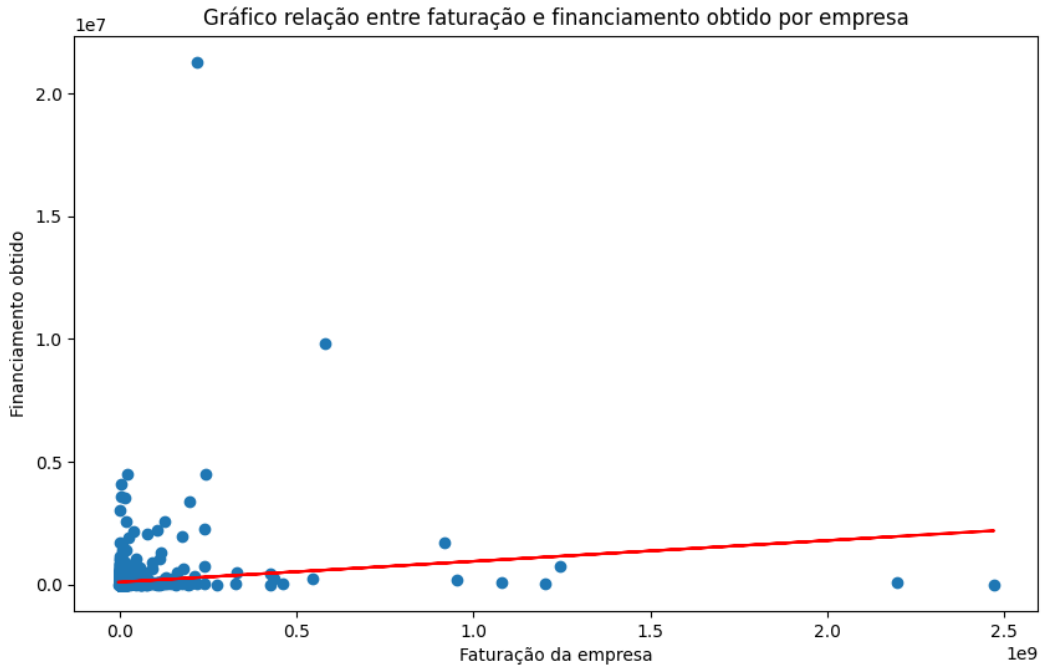


Figura 4.12: Gráfico de dispersão que relaciona a faturação das empresas com a soma do financiamento obtido por estas em contratos públicos de inovação

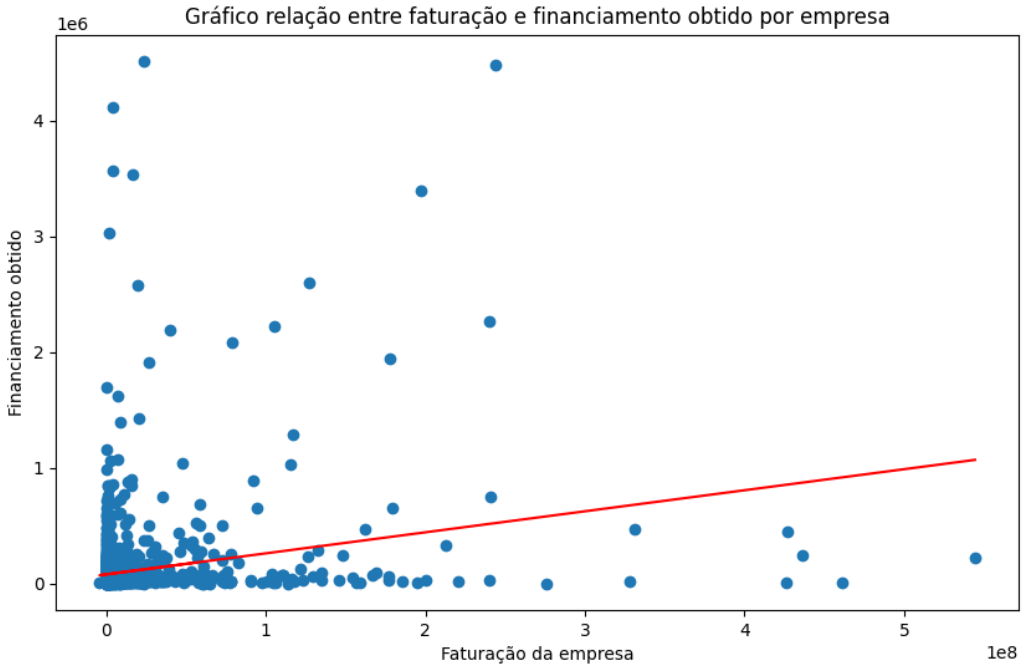


Figura 4.13: Gráfico de dispersão que relaciona a faturação das empresas com a soma do financiamento obtido por estas em contratos públicos de inovação, removendo os *outliers*

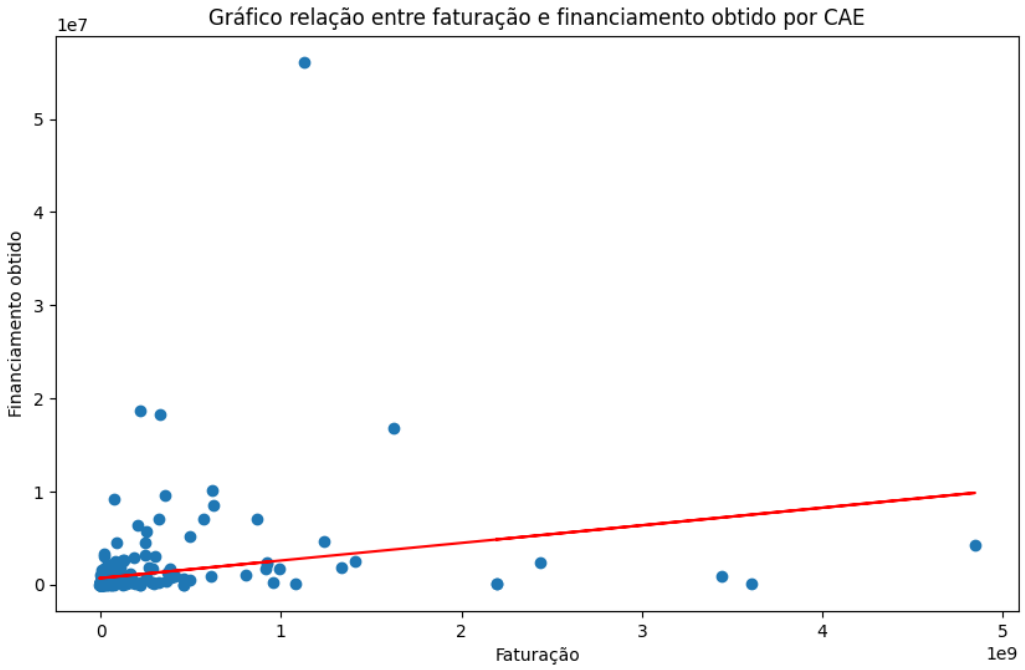


Figura 4.14: Gráfico de dispersão que relaciona a faturação das empresas com a soma do financiamento obtido por estas em contratos públicos de inovação, por CAE

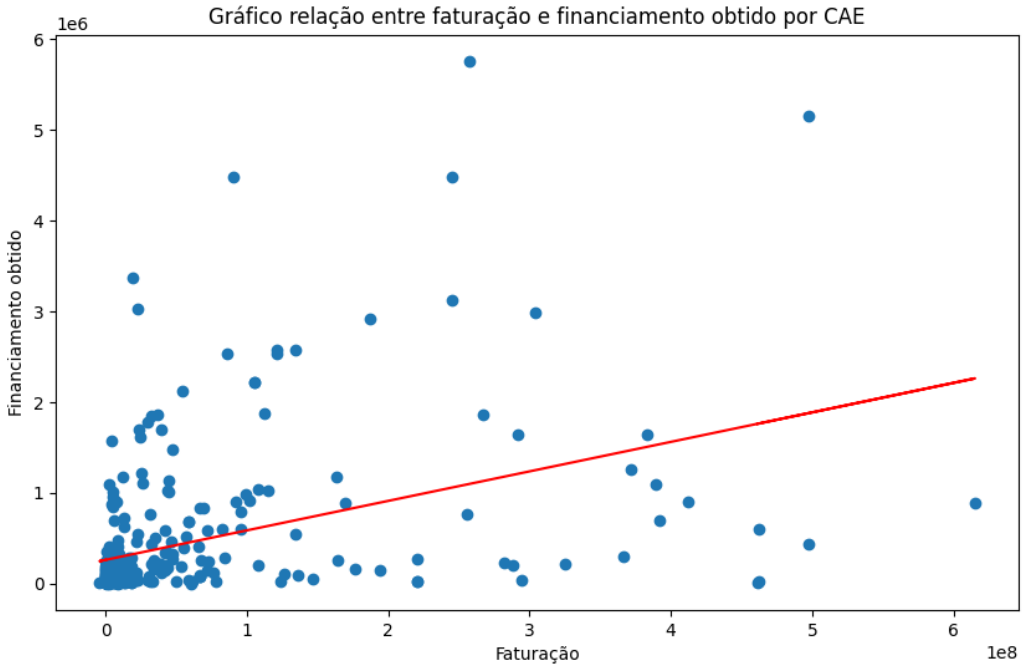


Figura 4.15: Gráfico de dispersão que relaciona a faturação das empresas com a soma do financiamento obtido por estas em contratos públicos de inovação, por CAE, removendo os *outliers*

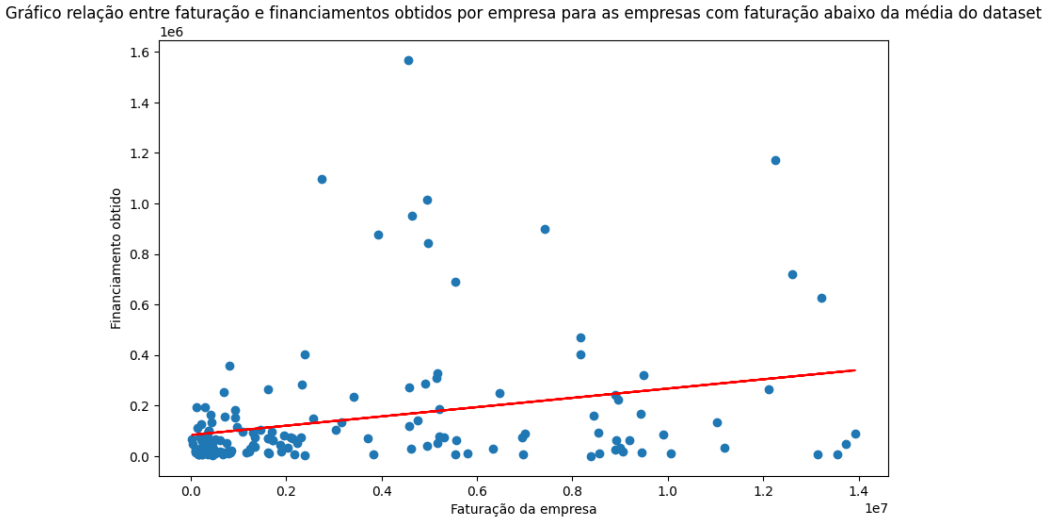


Figura 4.16: Gráfico de dispersão que relaciona a faturação das empresas com a soma do financiamento obtido por estas em contratos públicos de inovação, para empresas com faturação abaixo da média do *dataset* removendo os *outliers*

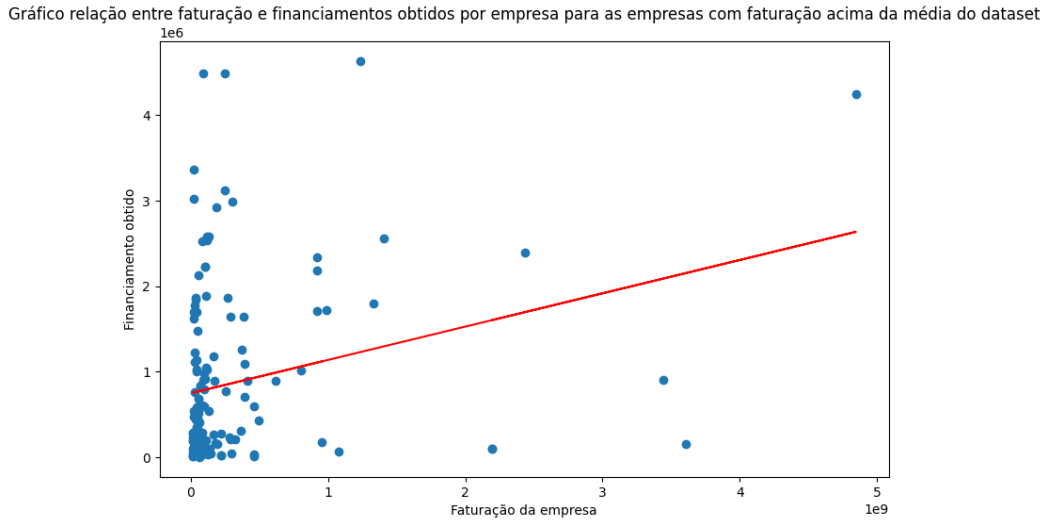


Figura 4.17: Gráfico de dispersão que relaciona a faturação das empresas com a soma do financiamento obtido por estas em contratos públicos de inovação, para empresas com faturação acima da média do *dataset* removendo os *outliers*

Com o objetivo de tentar comprovar as hipóteses anteriormente descritas quando se tentou relacionar a faturação das empresas que obtiveram contratos públicos de inovação, com o financiamento total que obtiveram através deste meio, em que se deu como hipótese que se está a financiar de igual forma tanto grandes empresas, como empresas menor dimensão, olhou-se agora na mesma para a dimensão das empresas, mas agora pelo número de colaboradores das mesmas e voltando a relacionar com o financiamento obtido. Sendo que a estatística descritiva básica das variáveis em questão já foram abordadas anteriormente e encontram-se nas tabelas 4.2 e 4.9, podemos partir

para a análise dos gráficos de dispersão e coeficientes de correlação formulados a partir destas variáveis. Como foi feito para os gráficos que relacionavam a faturação com o financiamento, agora para relacionar o número de colaboradores com o financiamento obtido por cada empresa, também se começou por formar gráficos de dispersão entre estas duas variáveis, primeiro agrupando por empresa e depois por CAE, respetivamente nas figuras 4.18 e 4.19, sendo que na primeira não se apresenta o gráfico sem remoção dos *outliers*, pois esta ação não resultou num melhor coeficiente de correlação entre as duas variáveis em estudo, ao contrário da segunda figura, onde se efetuou esta ação, por melhorar o coeficiente de correlação e ser, por isso, interessante apresentar nesta dissertação este gráfico. Contas feitas, para o gráfico agrupado por empresa obteve-se um coeficiente de correlação de 21% (que baixou para 13% com a tentativa de remoção dos *outliers*) e para aquele que foi agrupado por CAE o coeficiente foi de 31%, já depois de remover os *outliers* (antes desta ação registava o valor de 19%). Todos estes valores apresentam-se baixos, pelo que confirmam a hipótese anterior, de que não existe relação entre a dimensão das empresas e o financiamento obtido em contratos públicos de inovação, o que dá força à tese de que não existe um favorecimento óbvio na obtenção de contratos públicos tendo por base a dimensão da empresa com quem ele é celebrado e que existe diversificação na escolha das empresas com as quais se celebram contratos públicos de inovação em Portugal, tendo por base a dimensão da empresa, seja esta medida através da sua faturação ou do seu número de colaboradores.

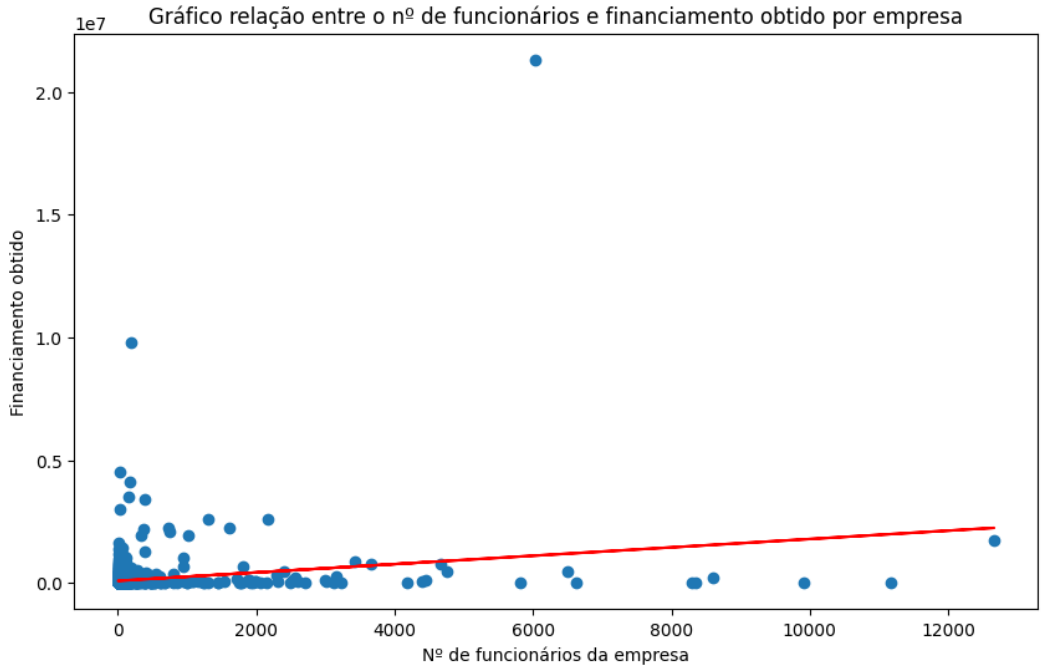


Figura 4.18: Gráfico de dispersão que relaciona o número de colaboradores das empresas com a soma do financiamento obtido por estas em contratos públicos de inovação

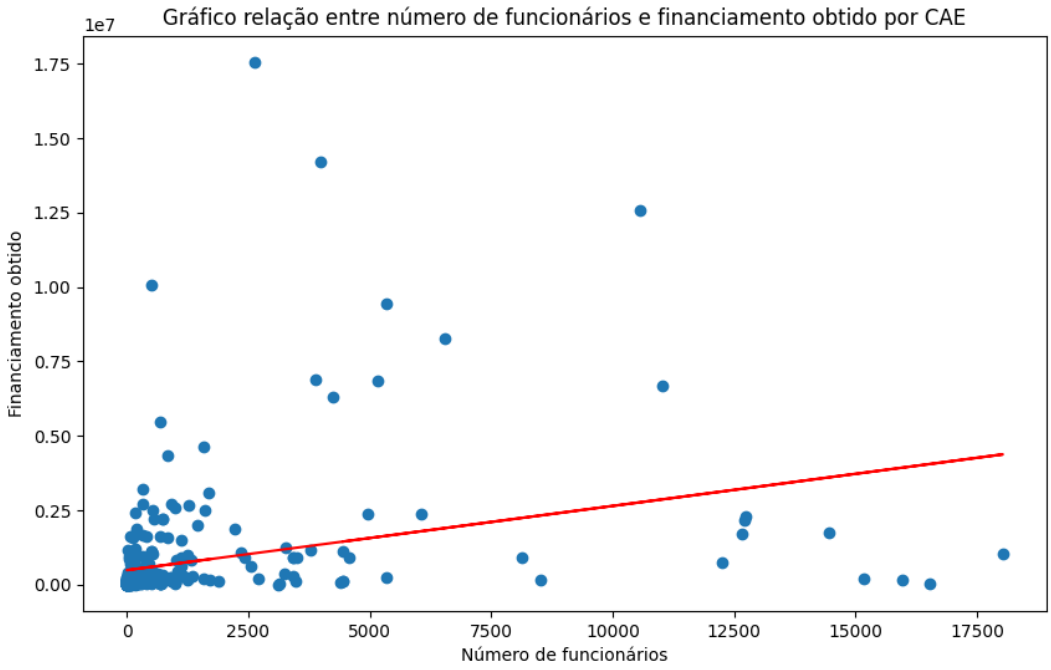


Figura 4.19: Gráfico de dispersão que relaciona o número de colaboradores das empresas com a soma do financiamento obtido por estas em contratos públicos de inovação, por CAE, removendo os *outliers*

Por último, é importante analisar todos os dados relativos a patentes, para isso analisou-se a relação entre as variáveis número de patentes registadas e financiamento total obtido em contratos públicos de inovação, quais os CAEs que registaram mais e menos patentes e os anos em que estas patentes foram registadas. A análise das estatísticas descritivas bases das variáveis número de patentes registadas e anos de registo de patentes já foram feitas anteriormente e estão apresentadas de forma resumida nas tabelas 4.6, 4.7 e 4.8, pelo que se avançará para a análise de gráficos de dispersão e respetivos coeficientes de correlação. Observando o gráfico de dispersão da figura 4.20, que relaciona o número de patentes registadas por cada empresa e o financiamento total obtido pela mesma em contratos públicos de inovação, não se observa uma relação clara entre as variáveis, mesmo quando se remove os *outliers* no gráfico da figura 4.20 e se agrupa as empresas por CAEs na figura 4.20. Analisando os coeficientes de correlação, para as variáveis em estudo quando agrupadas por empresas e sem remoção de *outliers*, obteve-se um coeficiente de correlação de Pearson de 48% e um de Spearman de 46%, valores próximos dos 50% e próximos um do outro. Quando se retira os *outliers*, a diferença entre os coeficientes de correlação de Pearson e Spearman já aumenta, sendo o primeiro de 62% e o segundo de 42%, o que por um lado é positivo para a relação linear entre as variáveis, mas não tanto para a robustez desta subida de coeficiente de correlação, ou seja, dá confiança em afirmar que existe uma correlação maior que 40% e talvez até próxima de 50%, contudo não deverá ser mais alta que isso. Quando se agrupa as empresas por CAE e se obtém um coeficiente de correlação de Pearson de 85% e um

de Spearman de 63%, o que dá ainda mais confiança à afirmação anterior de que existe relação entre as duas variáveis em estudo e que esta deverá ser próxima dos 50%, pelo que deverá ser uma boa relação, contudo não muito elevada.

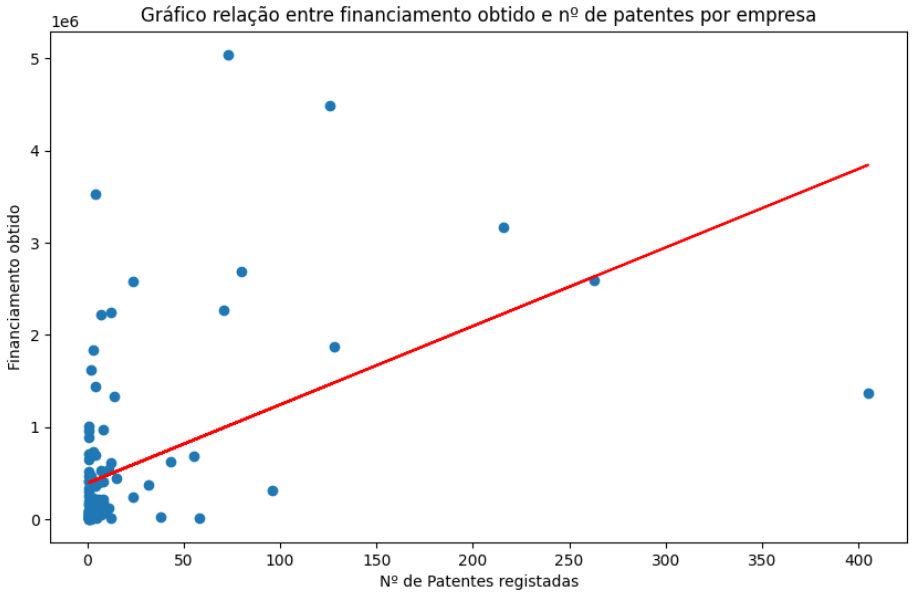


Figura 4.20: Gráfico de dispersão que relaciona o número de patentes registadas pelas empresas com a soma do financiamento obtido por estas em contratos públicos de inovação

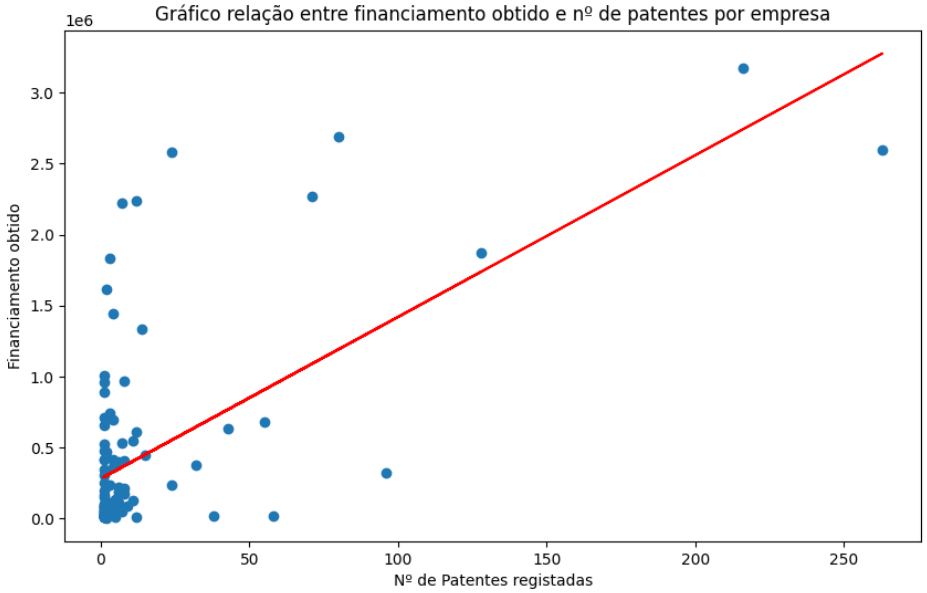


Figura 4.21: Gráfico de dispersão que relaciona o número de patentes registadas pelas empresas com a soma do financiamento obtido por estas em contratos públicos de inovação, removendo os *outliers*

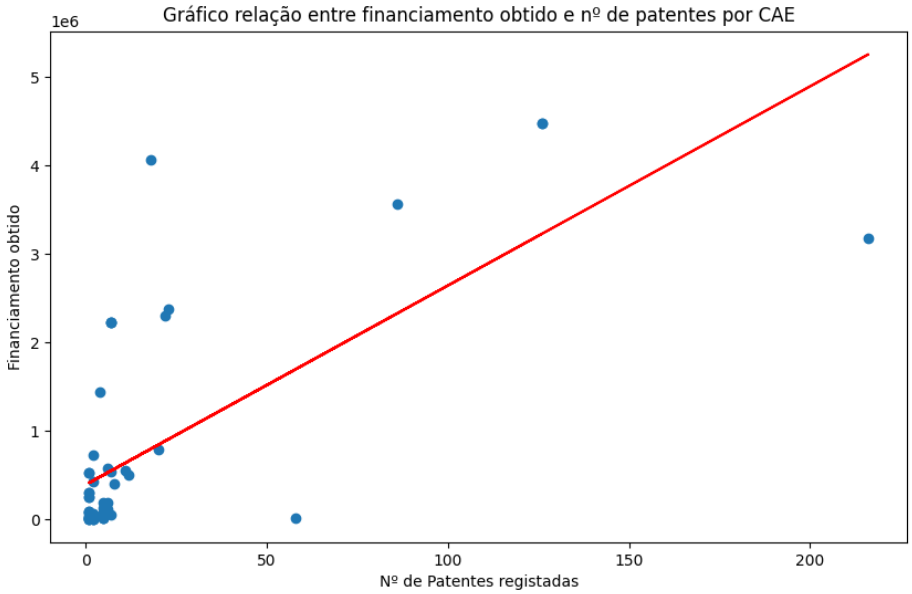
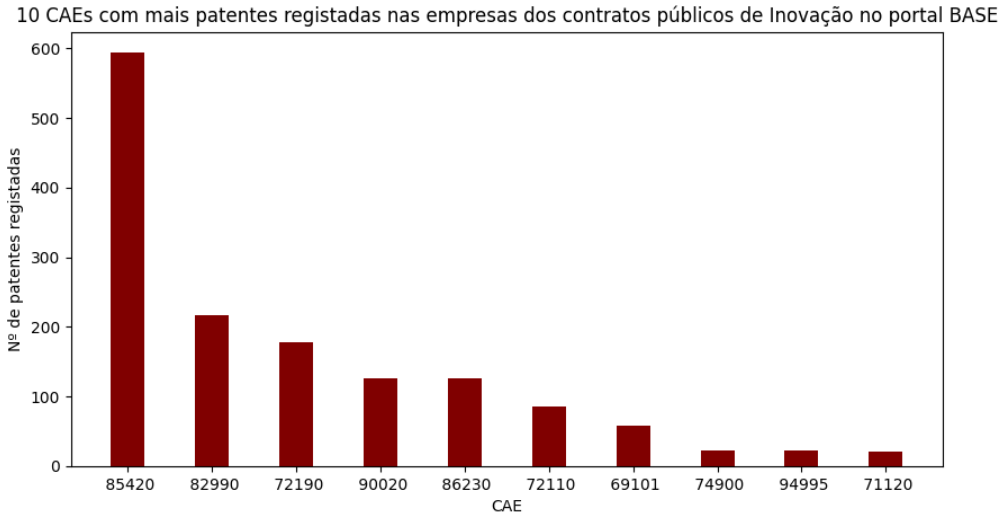


Figura 4.22: Gráfico de dispersão que relaciona o número de patentes registadas pelas empresas com a soma do financiamento obtido por estas em contratos públicos de inovação, por CAE

Após a análise dos gráficos de dispersão e dos respetivos coeficientes de correlação, passou-se à análise dos gráficos de barras com as frequências das variáveis em estudo, partindo de prismas diferentes para cada um. No gráfico da figura 4.23, é interessante compará-lo com o da figura 4.7 e perceber que o CAE 70220 já não figura nos CAEs com mais patentes registadas, contudo muitos outros relacionados com ensino, investigação e engenharia, mantém-se tanto no top 10 de CAEs com mais financiamento como no de maior registo de patentes. No número de patentes registadas o CAE que se apresenta com maior número é o 85420, ensino superior, o que faz todo o sentido, pois seria de esperar que as universidade e os centros de investigação fossem os locais por excelência de ocorrência de inovação e é o que se confirma pelos dados.

Olhando agora para os anos em que se registaram mais patentes e as mesmas empresas obtiveram mais contratos públicos, respetivamente nos gráficos de barras das figuras 4.24 e 4.25, podemos concluir que ambas têm os anos entre 2013 e 2020, como anos de maior frequência da sua variável, apenas faltando os anos de 2008 e 2009 nos anos com maior número de contratos obtidos por estas empresas que registaram patentes, o que faz sentido por serem os anos onde se começou a registar estes dados no Portal BASE, como já tinha sido abordado anteriormente e os anos de 2021 e 2022, presentes no gráfico do número de contratos, não se encontram no das patentes registadas, o que também faz sentido por nos encontrarmos em 2023 e o registo de patentes ser um processo que demora o seu tempo e mesmo a própria investigação demora o seu tempo, logo era de prever à priori que nos anos próximos daquele em que nos encontramos houvesse menos registo de patentes que noutros já consolidados pelo tempo.



85420 - Ensino superior
82990 - Outras actividades de serviços de apoio prestados às empresas
72190 - Outra investigação e desenvolvimento das ciências físicas e naturais
90020 - Actividades de apoio às artes do espectáculo
86230 - Medicina dentária e odontologia
72110 - Investigação e desenvolvimento em biotecnologia
69101 - Advogados
74900 - Outras actividades de consultoria, científicas, técnicas e similares
94995 - Outras actividades associativas
71120 - Engenharia e técnicas afins

Figura 4.23: Gráfico de barras com os 10 CAEs que registaram mais patentes

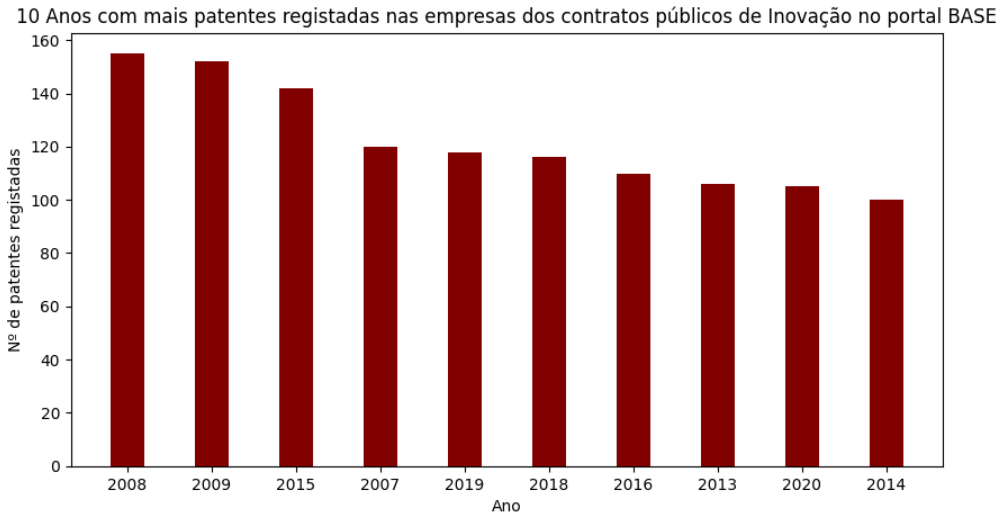


Figura 4.24: Gráfico de barras com os 10 anos onde as empresas registaram mais patentes

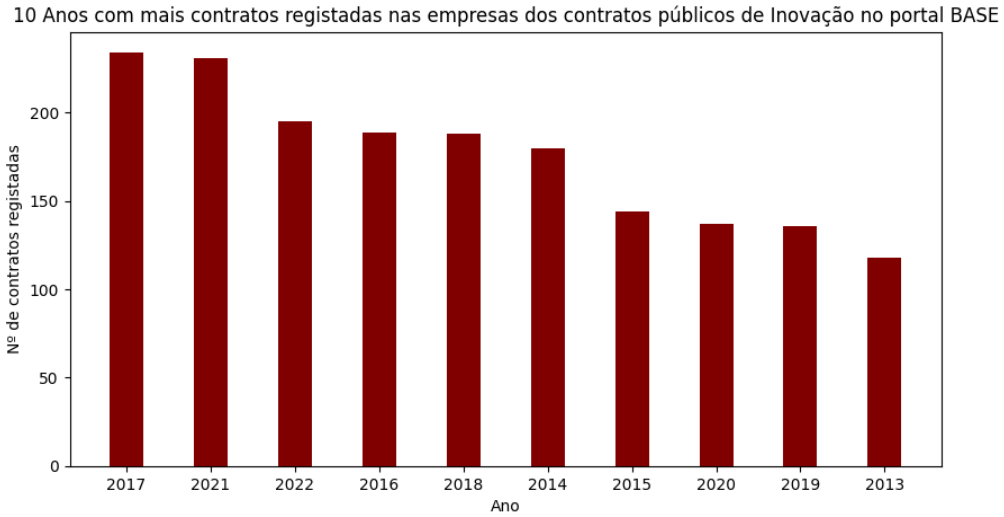


Figura 4.25: Gráfico de barras com os 10 anos onde as empresas obtiveram mais contratos públicos de inovação

Em seguida, olhando para os valores máximos dos anos em que foram registadas patentes e obtidos contratos públicos, respetivamente das tabelas 4.8 e 4.4, podemos concluir que estes foram respetivamente 2021 e 2023, por outro lado o valor mínimo de cada um, foi 2008 para o número de contratos obtidos e 1979 para o número de patentes registadas, pelo que se formaram gráficos de barras com as frequências por anos das duas variáveis entre os anos de 2008 e 2023, estes são apresentados nas figuras 4.26 e 4.27. Olhando para o gráfico do número de contratos, pode observar-se valores mais baixos entre 2008 e 2012 em comparação com os anos seguintes e volta a baixar em 2023, provavelmente devido à mesma razão que se tem vindo a mencionar nesta dissertação, de que é o ano em que nos encontramos atualmente e ainda não existem muitos contratos registados, dado os dados terem sido recolhidos, sensivelmente, a meio do mesmo. Já em relação ao número de patentes registadas, 2008 e 2009 foram os anos com mais patentes registadas, sendo que depois há um decréscimo em todos os anos seguintes, exceto 2015, que os números se aproximam mais dos de 2008 e 2009 e sendo este decréscimo mais significativo em 2017. Comparando o número de patentes que foram registadas em cada ano pelas empresas presentes no *dataset* com as que foram registadas por todas as empresas a nível nacional [45] e que se representou na figura 4.28, pode notar-se novamente um dos valores mais altos no ano de 2015, sendo que nacionalmente, é mesmo o mais alto, a nível nacional não se notou uma diferença tão grande nos anos de 2008 e 2009, que a este nível não é dos melhores anos em termos de registo de patentes, mas no *dataset* acaba por o ser, o mesmo para 2017, mas em sentido contrário a diferença que tem em relação aos outros anos nos dados do *dataset* desta dissertação, não se verifica nacionalmente. Olhando percentualmente, o ano em que as patentes registadas por empresas do *dataset* representam maior percentagem daquelas registadas pelas empresas a nível nacional é o de 2008, em que representam 33% das

patentes totais, desce para 22% em 2009, é sempre inferior a 20% a partir desse ano, atingindo a percentagem mínima de 10% em 2017 e terminando em 12% em 2021, último ano de que se obteve dados. Estas percentagens representam o máximo de influência que os contratos públicos de inovação podem ter tido no registo de patentes, percentagens essas relativamente baixas, pelo que poderá argumentar-se que caso os contratos públicos de inovação sejam bons promotores do registo de patentes a nível nacional, com o objetivo de aumentar o número de registos das mesmas, se poderia investir mais neste tipo de contratos, por outro lado pode argumentar-se também que se o que é investido no máximo tem este impacto, que, possivelmente, já se investe o suficiente ou até se poderia investir menos por ser uma estratégia pouco eficaz, contudo para tirar estas conclusões seria necessário fazer mais estudos, porque tudo depende da força da relação entre as variáveis número de patentes e financiamento, que vimos que é boa, mas não é alta.

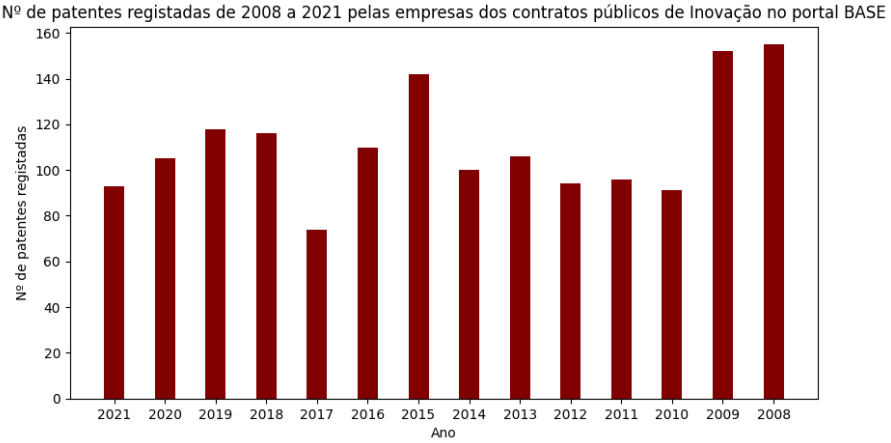


Figura 4.26: Gráfico de barras com o número de patentes registadas, entre 2008 e 2021, pelas empresas do *dataset* que tenham pelo menos uma patente registada

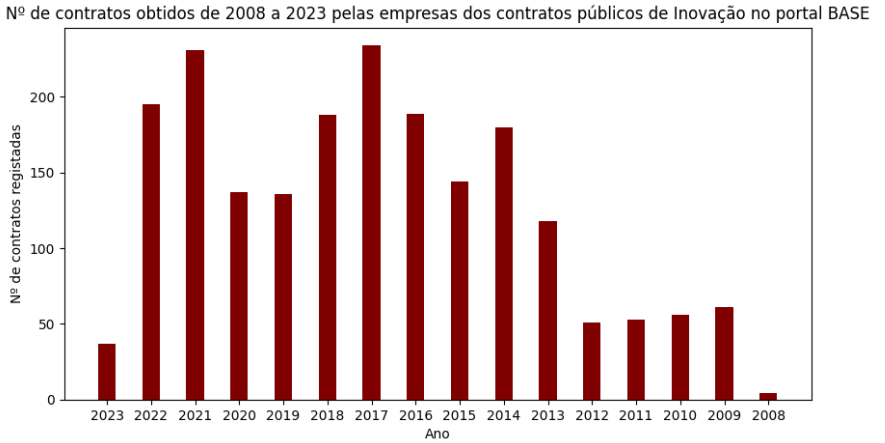


Figura 4.27: Gráfico de barras com o número de contratos públicos de inovação obtidos, entre 2008 e 2023, pelas empresas do *dataset* que tenham pelo menos uma patente registada

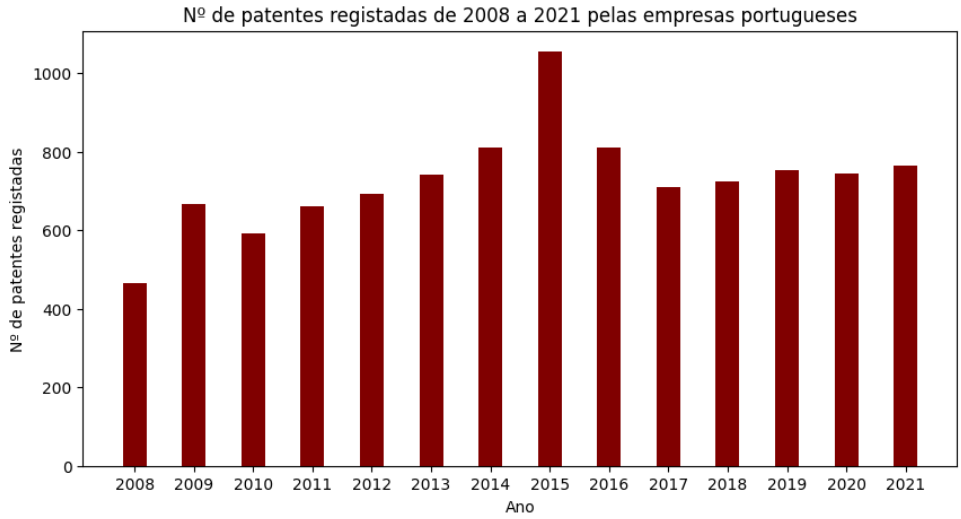


Figura 4.28: Gráfico de barras com o número de patentes registadas, entre 2008 e 2021, pelas empresas portuguesas

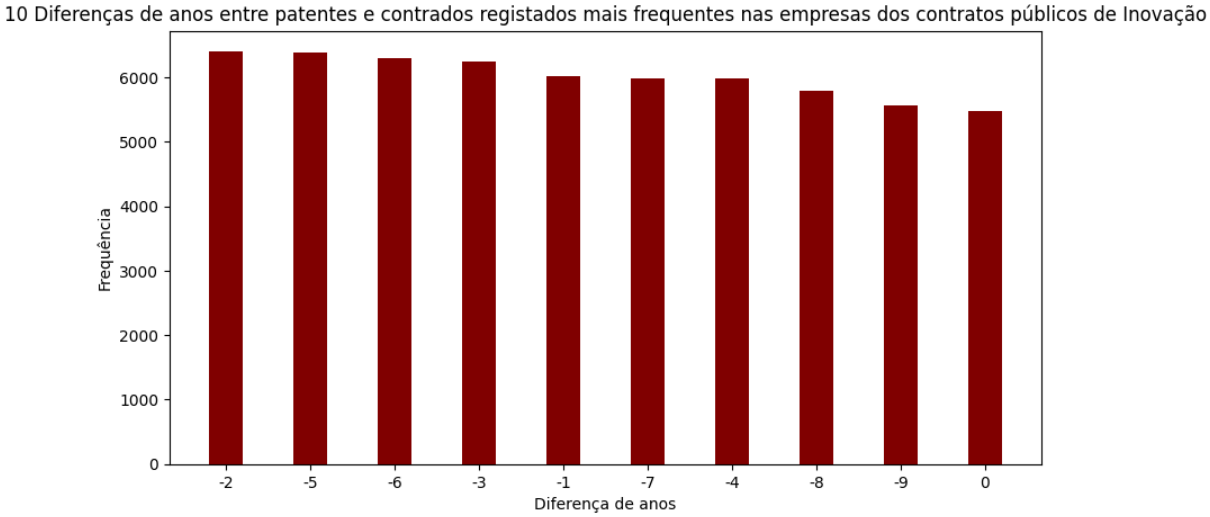


Figura 4.29: Gráfico de barras com as 10 diferenças mais frequentes entre os anos em que as empresas do *dataset* registaram patentes com os anos em que obtiveram contratos públicos de inovação

A última análise realizada foi a apresentada no gráfico da figura 4.29. Neste gráfico são apresentadas as dez diferenças de anos mais frequentes entre os anos de registo de patentes e os anos de obtenção de contratos públicos de inovação por cada uma das empresas presentes no *dataset*. Considerou-se uma diferença positiva aquela em que primeiro se obteve um contrato público de inovação e depois se registou uma patente, ou seja, quando o ano de obtenção de contrato era inferior ao ano de registo de patente. Por exemplo, se uma empresa registou patentes nos anos de 2015, 2016 e 2017 e obteve contratos públicos nos anos de 2016, 2017 e 2018, as diferenças obtidas irão ser 2015-2016=-1, 2016-2016=0, 2017-2016=1, 2015-2017=-2, 2016-2017=-1, 2017-2017=0,

2015-2018=-3, 2016-2018=-2 e 2017-2018=-1. Do exemplo anterior, podíamos concluir que as suas diferenças são, maioritariamente, negativas, pelo que essa empresa estará a receber financiamento, após o registo de patentes. Observando o gráfico mencionado anteriormente pode concluir-se o mesmo, ou seja, observa-se apenas diferenças negativas, excetuando o zero. Era de esperar que as diferenças mais comuns fossem as positivas, por indicar que se está a financiar as empresas através de contratos públicos de inovação, antes destas registarem patentes, contudo o que se sobressai nestes dados é exatamente o contrário. Isto pode querer indicar que se está a financiar empresas, após estas registarem patentes e não antes para as ajudar a registá-las, o que não tem necessariamente de ser algo mau, pois para algumas tarefas, pode querer-se empresas com provas dadas neste sentido, contudo como norma pode ser algo preocupante. De notar que esta é apenas uma hipótese que necessita de mais estudo e mais dados, para ser concretizada, mas que se colocou nesta dissertação, por ser algo interessante, dado resultar numa conclusão contra intuitiva.

4.4 Sumário do capítulo

Neste capítulo, que tratava a análise empírica, começou-se por descrever os dados, em primeiro lugar, descrevendo o tipo de dados atribuído a cada um dos campos das duas entidades da base de dados formulada, contratos e empresas e, em segundo lugar, analisando, para as variáveis relevantes, a estatística descritiva básica e alguns gráficos de barras e circulares, para aquelas variáveis em que este tipo de gráficos acrescentava à análise. Denotou-se que a mediana se encontrava, para várias variáveis, com uma diferença significativa da média, pelo que se estudou as percentagens de valores acima e abaixo da média e se deixou como alerta para a ocorrência de possíveis *outliers*. Constatou-se ainda pela análise dos gráficos circulares com as frequências de cada CAE e CPV nos contratos públicos de inovação do *dataset*, que em ambas as variáveis existia uma grande predominância de atividades relacionadas com consultoria. Terminou-se o capítulo abordando os resultados empíricos da análise. Confirmou-se a relação entre número de contratos obtidos e financiamento obtido em contratos públicos de inovação, com coeficientes de correlação sempre acima de 50% para as várias análises efetuadas, pelo que se conclui que as empresas que recebem mais contratos públicos de inovação são também aquelas que recebem maior financiamento. Na análise dos dez CAEs com mais e menos financiamento foi relevante entender novamente a predominância da consultoria nas atividades dos contratos obtidos pelas empresas presentes nos dois CAEs com mais financiamento e que após essas atividades vinham atividades relacionadas com investigação e desenvolvimento, sendo que novamente o CAE com mais financiamento se relacionava com consultoria e o segundo com atividades de investigação e desenvolvimento, contudo o primeiro enquadra-se num setor de atividade com mais expressão no VAB nacional, que o segundo, pelo que surge a questão de se se deveria financiar mais CAEs com mais expressão no VAB nacional ou CAEs com atividades, na

teoria, mais diretamente relacionadas com inovação? Além disso, será que esta menor expressão no VAB nacional, se deve ao menor financiamento ou há menor eficiência, pelo menos no imediato, destas atividades terem impacto no VAB nacional? Após este estudo, tentou-se entender a relação entre a dimensão das empresas com o financiamento obtido em contratos públicos de inovação, o que tanto medida pela faturação da empresa, como pelo número de colaboradores, resultou em baixos coeficientes de correlação, pelo que, por um lado, é um bom sinal de homogeneidade na atribuição desses fundos, tanto a empresas com maior dimensão, como empresas com menos expressão, o que é importante, mas, por outro lado, dado que em teoria é nas empresas mais pequenas, normalmente *start-ups*, que existe maior disrupção seria interessante ver um investimento mais forte neste tipo de empresas e também nas empresas de maior dimensão, mas estas últimas pelo motivo oposto, por normalmente estarem mais estabelecidas e não irem tão à procura desta inovação, pelo que seria importante financiá-la. Por último estudou-se a relação entre número de patentes registadas e financiamento obtido em contratos públicos de inovação, onde se observaram coeficientes de correlação sempre a rondar os 50% nas várias análises executadas e também se constatou que os CAEs mais financiados, em grande maioria, eram os CAEs que registavam mais patentes, pelo que dá mais força à relação entre as duas variáveis, mesmo esta não se afigurando muito alta, o que é coerente com a última hipótese apresentada, mas esta sim, necessitaria de ser ainda mais estudada, para ser confirmada, mas de que talvez se esteja a financiar mais as empresas após estas registarem patentes, ou seja, após estas mostrarem de alguma maneira o seu mérito e não a financiar as empresas para estas desenvolverem o trabalho necessário para depois registarem as suas patentes, dado que as 10 diferenças com mais representação no *dataset* são entre anos de registo de patentes inferiores aos anos de obtenção de contratos públicos de inovação. Para terminar esta análise ainda se compararam a percentagem de patentes no total nacional, que podem ter influência de contratos públicos de inovação e a grande maioria destas percentagens são todas inferiores a 30%, pelo que se pode argumentar tanto que estas baixas percentagens se devem a falta de investimento, como que os contratos públicos podem ser um mecanismo ineficaz para a estimulação do registo de patentes como forma de fomentar a inovação.

5

Conclusões e trabalho futuro

O trabalho produzido nesta dissertação permitiu desenvolver algoritmos de extração de dados de contratos públicos de inovação, de extração de empresas desses contratos, de recolha dos CAEs de cada empresa, de recolha do número de patentes registadas e anos de registo das mesmas pelas empresas resultantes das extrações anteriores e ainda recolha de dados económicos sobre as empresas. Permitiu ainda desenvolver métodos de limpeza, armazenamento e organização desses dados, como ainda métodos de análise empírica e visualização dos resultados dessas análises. Assim sendo, todo este trabalho, pretende demonstrar as possibilidades existentes através de técnicas de *web scraping*, de exploração das potencialidades de armazenamento e estruturação de dados das bases de dados relacionais e de métodos de limpeza, análise e visualização de dados para desenvolver estudos relevantes à economia portuguesa, como foi o que neste caso se elaborou em relação aos contratos públicos nacionais no que diz respeito ao seu potencial de fomentação da inovação em Portugal.

Além dos aspetos mais técnicos, através deles extrapolaram-se análises, hipóteses e conclusões. Sendo a mais forte de que as empresas que recebem mais contratos públicos de inovação são também aquelas que recebem maior financiamento. Contudo, também se deu os primeiros passos em hipóteses relacionadas com o forte financiamento de atividades relacionadas com consultoria e o questionamento deste facto, para no futuro compreender se esta é uma boa medida ou se se deveria tentar redirecionar o investimento para outras áreas de atividade. Também se observou uma baixa relação entre o financiamento obtido em contratos públicos de inovação e a dimensão da empresa, facto interessante no que diz respeito ao não favorecimento de empresas pela sua dimensão, mas seria interessante no futuro continuar este estudo, entrando em mais detalhe no tema, para responder à questão se esta será a maneira mais eficiente de utilizar este tipo de investimento e se não quem se deveria financiar, as empresas mais pequenas e, por norma, mais disruptivas ou as empresas maiores, com maiores necessidades de inovação? Por último, o estudo que se fez às patentes registadas a nível nacional foi interessante para inferir uma boa relação entre esta variável e o financiamento obtido pelas empresas em contratos públicos de inovação, mas seria

interessante continuar este estudo para tentar compreender se esta relação se mantém com maior, menor ou igual intensidade noutros universos de dados, mas também para tentar responder à questão de se o facto de as percentagens das patentes registadas a nível nacional que podem ser impactadas por contratos públicos de inovação serem inferiores a 30% é uma percentagem correta, pois mais investimento não faria com que aumentasse o número de patentes e talvez a inovação, por ser um método menos eficaz a potenciar este indicador ou, pelo contrário, se deveria investir mais neste sentido para que a percentagem aumentasse. E por último, também, seria interessante continuar a estudar as diferenças entre os anos de registo de patentes e os anos de obtenção de contratos públicos de inovação, para compreender se o padrão de o ano de registo de patentes ser, por norma, inferior ao ano de obtenção de contratos públicos de inovação, se mantém, noutros cenários, para assim tentar responder à questão de se se anda a financiar empresas, apenas após provas dadas, através de registo de patentes, quando, possivelmente, seria importante investir no contrário.

Com esta dissertação tentou-se contribuir não só com conhecimento técnico para estudos que possam vir no futuro no mesmo sentido do que se tentou estudar nesta dissertação, mas também contribuir com análises, hipóteses e questões, que podem servir de base para estudos ou dissertações futuras que se queiram debruçar em mais detalhe em cada temática apresentada nesta dissertação.

Bibliografia

[1] European Comission. *Public Procurement*. Jul. de 2022. URL: https://ec.europa.eu/growth/single-market/public-procurement_en (ver p. 1).

[2] Organisation for Economic Co-operation e Development. *Public Procurement*. Jul. de 2022. URL: https://www.oecd.org/governance/public-procurement/ (ver p. 1).

[3] European Comission. *Innovation procurement*. Jul. de 2022. URL: https://ec.europa.eu/growth/single-market/public-procurement/strategic-procurement/innovation-procurement_en (ver p. 1).

[4] Organisation for Economic Co-operation e Development. *Public Procurement for Innovation*. Jul. de 2022. URL: https://www.oecd.org/gov/public-procurement/innovation/ (ver p. 2).

[5] European Central Bank. *How does innovation lead to growth?* Jul. de 2023. URL: https://www.ecb.europa.eu/ecb/educational/explainers/tell-me-more/html/growth.en.html (ver pp. 2, 5).

[6] Daron Acemoglu e Joshua Linn. "Market Size in Innovation: Theory and Evidence from the Pharmaceutical Industry". Em: *The Quarterly Journal of Economics* 119 (ago. de 2004), pp. 1049–1090. URL: https://doi.org/10.1162/0033553041502144 (ver p. 2).

[7] Fábio Lazzarotti, Michael Samir Dalfovo e Valmir Emil Hoffmann. "A bibliometric study of innovation based on Schumpeter". Em: *Journal of technology management & innovation* 6.4 (2011), pp. 121–135 (ver p. 5).

[8] Peter F Drucker et al. "The discipline of innovation". Em: *Harvard business review* 76.6 (1998), pp. 149–157 (ver p. 5).

[9] McKinsey. *What is innovation?* Jul. de 2023. URL: https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-innovation (ver pp. 5, 6).

[10] M Rogers. *The definition and measurement of innovation*. English. WorkingPaper 10/98. Melbourne Institute of Applied Economic e Social Research, mai. de 1998 (ver pp. 5, 7–9).

[11] Anthony Arundel e Dorothea Huber. "From too little to too much innovation? Issues in measuring innovation in the public sector". Em: *Structural Change and Economic Dynamics* 27 (2013), pp. 146–159. ISSN: 0954-349X. DOI: https://doi.org/10.1016/j.strueco.2013.06.009. URL: https://www.sciencedirect.com/science/article/pii/S0954349X13000611 (ver pp. 6, 9).

[12] Harald Bathelt, Patrick Cohendet, Sebastian Henn e Laurent Simon. *The Elgar companion to innovation and knowledge creation*. Edward Elgar Publishing, 2017 (ver pp. 6–8).

[13] Amram R. Shapiro. "Measuring Innovation: Beyond Revenue From New Products". Em: *Research-Technology Management* 49.6 (2006), pp. 42–51. DOI: 10.1080/08956308.2006.11657407. eprint: https://doi.org/10.1080/08956308.2006.11657407. URL: https://doi.org/10.1080/08956308.2006.11657407 (ver p. 8).

[14] Riitta Katila. "Using patent data to measure innovation performance". Em: *International Journal of Business Performance Management* 2.1-3 (2000), pp. 180–193. DOI: 10.1504/IJBPM.2000.000072. URL: https://www.inderscienceonline.com/doi/abs/10.1504/IJBPM.2000.000072 (ver p. 8).

[15] Agustı Segarra-Blasco, Jose Garcia-Quevedo e Mercedes Teruel-Carrizosa. "Barriers to innovation and public policy in Catalonia". Em: *International entrepreneurship and management journal* 4 (2008), pp. 431–451 (ver p. 10).

[16] R. P. Oakey. "Funding innovation and growth in UK new technology-based firms: Some observations on contributions from the public and private sectors". Em: *Venture Capital* 5.2 (2003), pp. 161–179. DOI: 10.1080/1369106032000097049. eprint: https://doi.org/10.1080/1369106032000097049. URL: https://doi.org/10.1080/1369106032000097049 (ver p. 10).

[17] Rohmat Gunawan, Alam Rahmatulloh, Irfan Darmawan e Firman Firdaus. "Comparison of Web Scraping Techniques : Regular Expression, HTML DOM and Xpath". Em: *Proceedings of the 2018 International Conference on Industrial Enterprise and System Engineering (IcoIESE 2018)*. Atlantis Press, 2019/03, pp. 283–287. ISBN: 978-94-6252-689-1. DOI: https://doi.org/10.2991/icoiese-18.2019.50. URL: https://doi.org/10.2991/icoiese-18.2019.50 (ver pp. 10, 14, 15, 19–21).

[18] De S Sirisuriya. "A comparative study on web scraping". Em: (2015) (ver pp. 10, 12, 13).

[19] Daniel Glez-Peña, Anália Lourenço, Hugo López-Fernández, Miguel Reboiro-Jato e Florentino Fdez-Riverola. "Web scraping technologies in an API world". Em: *Briefings in Bioinformatics* 15.5 (abr. de 2013), pp. 788–797. ISSN: 1467-5463. DOI: 10.1093/bib/bbt026. eprint: https://academic.oup.com/bib/article-pdf/15/5/788/17488715/bbt026.pdf. URL: https://doi.org/10.1093/bib/bbt026 (ver p. 11).

[20] Sanjay Kumar Malik e SAM Rizvi. "Information Extraction Using Web Usage Mining, Web Scrapping and Semantic Annotation". Em: *2011 International Conference on Computational Intelligence and Communication Networks*. 2011, pp. 465–469. DOI: 10.1109/CICN.2011.97 (ver p. 13).

[21] Tobi Sam. *Easiest way to remember Regular Expressions (Regex)*. Set. de 2023. URL: https://towardsdatascience.com/easiest-way-to-remember-regular-expressions-regex-178ba518bebd (ver p. 14).

[22] NOVA School of Science e Technology. *Home Page*. Jul. de 2022. URL: https://www.fct.unl.pt/ (ver pp. 14, 15).

[23] Mihai Gheorghe, Florin-Cristian Mihai e Marian Dârdală. "Modern techniques of web scraping for data scientists". Em: *International Journal of User-System Interaction* 11.1 (2018), pp. 63–75 (ver pp. 16, 17).

[24] Instituto dos Mercados Públicos do Imobiliário e da Construção (IMPIC). *Potal BASE*. Mai. de 2023. URL: https://www.base.gov.pt/base4 (ver pp. 16, 32).

[25] Bo Zhao. "Web Scraping". Em: mai. de 2017, pp. 1–3. ISBN: 978-3-319-32001-4. DOI: 10.1007/978-3-319-32001-4_483-1 (ver pp. 18, 19).

[26] TaeHong Kim, YoonSeok Cha, ByeonChun Shin e ByungRae Cha. "Survey and Performance Test of Python-Based Libraries for Parallel Processing". Em: *The 9th International Conference on Smart Media and Applications*. SMA 2020. Jeju, Republic of Korea: Association for Computing Machinery, 2020, pp. 154–157. ISBN: 9781450389259. DOI: 10.1145/3426020.3426057. URL: https://doi.org/10.1145/3426020.3426057 (ver p. 20).

[27] Aldo Hernandez-Suarez, Gabriel Sánchez-Pérez, Karina Toscano-Medina, Victor Manuel Martínez-Hernández, Victor Sanchez e Héctor M. Pérez Meana. "A Web Scraping Methodology for Bypassing Twitter API Restrictions". Em: *ArXiv* abs/1803.09875 (2018) (ver pp. 21–23).

[28] Fatmasari Fatmasari, Yesi Kunang e Susan Purnamasari. "Web Scraping Techniques to Collect Weather Data in South Sumatera". Em: Pangkal, Indonesia, dez. de 2018. DOI: 10.1109/ICECOS.2018.8605202 (ver pp. 21, 22, 24).

[29] M. Rebaudengo, M. Sonza Reorda, M. Torchiano e M. Violante. "Soft-error detection through software fault-tolerance techniques". Em: *Proceedings 1999 IEEE International Symposium on Defect and Fault Tolerance in VLSI Systems (EFT'99)*. 1999, pp. 210–218. DOI: 10.1109/DFTVS.1999.802887 (ver p. 24).

[30] Oracle. *What Is a Database?* Jul. de 2022. URL: https://www.oracle.com/pt/database/what-is-database/ (ver p. 24).

[31] PostgreSQL. *pgAdmin*. Jul. de 2022. URL: https://www.pgadmin.org/ (ver pp. 25, 26).

[32] Nishtha Jatana, Sahil Puri, Mehak Ahuja, Ishita Kathuria e Dishant Gosain. "A survey and comparison of relational and non-relational database". Em: *International Journal of Engineering Research & Technology* 1.6 (2012), pp. 1–5 (ver pp. 26, 27).

[33] Cornelia Gyorödi, Robert Gyorödi e Roxana Sotoc. "A comparative study of relational and non-relational database models in a Web-based application". Em: *International Journal of Advanced Computer Science and Applications* 6.11 (2015), pp. 78–83 (ver pp. 27, 28).

[34] SAP. *What is data modeling?* Jul. de 2022. URL: https://www.sap.com/insights/what-is-data-modeling.html (ver p. 29).

[35] Peter Pin-Shan Chen. "The Entity-Relationship Model—toward a Unified View of Data". Em: *ACM Trans. Database Syst.* 1.1 (mar. de 1976), pp. 9–36. ISSN: 0362-5915. DOI: 10.1145/320434.320440. URL: https://doi.org/10.1145/320434.320440 (ver p. 29).

[36] Lucidchart. *What is a Database Model*. Jul. de 2022. URL: https://www.lucidchart.com/pages/database-diagram/database-models (ver p. 29).

[37] Wendy Hall e Thanassis Tiropanis. "Web evolution and Web Science". Em: *Computer Networks* 56.18 (2012). The WEB we live in, pp. 3859–3865. ISSN: 1389-1286. DOI: https://doi.org/10.1016/j.comnet.2012.10.004. URL: https://www.sciencedirect.com/science/article/pii/S1389128612003581 (ver p. 30).

[38] Daniel Schwabe, Guilherme Szundy, Sabrina Silva De Moura e Fernanda Lima. "Design and implementation of semantic web applications". Em: *WWW Workshop on Application Design, Development and Implementation Issues in the Semantic Web*. 2004 (ver p. 30).

[39] Jim Conallen. "Modeling web application architectures with UML". Em: *Communications of the ACM* 42.10 (1999), pp. 63–70 (ver p. 30).

[40] Kai Lei, Yining Ma e Zhi Tan. "Performance Comparison and Evaluation of Web Development Technologies in PHP, Python, and Node.js". Em: *2014 IEEE 17th International Conference on Computational Science and Engineering*. 2014, pp. 661–668. DOI: 10.1109/CSE.2014.142 (ver p. 30).

[41] Bureau van Dijk. *Orbis*. Jul. de 2023. URL: https://orbis-r1.bvdinfo.com/ (ver p. 33).

[42] Instituto Nacional da Propriedade Industrial (INPI). *Pesquisa de Patentes*. Mai. de 2023. URL: https://servicosonline.inpi.justica.gov.pt/pesquisas/main/patentes.jsp?lang=PT (ver p. 34).

[43] Pordata. *Empresas: total e por dimensão*. Jul. de 2023. URL: https://www.pordata.pt/portugal/empresas+total+e+por+dimensao-2857 (ver p. 41).

[44] Instituto Nacional de Estatística. *PIB a preços de mercado na ótica da produção - VAB por ramo de atividade, A38 (preços correntes; anual)*. Jul. de 2023. URL: https://www.ine.pt/ngt_server/attachfileu.jsp?look_parentBoui=97111807&att_display=n&att_download=y (ver p. 51).

[45] Pordata. *Patente, Design*. Jul. de 2023. URL: https://www.pordata.pt/portugal/invencoes+patentes+de+residentes+em+portugal+pedidos+e+concessoes+da+via+nacional++europeia+e+internacional-1279 (ver p. 64).