NOVA IMS Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

Using business data to assess the impact of Covid-19 in Portuguese Private Sector Activity

Beatriz Seca Freitas

Dissertation presented as partial requirement for obtaining the Master's degree in Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Using business data to assess the impact of COVID-19 in Portuguese Private Sector Activity

by

Beatriz Seca Freitas

Dissertation presented as partial requirement for obtaining the master's degree in Advanced Analytics

Advisor: *Prof Doutor* Flávio Luís Portas Pinheiro

May 2022

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the help of several people that supported me throughout this process.

First, I want to thank my parents for allowing me to enroll in this master's degree and supporting me throughout the entire process. My little brothers and sisters that, although not understanding what I was doing, always showcased faith in my work.

I want to thank my friends Laura, Catarina and André for helping me develop this thesis, always with good advice and patience. Maria, Rita and Madalena, which although distant always listened and supported me. Miguel, for being there throughout the worst times and pushing me until the finish line. My group of friends from the bachelor's degree which made the entire academic experience worthwhile.

I would like to thank my advisor, Flávio Luís Pinheiro, for guiding me through this process.

Thank you Edite and Manuel, for raising me and believing in me.

ABSTRACT

COVID-19’s containment measures resulted in a severe shock to all national businesses. The supply and demand were strongly impacted due to suspending the companies' typical behavior and the physical absence of customers. More than ever, there is an increased need for data to monitor these effects quickly and affordably. Therefore, this project proposes to assess the endogenous determinants of Portuguese private sector firms in their financial performance during the first year of Covid-19, while showcasing secondary data's potential within private institutions. The main goal is to extract historical data from one of the largest Banks in Portugal with features that reflect important business characteristics and deliver essential insights into the evolution of national enterprises. Several linear regression models were implemented to study which business attributes, from 2019, made companies more fragile or robust to COVID-19’s outbreak. Hence, delivering knowledge to build catered solutions to the vast spectrum of Portuguese firms and protect them from the next shock.

The results suggested that for the variation in total revenue, the variables responsible for the largest coefficients in absolute value were the accommodation and food services and the micro dimension category - both of which showcased a negative impact. Furthermore, the only positive coefficients were the features associated with exporting activity and the energy and manufactory sector. The number of employees was more impacted by the dimension of the company, whose negative impact was proportional to the size. The accommodation and food services, other services and wholesale and retail trading showcased having a significant adverse effect on this target variable, and exportation activity presented the only positive impact.

KEYWORDS

Firms; COVID-19; Linear Regression; Data partnership

INDEX

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

In 2019, the Portuguese business landscape was made of 1,335,006 companies, a growth of 3.1% since 2018 (INE, 2021). In the same year, the ecosystem of non-financial corporations accounted for 91.1% of the GDP from the entire private sector with a total turnover of 439,742 million Euros (INE, 2021). With the COVID-19 pandemic, there were meaningful social and economic changes that impacted not only social life, but also the normal activity of all kinds of businesses. Indeed, lockdown measures that have been implemented to contain the spread of COVID-19 had a strong impact both on the supply and the demand dynamics of economic activities (Manteu et al., 2020).

Surveys and financial reports are one of the main sources of information used to aid the government in understanding the efficacy of measures to control de pandemic. They have been widely used to obtain a picture of the behavior and condition of private institutions, which sometimes they do not get access to. Given that surveys collect data at a single point in time, to assess temporal changes we are required to repeat such surveys periodically. Such repetition is often expensive and time-consuming, making surveys as frequently as desirable - impractical (Woodruff, 2009). Although several studies have been done on the impact of COVID-19 on Portuguese businesses, like the IREE-COVID-19 from February (INE & Bank of Portugal, 2021), they were mainly descriptive and not inferential.

Morales et al. (2021) state that COVID-19 created a spike in demand for data to monitor the evolution of the pandemic and its effects. However, it has exacerbated funding gaps in national, regional, and global statistical offices, making the mobilization of international and domestic resources to support data more urgent than ever. When compiling official statistics, analysis of data held by companies is often more cost-efficient and can produce faster outcomes. Using such data also lowers the burden on companies and citizens by avoiding survey questionnaires (Morales et al., 2021). Hence, public-private data collaboration holds great promise for mitigating some of society's ongoing challenges.

This dissertation looks to assess the endogenous determinants of Portuguese private sector firms in their financial performance during the first year of Covid-19 Pandemic. To that end, and in contrast with previous studies, we use secondary data sourced one of the largest private Banks operating in Portugal. Firstly, I showcase that available data in the Banking institutions can offer several advantages over traditional sources (e.g., it is less time-consuming to compile, has a lower cost of acquisition, and lower overhead to obtain results). Moreover, like traditional data, it delivers insights to create and assess policies, develop better solutions, and improve decision-making. Secondly, using multivariate regression methods I explore which pre-pandemic factors - such as importing and exporting activity, dimension of the business, presence of a website and industry - might have been determinants to overcome the economic constraints created during the COVID-19 pandemic. In particular, estimations will deliver deliver a picture of which firms were better equipped to overcome the several limitations that the pandemic raised to the overall economic operations.

2. THEORETICAL BACKGROUND

Mobilizing data for social good is as ambitious as it is a necessary challenge. National statistical offices still rely heavily on the empirical base of surveys and other primary data collection approaches to develop and test social interventions and policies. However, according to Kabir (2016) obtaining the data is often the main expense in studies, and getting the data collected (personally or through others) is usually an extensive process. Meanwhile, much helpful knowledge remains in the silos of legacy information systems or commercial enterprises' digital footprints (Coulton et al., 2015). Secondary data, in other words, data collected for other purposes, does not create the hassles of data collection, which saves time and expenses (Ajayi, 2017)– and private corporations hold a great quantity of it.

Recent studies suggest that big data has become a useful, and perhaps necessary, addition to data sources and analysis methods used to examine social welfare practice and policy (Gillinghama, & Grahamb, 2017). Integrating administrative data across service agencies is expected to be the next frontier for generating quality evidence to inform public protocols and system reform. Also, such research might be possible in months rather than years and at a fraction of the cost compared to research based on primary data collection and qualitative methods. (Culhane, et al. 2010). As written by Avdiu & Gaurav (2020): “More systematic and timely evidence is needed to inform policy debates on how to best support firms during the COVID-19 pandemic, especially when countries are running out of the fiscal policy space to sustain stimulus packages”. However, there are some ethical issues with Big Data in social welfare. Individuals, businesses, and governments face common challenges when data are accessed and shared. OECD (2019) made a run-through of the three main challenges when enhancing access and sharing of data:

- Need for balancing the benefits of data “openness” with other legitimate interests, policy objectives, and risks. These include the risks of confidentiality and privacy breaches and the violation of other legitimate private interests, such as commercial interests.
- Trust and empowerment for the effective re-use of data across society. Trust can play an essential role in data access, sharing, and re-use across firms, sectors, and countries. However, it can be abused or erode over time, and restoring it can be challenging.

- The marginal costs of transmitting, copying, and processing data can be close to zero. However, substantial investments are often required to collect data and enable data sharing and re-use. Firms are investing a significant share of their capital in acquiring start-ups to secure access to data potentially critical for their business. They may also be needed for data cleaning and preparation, which is often beyond the scope and time frame of the activities for which the data were initially collected and used.

Jumia, a Nigerian e-commerce enterprise, offered its logistics and supply-chain network and collaborated with health departments in different African countries to use its website and mobile platforms to share COVID-19 related public service messages (Jumia Group, 2021). Another example is Meta Platforms, which has been assisting governments in reaching citizens with health alerts through its messenger and WhatsApp platforms (Nyman, 2020).

RELATED WORKS

The spread of COVID-19 in 2020 set corporations under tremendous stress to survive, demanding firms to develop appropriate reactions to the crisis. Businesses' resilience is considered to be "The capability of firms to readily respond to and quickly recover from disruptive events or crises faced" (Arıkan, 2022). The faster a firm reacts, the less its financial performance decreases. The financial performance of a company can be evaluated through fluctuations in different variables, such as the number of employees and the total revenue (cash flow made through sales of goods and services). To investigate the relationship between the financial performance and other business features we used a linear regression model. Regression models are a widely used statistical tool for analyzing the impact of the pandemic (such as cases and spreads), especially when investigating the correlation between COVID-19 and different variables (Almalki et al., 2021). The chosen factors for our regression models were collected from previous work about companies' vulnerability and resilience. The following section introduces this literature and gives a summary of the determinants that have been identified to impact firms' vulnerability and resilience to shocks.

INE performed multiple surveys to characterize the short-term economic impact of COVID-19. These were the base for reporting the main effects of the pandemic on companies' activity (in terms of revenue, human resources, financial conditions, and prices). These studies conducted by INE aimed at micro, small, medium, and large companies from a wide spectrum of sectors of activity. Although very relevant, these studies are only descriptive and not inferential. Contrarily to inferential statistics (such as regression analysis), descriptive statistics cannot infer about the population (Sutanapong & Louangrath, 2015). Nonetheless, reports like these inspired this dissertation and served as a baseline for our discussion.

Social isolation and containment measures did not allow for purchases to be made physically. According to SIBS analytics (2020), just in Portugal, about 8 billion euros were lost in transactions between March to June of 2020, which is equivalent to around 200 million transactions. Hence, another way to deliver services and goods is through e-commerce. OECD (2021c) considers that online presence can increase resilience to shocks: "Amidst the economic mayhem caused by the pandemic, digital technologies have emerged as a key element of economic and social resilience. By allowing businesses and governments to continue operating during lockdowns, digital technologies enabled faster economic responses to the emergency". However, more than 60% of Portuguese firms do not have a simple website or a presence on Facebook (Teixeira, 2020).

Importing and/or exporting goods and services can also be studied in relation to business resilience. As stated by OECD (2022) "The year 2020 was marked by some of the largest reductions in trade and output volumes since World War II.". Borino et al. (2021) states that exporting and/or importing can be a differentiating factor when analyzing firms' resilience. Impacts within the national market influence both domestic and international firms, while shocks that happen within international markets will only directly affect international firms. Hence, international enterprises are considered more sensitive to shocks than firms that only operate nationwide (Borino et al., 2021). More results from Borino et al. (2021) showcase that an international firm had more than a 60% probability of experiencing problems accessing goods and services. However, if the same firm had exclusively collected inputs domestically, it would have had only a 50% likelihood of encountering complications. This study also demonstrates that the probability of an enterprise which imports and/or exports to follow a resilient strategy is approximately 10% higher than a domestic firm.

Revenue-wise, a multinational firm had about 6% more chances of undergoing a decrease in sales compared with a domestic firm (Borino et al., 2021).

Regarding employment, it was found that international firms had a smaller likelihood of entering the layoff regime for their workers (Borino et al., 2021). Another research on this topic, Eppinger et al. (2018), showcased that in the 2008 financial crisis, exporters kept more jobs than other companies.

The size of the enterprise can also be associated with its resilience to shocks. SMEs are frequently considered more vulnerable than large corporations under diverse shocks (OECD 2021a, OECD 2020, McKinzey 2020). There is a large spectrum of different kinds of shocks and how they affect SMEs' response. As written by Miklian & Hoelscher (2022) "The pandemic's length also affects smaller firms more intensely as they generally lack sufficient resources to tolerate extended periods of upheaval, with many of them closing temporarily of permanently. Micro-firms are most vulnerable, representing the smallest percentage of businesses that had any 'rainy day' savings.". Sales-wise, smaller companies showcased greater losses than other firms, which "(…) severely affects their ability to function, and/or causes severe liquidity shortages" (OECD, 2020). McKinzey (2020) stated that two thirds of jobs are at risk in European SMEs and "more than 30 percent of all jobs at risk are found within microenterprises consisting of nine employees or fewer".

The literature establishes a connection between different industries and firm's resilience to shocks. According to Avdiu & Gaurav (2020), industries that rely on face-to-face contact, either by nature of the sector (like transportation) or not adapting to the digital world, were the most damaged. "The ease of doing home-based work is key during lockdown, while the intensity of face-to-face interactions assumes greater importance as restrictions ease. As non-essential services are increasingly suspended due to the COVID-19 outbreak, industries less amenable to home-based work are particularly vulnerable to immediate job losses." (Avdiu & Gaurav, 2020).

3. DATA AND METHODOLOGY

The focus of this dissertation is to analyze the influence of firms' endogenous determinants on their financial performance throughout the first year of the pandemic (2020). Given the data available, we focus on the following research questions:

1. Is the growth rate in total revenue impacted or related to importing and exporting, online presence, industry, or size of the business?
2. Is the growth rate in the number of employees impacted or related to importing and exporting, online presence, industry, or size of the business?

3.1 DATA

We use data sourced from the data repository of one of the largest Banking institutions operating in Portugal. The obtained dataset can be divided into two main sets:

1. Registration Data - Static information regarding different clients given the moment of registration in the bank. This table delivers insights about immutable characteristics like the economic sector, and geographic location of the client.
2. Activity Data - Dynamic information regarding the activity of the client throughout the time. This table will translate the changes throughout the different periods on aspects of the business, such as total revenue or number of employees fluctuations.

For security and privacy reasons, customer names and tax identification numbers were excluded at the moment of extraction. In their place, pseudo-unique identifiers were created to distinguish between different clients. The extracted dataset includes more than 196k records for the years of 2019 and 2020 and contains 9 columns: Customer ID; Total Revenue (in euros); Number of Employees; Net worth (in Euros); Website (url); Import and Export status; year, month, and day of the record. Additional information about the Code of Economic Activity (CAE) and the district were also available per company.

Next, we detail some of the pre-processing and exploration steps undertaken to obtain a working dataset and familiarity with its properties and characteristics.

DUPLICATES AND MISSING VALUES

Each record details the status of a firm in a month. However, several quantities of interest for our study (such as net worth, total revenue and number of employees) are updated only every so often (i.e., once per year, or once every quarter). As such, the dataset contains many repeated entries and only a few have the actual updated values.

Since this dissertation looks to assess the yearly accounting status (for 2019 and 2020), a bar plot was assembled to understand which dates are the most recurrent for updating the financial values with a year apart. The higher bars in **figure 1** represent the most frequent dates where the financial values changed: 2020-10-03 and 2021-08-02. Only the firms that changed values on these two dates were chosen, delivering 28 628 companies.

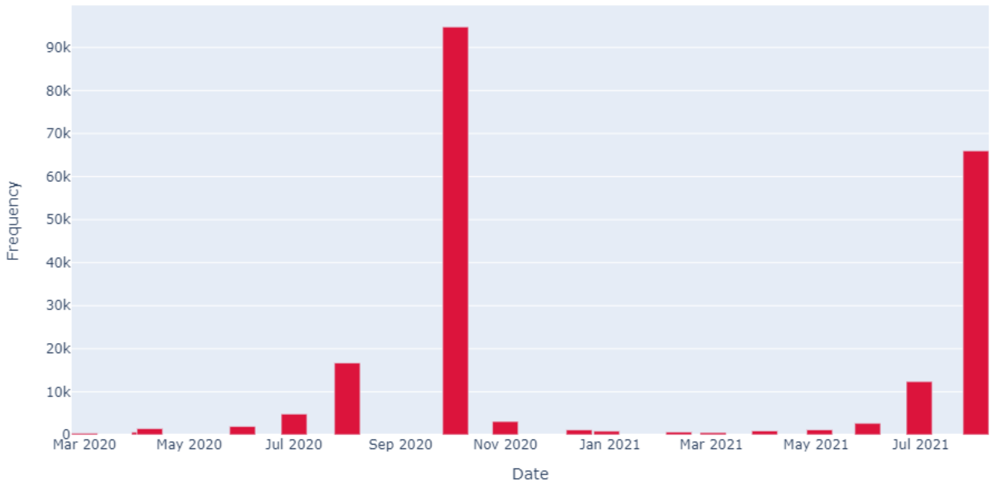

Figure 1 - Bar plot of most frequent dates for changing the total revenue values

For the variable Number of employees, a small portion of the values was missing having a 'nan' value instead. Discussing these records with Santander, it was reported that they had no valuable information - as such those records were discarded. For the variable CAE, there was a very small percentage of the population with records equal to '#'. Even though the literature sometimes advises replacing the missing values, it was not worth the risk of inducing bias into the data, given that they made a small portion of the records. Hence, these values of "#" were left in the data set.

CATEGORICAL VARIABLES

The dataset includes several categorical features, and additional ones were convenient to engineer in order to control for effects that would only manifest at firms of different sizes or industry sectors.

In that sense, firms were grouped into different categories depending on the business's total revenue and the number of employees. Therefore, according to INE's categorization of firms' dimensions (INE & Bank of Portugal, 2021), the variables total revenue and number of employees were employed to screen the following four sizes:

- **Microenterprise** - number of employees <10 and turnover ≤ 2 million euros.
- **Small enterprises** - number of employees <50, turnover ≤ 10 million euros and not classified as a micro-enterprise.
- **Medium-sized enterprises** - number of employees <250, turnover ≤50 million euros and not classified as micro or small enterprise.
- **Large enterprises** - number of employees ≥ 250 or turnover > 50 million euros.

Concerning the economic sector of activity of firms, INE & Bank of Portugal (2021) suggests these to be divided into the following groups:

- **Energy and manufacturing -** enterprises classified into sections B, C, D and E.
- **Construction and real estate activities -** enterprises classified into sections F and L.
- **Wholesale and retail trade -** section G.
- **Transportation and storage** – section H.
- **Accommodation and food service activities** – section I.
- **Information and communication** – section J.
- **Other services -** section M, N, P, Q, R and S.

The different grouped activities are based on the CAE (see Appendix 1).

Regarding the online presence, most the companies do not have a website. The gap is particularly large within micro-enterprises: 18k do not have an online presence compared with the almost 5k that do. For the remaining dimensions, the gap is not as extensive. There are 39 enterprises with a website for large companies and 9 firms with no website. Industry-wise most sectors have more firms with no online presence. The gap between these two categories is extensive for most, except for energy and manufacturing and information and communication. The latter is the only sector with more companies with a web platform.

Concerning the importing and exporting activity, its distribution varies throughout the different dimensions. For micro-companies, 9.2k do not import nor export, followed by companies that do both (6.2k) and 5.3k only import. For the medium-sized and small companies, most enterprises imported and exported, followed by firms that only import, and in third the one's that do neither. For large corporations, the greater portion imports and exports; the second largest portion has companies that only export, and lastly, enterprises that only import. We also analyzed the distribution of importing and exporting activity by industry. Accommodation and food services, construction and real estate, and other services have a larger portion of businesses that do not import or export; followed by companies that only import and companies that do both. The wholesale and retail industry has 5.2k companies that import and export, 2.8k that only import, and 1.8k that do neither. Energy and manufacturing, transportation and storage, and information and communication have companies that import and export as the largest category, followed by neither.

After exploring our sample, several steps were conducted to build features, address data quality issues, including cleaning and handling potential problems to ensure reliable results. Additionally, various techniques to study the relationship between features were employed to determine if there is a need for removing or transforming variables

Categorical data is non-numeric and frequently can be defined into groups. Thus, an encoding method can be required to turn these non-numeric categories into numbers. It is tempting to assign an integer, say from 1 to k, to each of k possible categories; however, it will create a hierarchical order which usually is not desirable. There are several types of encoding; However, the one we will use is the one-hot-encoding, which builds dummy variables for each category. (Casari & Zheng, 2018).

The categorical variable associated with the website was transformed into a binary variable called **Has_Website**, which assumes 1 if the company has a website and 0 otherwise. The industry feature was transformed into seven different dummies, one for each grouped activity; The dimension was unfolded into four different dummies and the IMP_EXP was also transformed into four different dummies. (See Appendix 2)

Description	Unique	Most Frequent	Freq
IMP_EXP	4	0	32146
WEBSITE	7306	0	30611

CAE	585	41200	3798
GCAE	353	home and building construction	4119
Division	48	47	13623
District	18	Lisbon	34921
Size	4	Micro	50256
Industry	7	Trading	19653

Table 1- General statistics of categorical features

CONTINUOUS VARIABLES

To create a linear regression model, we must guarantee that the dependent variables have a linear relationship with all independent variables. **Figure 2** (plot on the left) shows that none of the present variables have a linear relationship. A few observations appear to be distant from most data points creating a certain degree of skewness which can be associated with a log-normal distribution. The logarithm is the power to which (exponent) a base number must be raised to get the original number (Osborne, 2002). Transformations like these can address three different scenarios: errors non-normally distributed, the non-linearity relationship between X's and y, or both. Log-transform handles these sources of misspecification while remaining within the linear modeling framework and continuing with OLS estimation. Furthermore, the reduction in skewness might avoid a complex analysis of leverage (how much the outliers impact the estimate of the regression coefficients), and even avoid eliminating some of these extreme values (Pek et al., 2018). Hence, a transformation was applied to the total revenue and net worth (independent variables) to reduce skewness and achieve a linear relationship between the dependent and the explanatory variables. It was decided not to transform the number of employees for two reasons. First, this variable was used to create the dimension of the company and since it is a more relevant feature for the scope of the study than the number of employees per se, one could replace the other. Secondly, in later sections, the correlation amongst several features will be studied, and other variables are present in the dataset that highly correlate with the number of employees. Hence, it can be dropped without further analysis.

To create our target variables, the growth rate between 2019 and 2020 of the number of employees and total revenue was built using the following equation:

$$\frac{Y_T - Y_{T-1}}{Y_{T-1}} \qquad (1)$$

Where Y_T represents the data point in the period T and $Y_{T-1}$the data point in the period T-1.

After the transformation of the total revenue and net worth, the association between explanatory and target variables was still far from linear. It was decided to proceed with an outlier analysis to investigate if the removal of extreme values would result in a linear relationship between response and explanatory features.

The Tukey's method (Salgado et al., 2016) uses several statistics such as median and the values of the lower (25%) and the upper (75%) quartiles to provide a robust measurement of the data and detect extreme values. The interquartile range (IQR) is the length between the lower (Q1) and upper (Q3) quartiles in the boxplot. Inner fences are defined as 1.5 times IQR below Q1 and above Q3, and outer fences are defined as 3 times IQR below Q1 and above Q3 (Nascimento et al., 2021). An observation within the inner and outer fences is a possible outlier, and a value outside the outer limit is a probable outlier. Removing the probable outliers is referred to as Tukey's method. (Salgado et al., 2016)

According to Tukey's method the percentage growth rate of total revenue has 162 outliers, from which 111 are possible outliers, and 51 are probable outliers. For the percentage growth rate of the number of employees, the inner and outer fences are the same since Q1=0 and Q2=0. Hence, the total number of outliers is 10363. There are 327 outliers, 13 probable and 314 possible for the total revenue, for the net worth 450, 19 probable, and 431 possible. When treating outliers, one should be careful and avoid deleting data as much as possible since the variability caused by these extreme values can represent the population at study. Finding a balance between the benefits of removal and the downsides of loss in variability is essential. To remove as few records as possible, I've only removed the probable outliers from the total revenue's growth rate, which delivered the results from **figure 2** (plot on the right) where the relationship between the independent and dependent variables appears to be closer to a linear association.

Description	Total Revenue	Number of Employees	Net Worth	Total revenue growth rate	Number employees Growth rate
Mean	788446.76	7.61	447189.32	1.1169	-0.0034
Std	5554189.22	124.24	5302133.80	133.4282	0.3317
Coefficient of Variation (CV) = $\frac{Std}{Mean}$	7.044469585	16.32588699	11.8565752	119,46	-98,51
Min	0.00	0.00	-73108581.82	-1	-1
25%	66011.42	1.00	15092.05	-0.3076	0
50%	170180.95	3.00	76033.23	-0.0938	0
75%	451944.53	6.00	254948.97	0.0789	0
Max	542432143.41	13938.00	647496181.97	20709	8

Table 2 - General Statistics of numerical features

To understand the behavior of the continuous variables **table 2** was built. The mean values and all three percentiles for the total revenue and for the number of employees are quite low. In other words, on average the sample is made of micro-companies, and at least 75% of the sample is made of firms from this dimension. Moreover, for these two features plus the net worth, the coefficient of variation is very far from one, a symptom of the sample's high variability.

Concerning the growth rates of the total revenue and of the number of employees the mean showcases that, on average per company, there was an increase in 111.69% of total revenue and a decrease in the number of employees by 0.34%. Concerning total revenue' and number of employees' growth rates the CV is 119.46 and -98.51, respectively. These values are considered quite large and illustrate the variability of the sample. The quartiles showcase that for the total revenue's growth rate 50% of the values are negative – at least half of the companies decreased its total revenue between 2019 and 2020. The growth rate of the number of employees has every quartile equal to zero – which means that 75% of the population did not experience any changes in the number of employees.

To avoid collinearity issues, we used a correlation matrix with the Pearson coefficients to analyze the relationships between features. Large correlation coefficients can be a symptom of high collinearity. A rule of thumb is introduced to define a threshold beyond which it should be considered removing variables. Correlation values higher than the absolute value of 0.70 are studied and if needed removed. According to the correlation matrix (see Appendix 3) the Number of Employees and the Total Revenue have a coefficient of 0.73. Moreover, the micro and small dimensions are highly correlated - a phenomenon also known as 'dummy trap.' A dummy trap occurs when binary variables born from the same feature are heavily correlated. Hence, the small dimension variable and the Number of Employees were deleted.

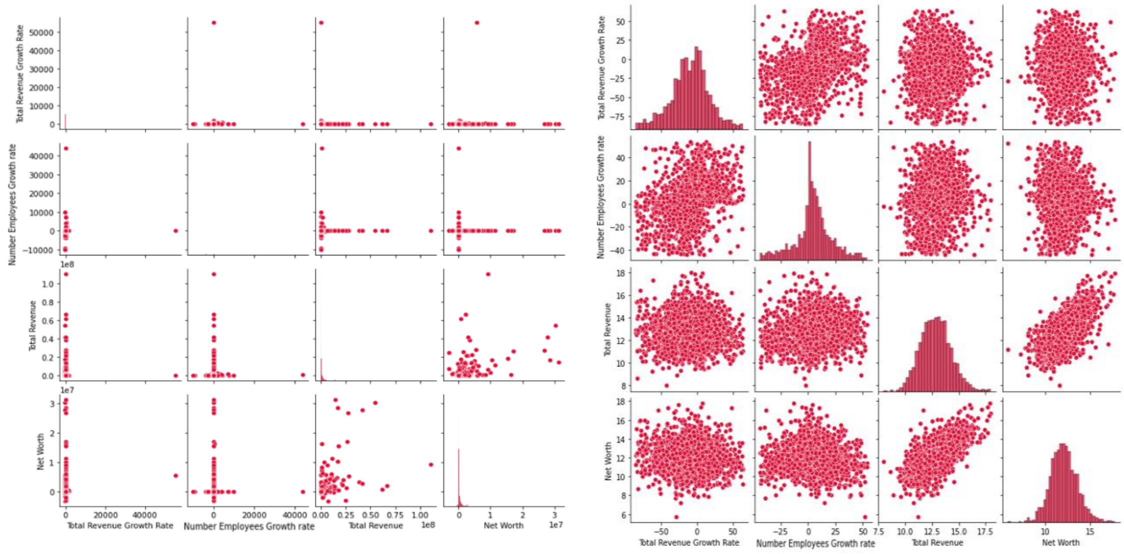

Figure 2 - Pair-Plot showing in the diagonal the distribution of observations of each variable and in the non-diagonal cells the relationship between each pair of variables. The left plot shows the relationships before the executed data transformation steps (log-transformation and outlier removal).

DATA EXPLORATION

After the data treatment stage, we are left with 56256 records.

The growth of the Total Revenue, Net Worth, and Number of Employees between 2019 and 2020 will be the target of the regression analysis. Hence it is important to deliver an overview of their evolution and a deeper understanding of the sample's distribution amongst the different categorical features - which will later serve as predictors.

There is a total of 27628 firms that fit into a size category. The micro-enterprise class has the highest concentration of entities with around 94% of the total sample, unlike large corporations, which only portray around 0.5% of the population. Small and medium-sized companies match around 4% and 1.5% of the sample, respectively.

Concerning the number of employees according to size more than 41% of the sample belongs to micro-companies, followed by large companies with 36%. Medium companies have the minor portion with 7% of intuitions, while small companies have 16%. In terms of total revenue 36% comes from micro companies, followed by large (26%), small (25%) and finally medium companies, which represent 13% of the total sample. For the Net Worth distribution, 49% belonged to the micro companies' category, followed by 26% belonging to small companies and the remainder are divided almost equally within other sizes, with medium companies having a larger quantity.

Regarding activity, the trading sector stands out with around 38% of the companies, as opposed to the energy and manufacturing, and the information and communication sectors, equivalent to 1% and 3% of firms, respectively. Transportation and storage, accommodation and food service activities account for around 5% and 7%; construction and real estate activities along with other services represent around 20% and 26% of companies.

For the total revenue, more than a half belongs to trading (about 55%), followed by other services (15%) and in third place, construction, and real estate with 12%. On the other side of the spectrum, there is energy and manufacturing with 2% of the sample; progressively, there is accommodation and food services (3%), information and communication (4%) and lastly, transportation and storage with nine percentage values. Regarding the number of employees, the largest proportion is within in the other services (42%), followed by the wholesale and retail trading (30%) and thirdly construction and real estate activities (13%). In contrast, there is energy and manufacturing with 5%, info and communication with 4%, and accommodation and food services plus transportation and storage with approximately 6%.

For the net worth, the vast majority is divided amongst construction and real estate, and wholesale and the retail industry. On the other side of the spectrum there is energy and manufactory and information and communication.

3. Results

The empirical results of the OLS estimates for the coefficients will be presented in this chapter. The historical data from 2019 plus the growth rates between 2019 and 2020 will be the canvas to various models to showcase how certain factors impact the dependent variable and which variables are more relevant to explain it. The value of the coefficient relating explanatory variables with the target is calculated, taking into account the correlations of all other independent variables among themselves and with the dependent variable. That is, a positive coefficient between total revenue's growth rate and exportation activity, for example, indicates that after considering all associations between this variable and the other independent variables, as well as the association between the latter and the target, there is a positive association between the exportation activity and growth rate of total revenue. Breakdowns like this are a much more demanding test than the mere significance of a simple correlation coefficient. The coefficient's assessment will be made under the Ceteris Paribus condition. All changes within the target caused by the change of one explanatory variable are only considered with other factors (variables) held constant.

To study which endogenous factors impact the total revenue the following model was designed:

$$\begin{aligned}
&Total\ Revenue\ Growth\ Rate = \beta_0 + \beta_1 IMP\ EXP\ Exports + \\
&\beta_2 IMP\ EXP\ Imports + \beta_3 IMP\ EXP\ Imports \text{ and } Exports + \\
&\beta_4 IMP\ EXP\ Neither + \\
&\beta_5 Industry\ Accomodation\ and\ food\ service\ actvities + \\
&\beta_6 Industry\ Energy\ and\ manufacturing + \beta_7 Industry\ other\ services + \\
&\beta_8 Industry\ Transportation\ and\ storage + \\
&\beta_9 Industry\ Information\ and\ communication + \\
&\beta_{10} Industry\ Construction\ and\ real\ estate\ activities + \\
&\beta_{11}\ Industry\ Wholesale\ and\ retail\ trading + \beta_{12} Dimension\ Micro + \\
&\beta_{13} Dimension\ Medium\ Sized + \beta_{14} Dimension\ Large + \beta_{15} Has\ Website + \\
&\textstyle\sum_{i=16}^{33} \beta_i\ District_i + \epsilon
\end{aligned} \tag{2}$$

Also, to study which endogenous factors impact the variation in number of employees the following model was developed:

$$\begin{aligned}
&Number\ of\ Employees\ Growth\ Rate = \beta_0 + \beta_1 IMP\ EXP\ Exports + \\
&\beta_2 IMP\ EXP\ Imports + \beta_3 IMP\ EXP\ Imports\ and\ Exports + \\
&\beta_4 IMP\ EXP\ Neither + \\
&\beta_5 Industry\ Accomodation\ and\ food\ service\ actvities +
\end{aligned} \tag{3}$$

$$\beta_6 Industry\ Energy\ and\ manufacturing + \beta_7 Industry\ other\ services + \beta_8 Industry\ Transportation\ and\ storage + \beta_9 Industry\ Information\ and\ communication + \beta_{10} Industry\ Construction\ and\ real\ estate\ activities + \beta_{11}\ Industry\ Wholesale\ and\ retail\ trading + \beta_{12} Dimension\ Micro + \beta_{13} Dimension\ Medium\ Sized + \beta_{14} Dimension\ Large + \beta_{15} Has\ Website + \sum_{i=16}^{33} \beta_i\ District_i + \epsilon$$

the $\sum_{i=16}^{33} \beta_i\ District_i$ present in equations (1) and (2) represents the dummies for all eighteen districts, each with its own parameter. These dummies will serve as control variables, to make sure that the effects of the different districts are accounted for and do not inflate or deflate the estimates of the other features. The ϵ term represents the error.

The proposed model factors were chosen based on the literature relating resilience and vulnerability with different business characteristics. This fragility is interpreted as fluctuations in factors such as the number of employees and total revenue. If a firm, in response to shock, experience largely negative variations during the year (when compared with the remainder), it is considered a vulnerable firm.

The dummies related with importing and exporting activity were based on the literature of Borino et al. (2021). This article stated that importing and exporting is considered a differentiating factor when studying business resilience and vulnerability to shocks. The results showcased total amount of sales is more likely to decrease for a multinational firm than for a domestic firm and that international firms had a smaller likelihood of entering the layoff regime for their workers (Borino et al., 2021). Another research on this topic, Eppinger et al. (2018), showcased that in the 2008 financial crisis, exporters kept more jobs than other companies.

The dummies regarding the size of the enterprise were also included. Miklian & Hoelscher (2022) and OECD (2020) wrote that the dimension of a firm is related with its vulnerability to impacts, specifically, SMEs are frequently more vulnerable than large corporations under diverse shocks. Their decrease in revenue and number of workers when compared with bigger businesses is considered a symptom of their fragility (OECD, 2020).

For the industry dummies, the literature (OECD, 2021b) stated that different levels of vulnerability to shocks are related with different economic sectors. The response varies with the industry’s characteristics, which may amplify or reduce the impact of containment measures and restrictions on activity caused by the pandemic.

The online activity is also considered a significant determinant of resilience since it delivers an alternative for production and market exchanges when shocks cause severe disruptions to physical economic activity (OECD, 2021b).

Model (16) (**table 3**) showcases the results of implementing the model in equation (2). The district control variables were hidden since they are not the features of interest to our research.

Results

	Dependent variable:															
	Total Revenue Growth Rate															
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Industry_Accommodation and food service activities	-26.957***															-30.083***
	(0.795)															(0.958)
Industry_Construction and real estate activities		3.402***														-1.200
		(0.526)														(0.878)
Industry_Energy and manufactoring			8.753***													2.144**
			(2.218)													(2.879)
Industry_Transportation and storage				1.075												-3.978***
				(0.845)												(0.943)
Industry_Other services					-1.870***											-4.946***
					(0.454)											(0.776)
Industry_Information and communication						4.050***										0.569
						(1.203)										(1.543)
Industry_Wholesale and retail trade							5.031***									1.233
							(0.403)									(0.923)
Dimension_Micro								-3.223***								-8.569***
								(0.465)								(0.899)
Dimension_Large									5.832							0.579
									(4.394)							(5.004)
Dimension_Medium-sized										-1.679						-3.589**
										(1.362)						(2.456)
GIMP_EXP_Exports											2.258***					3.007***
											(0.749)					(0.994)
GIMP_EXP_Neither												-2.225***				-0.435
												(0.429)				(0.600)
GIMP_EXP_Imports													-2.036***			-0.914
													(0.471)			(0.912)
GIMP_EXP_Imports and Exports														2.893***		1.245
														(0.409)		(0.234)
Has Website															0.910**	2.446
															(0.428)	(0.312)
Constant	-8.694***	-10.971***	-10.456***	-10.448***	-9.900***	-10.499***	-12.450***	-7.936***	-10.394***	-10.348***	-10.557***	-9.690***	-9.913***	-11.483***	-10.673***	-6.675***
	(0.199)	(0.218)	(0.200)	(0.205)	(0.231)	(0.202)	(0.258)	(0.405)	(0.199)	(0.201)	(0.207)	(0.240)	(0.227)	(0.252)	(0.241)	(0.680)
Observations	27,628	27,628	27,628	27,628	27,628	27,628	27,628	27,628	27,628	27,628	27,628	27,628	27,628	27,628	27,628	27,628
R^2	0.06	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10
Adjusted R^2	0.06	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10
Residual Std. Error	25.512	26.305	26.324	26.335	26.323	26.328	26.220	26.300	26.335	26.335	26.329	26.316	26.322	26.299	26.333	22.352
F Statistic	1,150.068***	41.829***	15.571***	1.618	16.988***	11.339***	156.168***	48.065***	1.074	1.519	9.087***	26.907***	18.678***	49.940***	4.525**	150.574***
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$															

Table 3 - Regression table for the target variable Total Revenue Growth Rate with univariate and multivariate regression models.

Model (16) (**table 3**) only explains 0.10 of the variation in the total revenue's growth rate.

The average growth rate in total revenue from a company with all features equal to zero is -6.675%. Furthermore, not all variables are deemed relevant for explaining the target variable in model (16). Only the following factors were considered significant. A business that only exported in 2019, is expected to increase its total revenue by 3%. Moreover, belonging to the energy and manufacturing industry is estimated to increase a company's total revenue by 2.144%. On the other hand, if the enterprise belongs to the accommodation and food services sector, the estimated decrease would be around 30%.

For the remaining industries, businesses from other services and transportation and storage, are expected decrease its total revenue by 4.946% and 3.978%, respectively. If the firm was a micro company the estimated decrease would be 8.569%.

For the remaining models in **table 3** they showcase the individual importance of each feature in explaining the total revenue. All variables are statistically significant except for the transportation and storage, medium-sized and large dimension dummies.

We can now answer the question "Is the growth rate in total revenue impacted or related to importing and exporting, online presence, industry, or size of the business?". According to the results, size-wise only micro-companies are estimated to generate a significant impact in the total revenue. Regarding industry, only specific sectors such as accommodation and food service activities, transportation and storage, other services and energy and manufacturing deliver significant impact to the variation in total revenue. For the exporting and importing activity only exporting is considered a relevant determinant in explaining the amount in sales and services sold by a company. The online presence is not a significant factor in explaining the growth rate of total revenue.

Model 16 **table 4** illustrates the results of applying the model in equation (3) into our sample. The district control variables were hidden since they are not the features of interest to our research.

Results																
	Dependent variable:															
	Number of employees Growth rate															
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Industry_Accommodation and food service activities	-7.085***															-8.978***
	(1.116)															(2.434)
Industry_Construction and real estate activities		2.005***														0.638
		(0.717)														(0.896)
Industry_Energy and manufactoring			2.563													2.141
			(3.021)													(3.999)
Industry_Transportation and storage				2.814**												2.009
				(1.150)												(1.529)
Industry_Other services					-0.371											-2.500***
					(0.618)											(0.800)
Industry_Information.and.communication						2.117										3.422
						(1.638)										(1.994)
Industry_Wholesale and retail trade							-0.136									-1.900***
							(0.551)									(4.502)
Dimension_Micro								-4.220***								-6.078***
								(0.633)								(0.768)
Dimension_Large									-8.157							-13.023**
									(5.982)							(6.706)
Dimension_Medium-sized										-5.327***						-10.352***
										(1.855)						(1.942)
GIMP_EXP_Exports											3.973***					4.392***
											(1.020)					(1.358)
GIMP_EXP_Neither												-0.417				0.761
												(0.585)				(1.496)
GIMP_EXP_Imports													-0.878			0.375
													(0.642)			(0.977)
GIMP_EXP_Imports and Exports														-0.145		1.856
														(0.558)		(0.751)
Has_Website															-1.451**	-3.670
															(0.582)	(1.888)
Constant	1.349***	0.559*	0.884***	0.739***	1.001***	0.845***	0.960***	4.111***	0.921***	1.021***	0.601**	1.035***	1.108***	0.960***	1.363***	8.001***
	(0.279)	(0.298)	(0.272)	(0.279)	(0.315)	(0.275)	(0.353)	(0.552)	(0.271)	(0.274)	(0.282)	(0.327)	(0.309)	(0.344)	(0.328)	(1.978)
Observations	27,628	27,628	27,628	27,628	27,628	27,628	27,628	27,628	27,628	27,628	27,628	27,628	27,628	27,628	27,628	27,628
R^2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09
Adjusted R^2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09
Residual Std. Error	35.818	35.852	35.859	35.853	35.859	35.858	35.860	35.814	35.858	35.851	35.844	35.859	35.858	35.859	35.853	34.047
F Statistic	40.306***	7.818***	0.720	5.984**	0.359	1.670	0.061	44.434***	1.859	8.249***	15.181***	0.509	1.873	0.068	6.207**	32.500***
Note:																$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 4 - Regression table for target variable Number of Employees Growth Rate with univariate and multivariate regression models.

table 4 model (16) shows that the model for all firms explain only 9% of the variation in the number of employees' growth rate.

If a company had none of the present factors, the expected variation in number of workers would be 8.001%. Only the following variables were considered significant predictors to our target variables. Firms who are either micro, medium-sized, and large are expected to decrease their number of employees by 6.078%, 10.352%, and 13.023%, respectively. Industry-wise, a company from the accommodation and food services industry has an expected decrease of 8.978%, while a firm belonging to other services and wholesale and retail trading are estimated to decrease their personnel by 2.50% and 1.90%, respectively. Furthermore, exporting companies are estimated to grow their employees by 4.392%.

Individually, half of the features are significant in explaining the number of employees (**table 4** models (1), (2), (4), (8), (10), (11) and (15)).

Model (16) (**table 4**) allows us to answer the question “Is the growth rate in the number of employees impacted or related to importing and exporting, online presence, industry, or size of the business?”. According to the results, the size of the company influences the number of employees of a firm; Concerning the industry, only specific sectors such as accommodation and food service activities, other services and wholesale and retail trading are estimated to create an impact. For the exporting and importing activity only exporting is considered relevant, and the online presence is not a significant factor in explaining the variation in personnel.

4. DISCUSSION

The results from this study, delivered important knowledge regarding the endogenous determinants that affected national enterprise's financial performance during the first year of covid.

The IREE-COVID-19 survey (INE & Bank of Portugal, 2021) is the center of this discussion since it delivered descriptive results of the pandemic's impact in a sample of national firms. However, it is important to explain the main differences and benefits between this dissertation and their work. INE delivers descriptive and not inferential results, which means that they can only describe what's happening within their data but cannot extrapolate for the entire population. In this dissertation the regression method allows to infer about the population. We analyze which factors are important for explaining fluctuations of certain financial elements and we can estimate the variation that a certain feature can induce in a company. Something that the IREE-COVID-19 did not do.

Previous research showed that a firm's economic sector is a determinant factor in explaining the type of response to shocks (OECD, 2021b). Our results confirmed that and added that only a few industries have a significant impact in explaining these responses.

The reports from INE confirm that the effect caused by COVID-19 varied throughout different sectors, and that the industry of accommodation and food services, followed by the transportation and storage, and the trading sector felt the most negative impact on the income from sales (INE & Bank of Portugal, 2021). This dissertation results were a little different. Although accommodation and food service activities have the largest decrease and represent a significant predictor of the total revenue's variation, the next industries to have a negative impact are not the same. Our findings showcase that other services is the next sector with a decrease of 5% followed by transportation and storage with an effect on the total revenue of -4%. The IREE-COVID-19 from February also stated that the trading sector, followed by the energy and manufacturing, were the top 2 industries which increased their total revenue since pre-pandemic season. Our results showcase that the wholesale and retail trading is not a significant predictor of the variation in total revenue; However, the energy and manufactory is, and it is the only feature which increases the target variable.

In terms of fluctuations in employment, the literature establishes a connection between industries which are considered non-essential and rely on face-to-face contact, with higher vulnerability to immediate job losses (Avdiu & Gaurav, 2020). This is confirmed by our results, which showcase that only three industries have a significant negative impact in explaining the variation in the number of employees: accommodation and food services (-9%), other services (-2.5%) and trading (-2%) – which are also considered non-essential industries that most demand face to face contact (Avdiu & Gaurav, 2020). INE & Bank of Portugal (2021) stated that the accommodation and food services sector had the largest portion of companies which lost employees. The wholesale and retail trading and other services were not in the top three worst industries regarding employee losses.

Regarding the size of the firm, evidence showcase that the dimension of the enterprise is related to its vulnerability. In particular, SMEs were more damaged in terms of financial performance by COVID-19 than large firms, demonstrating greater fragility to external shocks (OECD, 2021a).

Revenue-wise OECD (2021a) stated that, "these vulnerabilities translated in sharp drops in revenues from the outset of the crisis at a faster rate than they were able to cut operating costs". Our results showcase that although the dimension of a firm has an important role, only the micro size variable is considered a relevant predictor which decreases the total revenue. Taking a closer look into INE & Bank of Portugal (2021) also demonstrated that micro-enterprises suffered large losses in sales. In the first half of February 2021, micro-companies had the largest number of firms reporting an inferior revenue to the first confinement (42%), followed by small, medium and lastly large companies with 23% (the larger the company, the smaller the percentage).

The literature also states that the losses in employees vary according to the dimension of the firm. McKinzey (2020) stated that two thirds of jobs are at risk in European SMEs and "more than 30 percent of all jobs at risk are found within microenterprises consisting of nine employees or fewer". This dissertation confirmed that the dimension of a company is a significant predictor of the variation in number of employees; However, in our results large companies are associated with greater losses (-13%), followed by medium-sized companies (-10%) and lastly micro companies (-6%). INE & Bank of Portugal showed that the percentage of companies reporting a reduction in staff in the first fortnight of February increases with company size. Thus, our results go accordingly with the Portuguese results from INE which are far from the European results obtained from OECD (2021a) and McKinzey (2020).

Importing and exporting is considered a differentiating factor when studying business resilience and vulnerability to shocks (Borino et. al 2020, Kurz and Senses 2016).

Borino et al. (2021) considered that firms which imported and/or exported were more likely to decrease its total revenue compared with firms that only operate nationally; However, in the long run they showcased greater resilience. Our outcomes demonstrate that the exporting factor is significant and increases the total revenue. INE & Portuguese Bank (2021) do not have data about exporting and importing companies, hence we cannot compare our results.

International firms are considered to be less probable of employing layoff measures. (Borino et al., 2021). Eppinger et al. (2018), also wrote that in previous crisis exporters were able to keep more jobs than other companies. Our results confirm these conclusions demonstrating that the exporting activity is a significant predictor of number of employees' variation, generating a positive impact of almost 4%.

Lastly, the online presence is considered by the literature as a viable alternative for firms to redirect their businesses when COVID-19 reduced face-to-face contact (OECD, 2021c). Hence, companies that have a viable alternative to conduct their business, such as online presence, are considered less vulnerable to the restrictions to physical purchases. However, our results demonstrated that the online presence was not a significant predictor of neither the variation of the number of employees nor the total revenue.

Another relevant topic to be discussed is the low R-squared. The two models that were designed explain less than 11% of the total variability of the sample. Colton & Bower (2002) wrote that a common misconception about r squared was that if it is low then the relationship between target and explanatory variable is not reliable: "The R2 statistic can be small, yet one or more of the regression coefficient p-values can be statistically significant. Such a relationship between predictors and the response may be very important, even though it may not explain a large amount of variation in the response.". Although the proportion of variation in our sample explained by the models is small, the relationships between variables is still significant.

5. CONCLUSIONS

This chapter will begin with a summary of the project and main results and afterward, it will be divided into limitations, lessons learned, and future work.

COVID-19 and the necessary containment measures resulted in a severe shock to the Portuguese business creating a need for data that can shape policy decisions and enhance public services. To fulfill that need, this dissertation intended to study the outcomes of the pandemic in Portuguese businesses with open activity in one of the largest Banks in Portugal. The sample was extracted from the bank's data repository with features from different tables quantifying the condition of various business elements. After the data collection stage, a thorough analysis was performed to understand the underlining patterns and behaviors and assemble variables that illustrated the companies' evolution. The data was then transformed, and the outliers deleted to prepare the sample for the modelling stage.

The main goal was to study the endogenous determinants of Portuguese private sector firms in their financial performance during the first year of Covid-19 Pandemic. The second goal was to understand the degree of the impact by analyzing the coefficients.

The first feature to be analyzed was the total revenue growth rate. In sum, from all the analyzed factors, the industries have a higher percentage of impact in explaining the variation in total revenue, largely because of the accommodation and food activities which is responsible for a 30% decrease. The vast majority of the sectors generate a negative effect, and only the energy and manufactory has a positive impact. The company's size also has an important role, specifically whether a company is a micro-business or not. Lastly, the exportation activity has a small but positive influence on total revenue.

The variation in the number of employees had a larger number of features involved. By descending order of absolute value, the first coefficient is the one associated with the dimension large which was expected to generate a decrease of approximately 13%, followed by accommodation and food service activities with an expected decrease of 9%. The following variables are related to dimension: The estimations showcase that the impact caused by different dimensions ascends with the size - thus, the largest the company, the largest the loss. Lastly, by descending order of absolute value, there is the exporting dummy (the only variable with a positive impact) followed by other services, and the wholesale and retail trade both inducing a slightly negative impact.

All in all, these conclusions serve as a foundation for shaping pandemic response policies or guarantee that existing measures are operating as predicted. Moreover, splitting the companies in size and industry, delivering different feature growth rates, and specifying other characteristics such as importing or exporting activity or online presence delivers insight of which factors make companies resilient or vulnerable to shocks. Hence, helping deliver more catered solutions to the vast spectrum of Portuguese enterprises.

5.1 LIMITATIONS

The main challenges were primarily associated with the data collection stage.

With the pandemic and consequently remote work the inter-department communication and coordination got worse, which made it difficult to acquire some of the dataset's details. Additionally, some business principles necessary for understanding the dataset were usually secured by a small fraction of specialists which were not easy to contact.

Furthermore, a preliminary understanding of the data's distribution highlighted some practical inconsistencies. Given its dimension, scarcity of quality control, maintenance, and metadata regarding each table, the repository can retain unnecessary information, thus, translating into tables that are no longer helpful but still updated by a routine.

5.2 LESSONS LEARNED

In the theoretical framework, ethical issues for sharing data were mentioned, such as risks of confidentiality and privacy breaches; Other concerns surrounded the substantial investments often required to collect data and enable data sharing and re-use.

Firstly, it is possible to develop results without compromising the privacy of the clients using only aggregated non-identified data. Regarding investment, a significant advantage is that the data employed to build this project comes from a secondary data collection method; Hence, it does not create the hassles of primary data collection, saving time and expenses. There are various benefits for the organization. To start with, the Bank acquired new insights about what type of companies make most of its clientele and their evolution throughout time. This knowledge can then be used to develop better solutions and improve decision-making by identifying new sources of value, leading to unique product offerings. All in all, the institution portrays a positive role in the Portuguese economy and still boost its business.

This dissertation attempted to fill the gap between data science within the banking industry and social commonweal and showcased that enterprises such as banks and other private parties can deliver valuable knowledge to the community without compromising one of its most significant assets: data.

5.3 Future Work

COVID-19 created an urgency for reliable knowledge to be delivered quickly to develop and test policies; As previously mentioned, National statistical offices (such as INE) still rely heavily on the empirical base of surveys, and other primary data collection, which, is quite time-consuming and expensive. The first direction would be developing intelligence products. Intelligence products are a tool, dashboard, report, app, or another technical device to support a public or humanitarian cause, built on top of shared (often aggregated) corporate data - such as the sample from this dissertation. The Directorate-General of health developed an example of this type of tool. This organization created a dashboard with information regarding infections, deaths, vaccination, and other factors, all with a location filter. Following this path, and with helpful information from firms which are willing to share its information safely for social welfare, different intelligence products could be developed that delivers the status-quo of the Portuguese businesses. The data that serves as a foundation would be aggregated to avoid privacy issues; Hence delivering quickly accurate information economically.

6. BIBLIOGRAPHY

100 days of Pandemic - Sibs Analytics. (2020). Retrieved from https://www.sibsanalytics.com/wp-content/uploads/2020/06/20200629_Report-100-dias-COVID-19_SIBS_Analytics.pdf

AbuHassan, H. (2020). Using Multiple Linear Regression and Factor Analysis to Explore the Determinants of Students Success at the University.

Almalki A, Gokaraju B, Acquaah Y, Turlapaty A. Regression Analysis for COVID-19 Infections and Deaths Based on Food Access and Health Issues. Healthcare. 2022; 10(2):324. https://doi.org/10.3390/healthcare10020324

Ajayi, Victor. (2017). Primary Sources of Data and Secondary Sources of Data. 10.13140/RG.2.2.24292.68481.

Arıkan, E. (2022). The Effects of the COVID-19 Pandemic on the Hospitality Industry: Strategies for Enhancing Resilience and Recovery. In Ş. Omeraki Çekirdekci, Ö. İngün Karkış, & S. Gönültaş (Eds.), Handbook of Research on Interdisciplinary Perspectives on the Threats and Impacts of Pandemics (pp. 24-45). IGI Global. https://doi.org/10.4018/978-1-7998-8674-7.ch002

Avdiu, Besart & Nayyar, Gaurav, 2020. "When face-to-face interactions become an occupational hazard: Jobs in the time of COVID-19," Economics Letters, Elsevier, vol. 197(C).

Batista, Gustavo & Monard, Maria-Carolina. (2003). An Analysis of Four Missing Data Treatment Methods for Supervised Learning. Applied Artificial Intelligence.

Blatná, Dagmar. (2006). Outliers in Regression.

Borino, F., Carlson, E., Rollo, V., & Solleder, O. (2021, April 30). International firms: More exposed but more resilient during COVID-19. VOX, CEPR Policy Portal. Retrieved from https://voxeu.org/article/international-firms-more-exposed-more-resilient-during-COVID-19

Casari, Amanda & Zheng, Alice. 2018. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists (1st. ed.). O'Reilly Media, Inc. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T.P., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide.

Colton, James & Bower, Keith M.. (2002). Some Misconceptions About R-Square.

Coulton, C. J., Goerge, R., Putnam-Hornstein, E., & de Haan, B. (2015). Harnessing big data for social good: A grand challenge for social work (Grand Challenges for Social Work Initiative Working Paper No. 11). Cleveland, OH: American Academy of Social Work and Social Welfare.

Culhane, D. P., Fantuzzo, J., Rouse, H. L., Tam, V., & Lukens, J. (2010). Connecting the Dots: The Promise of Integrated Data Systems for Policy Analysis and Systems Reform. Intelligence for Social Policy, Retrieved from https://repository.upenn.edu/spp_papers/146

Eppinger, P S, N Meythaler, M Sindlinger and M Smolka (2018), "The great trade collapse and the Spanish export miracle: Firm-level evidence from the crisis", The World Economy 41(2): 457–493.

Gillingham, Philip & Graham, Timothy. (2017) Big Data in Social Welfare: The Development of a Critical Perspective on Social Work's Latest "Electronic Turn", Australian Social Work, 70:2, 135-147, DOI: 10.1080/0312407X.2015.1134606

Hoffman, W. H., Bick, R. B., Boral, A. B., Henke, N. H., Olukoya, D. O., Rifai, K. R., & Roth, M. R. (2019, May 30). Collaborating for the common good: Navigating public-private data partnerships. McKinsey Analytics.

INE. (2007). Portuguese Classification of Economic Activities- CAE - Rev. 3. Retrieved from: https://www.ine.pt/ine_novidades/semin/cae/CAE_REV_3.pdf

INE. (2021). Enterprises in Portugal 2019. Lisbon. Retrieved from www: <url:https://www.ine.pt/xurl/pub/437546392>. ISSN 0872-9514. ISBN 978-989-25-0561-9

INE & Bank of Portugal (2021). Fast and exceptional enterprise survey - COVID-19. February 2021. How was the lockdown situation reflected in the enterprise's activity? Retrieved from: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_destaques&DESTAQUESdest_boui=473859120&DESTAQUESmodo=2

Jumia Group. (2021, March 31). Jumia Group | Jumia Expand Your Horizons. Retrieved from: https://group.jumia.com/community-support/covid19-jumia-offers-african-governments-last-mile-delivery-solution-for-medical-supplies-trade-e-commerce

Kabir, Syed Muhammad. (2016). METHODS OF DATA COLLECTION.

Kurz, C and M Z Senses (2016), "Importing, exporting, and firm-level employment volatility", *Journal of International Economics* 98: 160–175.

Nyman, S. N. (2020, July 14). COVID-19, tech firms, and the case for data sharing. World Bank Blogs. Retrieved from: https://blogs.worldbank.org/psd/covid-19-tech-firms-and-case-data-sharing

Manteu, C., Monteiro, N., & Sequeira, A. (2020). The short-term impact of the COVID-19 pandemic on Portuguese companies. www.bportugal.pt

McKinsey. (2020). Safeguarding Europe's livelihoods | McKinsey, https://www.mckinsey.com/industries/public-sector/our-insights/safeguarding-europes-livelihoods-mitigating-the-employment-impact-of-covid-19 (accessed on 1 July 2020).

Miklian, J., & Hoelscher, K. (2022). SMEs and exogenous shocks: A conceptual literature review and forward research agenda. International Small Business Journal, 40(2), 178–204. https://doi.org/10.1177/02662426211050796

Morales, L. G. G., Perucci, F. P., & United Nations Statistics Division. (2021, March 26). How the data and statistical community stepped up to the new challenges. UN/DESA Policy Brief: COVID-19, 96.

Nascimento, G. F. M., Wurtz, F., Kuo-Peng, P., Delinchant, B., & Batistela, N. J. (2021). Outlier detection in buildings' power consumption data using forecast error. Energies, 14(24). https://doi.org/10.3390/en14248325

OECD (2019), Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies, OECD Publishing, Paris, https://doi.org/10.1787/276aaca8-en.

OECD (2020), "Coronavirus (COVID-19): SME policy responses", OECD Policy Responses to Coronavirus (COVID-19), OECD Publishing, Paris, https://doi.org/10.1787/04440101-en. Retrieved from: https://www.oecd-ilibrary.org/sites/d17b89dc-en/index.html?itemId=/content/component/d17b89dc-en

OECD (2021a), "One year of SME and entrepreneurship policy responses to COVID-19: Lessons learned to “build back better”", OECD Policy Responses to Coronavirus (COVID-19), OECD Publishing, Paris, https://doi.org/10.1787/9a230220-en.

OECD (2021b), Strengthening Economic Resilience Following the COVID-19 Crisis: A Firm and Industry Perspective, OECD Publishing, Paris, https://doi.org/10.1787/2a7081d8-en. Retrieved from: https://www.oecd-ilibrary.org/sites/d17b89dc-en/index.html?itemId=/content/component/d17b89dc-en

OECD (2021c), "The role of online platforms in weathering the COVID-19 shock", OECD Policy Responses to Coronavirus (COVID-19), OECD Publishing, Paris, https://doi.org/10.1787/2a3b8434-en.

OECD (2022), "International trade during the COVID-19 pandemic: Big shifts and uncertainty", OECD Policy Responses to Coronavirus (COVID-19), OECD Publishing, Paris, https://doi.org/10.1787/d1131663-en.

Osborne, Jason. (2002). Notes on the Use of Data Transformations. Practical Assessment, Research & Evaluation. 8.

Pek, Jolynn & Wong, Octavia & Wong, Augustine. (2018). Addressing non-normality: A taxonomy of approaches, reviewed and illustrated. Frontiers in Psychology. 9. 10.3389/fpsyg.2018.02104.

Salgado, C.M., Azevedo, C., Proença, H., Vieira, S.M. (2016). Noise Versus Outliers. In: Secondary Analysis of Electronic Health Records. Springer, Cham. https://doi.org/10.1007/978-3-319-43742-2_14

Sutanapong, C., & Louangrath, P. I. (2015). Descriptive and inferential statistics. International Journal of Research & Methodology in Social Science, 1(1), 22-35.

Teixeira, A. (2020, January 13). The current state of e-commerce in Portugal. ECL. Retrieved from: https://escolacomerciolisboa.pt/2020/01/o-estado-atual-do-e-commerce-em-portugal/

Woodruff, W.B., Bornemisza, O.B., Checchi, C.F., Global Peace and Security Fund of the Department of Foreign Affairs and International Trade, Government of Canada, & London School of Hygiene and Tropical Medicine. (2009). The Use of Epidemiological Tools in Conflict-Affected Populations: Open-Access Educational Resources for Policy-Makers. Retrieved from: http://conflict.lshtm.ac.uk/page_26.html

7. Appendix

7.1 Appendix 1

Association between Grouped activity, Section and Division (INE, 2007)

Grouped Activity	Section	Division
Energy and manufacturing	B, C, D and E	05,06,07,08,09,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25, 26,27, 28,29,30,31,32,33,35, 36,37,38,39
Construction and real estate activities	F and L	41,42,43,68
Wholesale and retail trade	G	45,46,47
Transportation and storage	H	49,50,51,52,53
Accommodation and food service activities	I	55,56
Information and communication	J	58,59,60,61,62,63
Other Services	M, N, P, Q, and S	69,70,71,72,73,74,75, 77,78,79,80,81,82,94,95,96

7.2 APPENDIX 2

List of dummies

Variable	Description
Industry_Construction and real estate activities	Dummy that assumes 1 if the company belongs to the construction and real estate activities sector, 0 otherwise.
Industry_Energy and manufactoring	Dummy that assumes 1 if the company belongs to the energy and manufactoring sector, 0 otherwise.
Industry_Transportation and storage	Dummy that assumes 1 if the company belongs to the transportation and storage sector, 0 otherwise.
Industry_Other services	Dummy that assumes 1 if the company belongs to the other services sector, 0 otherwise
Industry_Information and communication	Dummy that assumes 1 if the company belongs to the information and communication sector, 0 otherwise.
Industry_Wholesale and retail trade	Dummy that assumes 1 if the company belongs to the wholesale and retail trade sector, 0 otherwise.
Industry_Accommodation and food service activities	Dummy that assumes 1 if the company belongs to the accommodation and food service activities sector, 0 otherwise.
Dimension_Micro	Dummy that assumes 1 if the company is a micro-enterprise, 0 otherwise.
Dimension_Small	Dummy that assumes 1 if the company is a small enterprise, 0 otherwise.
Dimension_Medium-sized	Dummy that assumes 1 if the company is a medium-sized enterprise, 0 otherwise.
Dimension_Large	Dummy that assumes 1 if the company is a large enterprise, 0 otherwise.
GIMP_EXP_Exports	Dummy that assumes 1 if the company only exports, 0 otherwise.
GIMP_EXP_Neither	Dummy that assumes 1 if the company does not export or import, 0 otherwise.
GIMP_EXP_Imports	Dummy that assumes 1 if the company only imports, 0 otherwise.
GIMP_EXP_Imports and Exports	Dummy that assumes 1 if the company imports and exports, 0 otherwise.
Has_Website	Dummy that assumes 1 if the company has a website, 0 otherwise.

7.3 Appendix 3

<u>Correlation Matrix</u>