A Work Project, presented as part of the requirements for the Award of a master's degree in business Analytics from the Nova School of Business and Economics.

The Relationship between Programming Literacy and Economic Growth in Europe

Exploration of New Variables and Their Impat

Carlos Ferrufino – 53276

Work project carried out under the supervision of:

Michael Kummer

20-12-2023

Abstract

The thesis investigates the link between programming literacy and economic growth across European regions, aiming to understand their interplay more comprehensively. Using statistical models, the research predicts the economic effects of programming literacy, revealing positive correlations between literacy and economic prosperity. These findings are extended by an additional panel data analysis. The study also presents a nuanced case study on Poland. Furthermore, by shifting the focus on NUTS-2 level and incorporating new variables, deeper insights into the dynamics between programming literacy and regional economic growth are gained. Lastly, economic resilience and programming literacy are regressed, again displaying positive correlations.

Keywords

[Programming Literacy, Economic Growth, Stack Overflow, NUTS-3 Regions, Tech Hubs]

This work used infrastructure and resources funded by Fundacao para a Ciencia e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

Table of Content

List of Figures

List of Tables

1 Introduction

In today's digital age, the importance of programming literacy in driving economic growth cannot be understated. This Thesis examines the intersection between programming skills and economic development within Europe, with a specific focus on analyzing data at a granular level. By shedding light on how programming proficiency impacts regional economies, this research aims to bridge a critical gap in our understanding of how digital competencies contribute to overall prosperity.

As European nations transition into "new economies" that heavily rely on digital capabilities, it is essential to explore the transformative role of programming literacy. These once niche skills have now become indispensable for sustainable economic growth—comparable to traditional literacy during previous centuries. Through active participation in platforms like Stack Overflow, individuals demonstrate their fluency in coding languages as well as their ability to problem-solve collaboratively—an aspect crucial for thriving in today's interconnected world.

This study seeks answers through several interrelated questions:

- What factors contribute to the emergence of programming hubs, influencing the distribution and growth of programming literacy across diverse regions?
- To what degree does the concentration of programmers in specific regions correlate with economic indicators such as GDP, GVA, and employment rates?
- Can economic factors, such as GDP and employment, be considered significant predictors of programming literacy in specific regions?

The aim of our research is twofold: to examine the relationship between digital skills and regional economic growth, and to provide insights for policymakers to foster environments that nurture these skills.

Employing a comprehensive methodological approach, the thesis analyzes a confluence of programming activity and economic data. It utilizes advanced statistical models including OLS,

Ridge, Lasso, Random Forest, and ARIMA for outlier analysis, aiming to predict and understand the economic impact of programming literacy. This approach not only assesses the direct correlation between programming skills and economic factors but also explores the predictive capability of these variables in forecasting regional economic trends.

This research contributes significantly to our understanding of how programming literacy impact regional economies. By identifying key economic factors that influence programming literacy levels and assessing their potential for forecasting purposes, this study offers valuable insights into strategic decision-making during Europe's ongoing digital transformation. The findings are particularly relevant for policymakers seeking to shape educational curricula aligned with industry needs as well as business leaders aiming to leverage coding expertise as a driver of sustainable economic development.

The paper's structure includes introductory chapters, methodology explanation, data exploration of programming literacy on country level as well as on NUTS-3 level. A regression model analysis for predicting programming literacy based on economic metrics follows. The study delves into outlier analysis and includes five additional papers. These include a panel data analysis to compare the efficacy of linear and machine learning models in economic forecasting, a case study on Poland examining the nexus of programming skills and economic growth, and an analysis of the creation and growth of programming hubs within Poland. The study also contemplates how the inclusion of varied economic and technological factors could refine predictive models. The concluding chapter synthesizes these discussions, probing the potential relationship between programming literacy and economic resilience, thereby providing strategic insights into the sustenance and fortification of regional economies in the face of adversities.

2 Theoretical Foundations

2.1 Definitions

In this subchapter, we define relevant terms to clarify their scope within the context of this paper, ensuring that their meaning is precisely communicated and understood.

Programming Literacy: Balanskat and Engelhardt (2014) define computer programming as „the process of developing and implementing various sets of instructions to enable a computer to perform a certain task, solve problems, and provide human interactivity“ (21). However, when referring to programming literacy, the term expands beyond a mere technical skill limited to few specialists. We define it as a comprehensive understanding of computational processes, a widely held ability integral to communication and societal interaction in the digital age, reflecting its extensive applications and significance (Vee 2013, 46-47).

Economic Growth: Simply put, economic growth refers to “an increase in the quantity and quality of the economic goods and services that a society produces” (Roser 2023). Through this process economic well-being and material quality of life are improved (Investopedia 2023). Various metrics to measure economic growth exist. For the scope of the thesis, our dataset focuses on population size, Gross Domestic Product, Gross Value Added and employment rates.

- *Gross Domestic Product (GDP):* GDP is the total monetary value of all goods and services produced within a country's borders in a specific time period. It encompasses not only the income generated from production but also the aggregate expenditure on final goods and services, accounting for the deduction of imports. However, GDP has limitations for temporal comparisons, as its variations over time may be attributed not only to actual

economic growth but also to fluctuations in prices and purchasing power parities (PPPs). (OECD n.d.-a)

- *Gross Value Added (GVA):* GVA is an economic metric that quantifies the value of goods and services produced in a specific area, industry, or economic sector. It is calculated by subtracting the cost of intermediate consumption from the total output value, representing the net contribution of labor and capital in the production process. This measure also indicates the income generated from these contributions. GVA by activity highlights the unique value added by different sectors like agriculture, industry, and services, often expressed as a percentage of the economy's total value added. (OECD n.d.-b)
- *Employment rates:* Employment rates are indicators that assess the utilization of available labor resources, specifically individuals who are ready and able to work. These rates are determined by calculating the proportion of employed individuals relative to the population within the working age bracket, typically defined as ages 15 to 64. The ratio serves as a gauge for understanding how effectively a society is employing its potential workforce. (OECD n.d.-c)

Nomenclature of Territorial Units for Statistics (NUTS): To measure geographical developments of programming literacy across Europe, our dataset works with a Eurostat classification known as NUTS (Eurostat 2022). This classification divides regions within the European Union using a hierarchical, three-level system based on population sizes. The purpose of this system is to provide a single, uniform breakdown for the collection, development, and harmonization of EU regional statistics. Table 1 captures the 3-level categorization. Each level is designed to provide a standardized comparison between the areas. By utilizing the NUTS classification, our research can achieve a more granular understanding of how programming literacy varies not only between countries but also within different regions of a single country.

Table 1 Categorization of the Nomenclature of Territorial Units for Statistics Regions

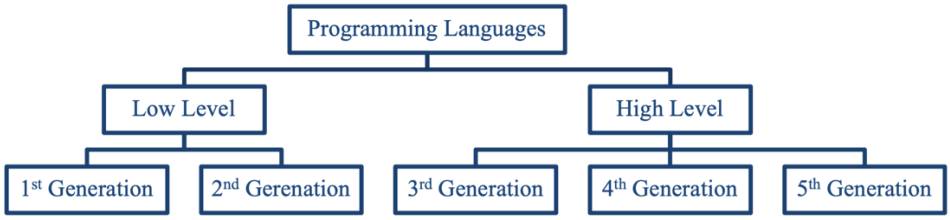
NUTS-1	These are major socio-economic regions. In larger countries, these might be large regions or groups of federal states, while in smaller countries, it can be the country itself. The population of each NUTS-1 region is typically between 3 million and 7 million people.
NUTS-2	This level corresponds to basic regions for the application of regional policies. They can be provinces or smaller states within a country. The aim is to provide a regional breakdown that is consistent for economic analyses. NUTS-2 regions usually have populations between 800,000 and 3 million people.
NUTS-3	These are smaller regions, often for more local administrative purposes, such as counties or larger urban districts. The NUTS-3 level is used for more detailed regional planning and comparisons. These areas usually have populations ranging from 150,000 to 800,000 people.

Stack Overflow: Stack Overflow is a question-and-answer platform for knowledge sharing across programming communities. The platform serves as a repository of information about computer programming. Stack Overflow facilitates collaboration, problem-solving, and knowledge-sharing among individuals, groups, and organizations (Sartore 2022). Established in 2008, Stack Overflow has emerged as a digital forum for programmers worldwide. The regional activity on Stack Overflow is going to serve as our primary metric of programming literacy.

2.2 History of Programming Languages

GitHub's latest October report stated that in 2022 almost 500 primary languages were used on their platform to build software (GitHub 2023). However, the history of programming languages spans over a century of innovation, reflecting the evolving needs of computation, industry practices, and academic research. An effective approach to illustrate these major advancements is by categorizing them into five distinct generations, each characterized by increasing levels of abstraction and sophistication (see Figure 1).

Figure 1 Overview Programming Languages Generations



Note. Adapted from *Generation of Programing Languages* by GeeksforGeeks, 2023. Retrieved from https://www.geeksforgeeks.org/generation-programming-languages/

1st Generation Languages (1940s-1950s) – Machine Languages: The advent of programming languages can be traced back to the 1940s and 1950s with the development of machine languages, also called low-level languages, the most fundamental form of programming languages (CSDH 2022). These first-generation languages are characterized by their use of binary code, directly executable by the computer's central processing unit (CPU), making them fast and efficient with no translator needed (GeeksforGeeks 2023).

2nd Generation Languages (1950s-1960s) – Assembly Language: Emerging in the 1950s and 1960s, the second generation of programming languages introduced assembly languages. These languages represented a significant step forward, utilizing mnemonic codes that were more accessible than the binary code of the first generation. Assembly languages allow programmers to write instructions in symbolic code, which an assembler would then convert to machine code. (Medewar n.d.; GeeksforGeeks 2023)

3rd Generation Languages (1960s-1970s) – High-Level Languages: The third generation of programming languages, emerging around the 1960s, marked a significant evolution in the field, as they are machine-independent, meaning one can run the same code on different machines (Jain 2018). However, this means that for translating them into machine language a compiler is needed, and different machines use different compilers. Therefore, developers must use compilers accordingly (Medewar n.d.). These high-level languages, such as FORTRAN,

PASCAL, ALGOL, and COBOL, shifted closer to human language in their syntax, including English-like keywords and hence greatly enhancing the efficiency and accessibility of programming (Chandra n.d.; Jain 2018).

4th Generation Languages (1970s-1980s) – Very High-Level Languages: The fourth generation of programming languages, arising in the 1970s, focused on further increasing programmer productivity and efficiency. These very high-level languages, exemplified by SQL and MATLAB, are often domain-specific, targeting areas such as database management and report generation as well as scientific computation (Chandra n.d.; GeeksforGeeks 2023).

5th Generation Languages (1980s-Present) – Artificial Intelligence Languages: The fifth and most recent generation of programming languages make use of visual tools to help develop a program. A popular example for this is Visual Basic. This generation centers on artificial intelligence (AI) and constraint programming. This approach is particularly prevalent in AI and machine learning research. While 5th generation languages are crucial in these advanced fields, their use is not as widespread in general-purpose programming, reflecting their specialized nature. (Jain 2018)

6th Generation? – In the forthcoming years, we expect programming languages to evolve in response to technological advancements, industry challenges, growth of automation, AI and machine learning as well as the increased need for collaboration due to increasing (semi-) remote work and lastly economic considerations (Schaefer 2023). The next decade will likely see a trend towards specialization, with programming languages tailored to unique domains such as quantum computing and augmented reality.

The retrospective of programming languages, from their genesis in machine code to the sophisticated AI and domain-specific languages of today, sheds light on the exponential

trajectory of programming literacy. It is a journey marked by the relentless pursuit of efficiency, abstraction, and the bridging of the human-computer divide. As each generation of languages abstracted complexity and made programming more accessible, the skills to harness them became increasingly prevalent. Today, programming literacy is not only widespread but also integral to innovation and economic development, reflecting the demand for digital competency in an increasingly tech-driven world. The future promises even greater accessibility, with languages evolving to meet the needs of new technological frontiers.

2.3 Literature Overview

In the contemporary global economy, the role of programming and digital skills, collectively referred to as Information and Communications Technologies (ICTs), has emerged as a critical factor influencing economic growth. Numerous studies worldwide have explored the correlation between digital skills and the economic development of specific countries. Global research consistently confirms a correlation between the development of digital skills and the economic growth of nations. Scholars have explored this linkage using diverse methodologies and frameworks (Sein et al. 2018).

For instance, Laitsou, Kargas, and Varoutas (2020) conducted a comprehensive study spanning a 20-year period, with a special focus on the economic crisis period (2008–2016) in the European Union. Using growth accounting methodology and regression analysis the study revealed that investments in ICT, especially during economic crises, play a pivotal role in fostering economic growth in the Eurozone, and specifically growth of GDP in Greece in the investigated timeframe. This suggests the relation between economic growth and programming literacy this paper investigates.

Urbančíková, Manakova, and Bielcheva (2017) delved into the European context, specifically focusing on the impact of digital skills on the local economy in Slovakia. They assessed the association between education, income and household type and digital skills. The paper proved that those factors are significantly affecting digital literacy of the society, rather than the size of the economy or the economic growth of the technological industry in the country.

Building on this, Bălăcescu (2019) investigated the relationship between economic growth and digital skills across EU member states. Using GDP per capita and digital skills, defined as BOSE index (basic overall digital skill), the research identified similarities and dissimilarities among EU countries. The findings provide valuable insights for decision-makers, proving that development of digital skills can hinge the growth of European economies.

As evidenced by the research, the symbiotic relationship between programming literacy and economic growth is clear, with digital skills serving as both a catalyst for and an indicator of a nation's economic vitality. Our study builds upon this foundation to empirically explore the specific correlation between programming literacy and economic development in NUTS-3 regions of six European countries.

2.4 The Rise of European Tech Hubs

In 2019, Germany led the European landscape with over 901,400 professional developers, followed by the UK with around 849,600. Other significant contributors included France, Italy, the Netherlands, and Poland (Statista 2019). Highlighting Europe's emergence as a technological leader, recent data reveals a significant and sustained rise in its developer workforce. With around 6.1 million professional developers in 2019, Europe has overtaken the United States in this metric. This marked growth in developer populations across European regions not only cements Europe's position at the forefront of global tech innovation but also

mirrors the burgeoning concentration of tech professionals within rapidly growing tech hubs (Atomico, Slush, and Orrick 2019).

Central to this development has been the role of government policies and incentives. European governments have increasingly recognized the importance of the digital economy, enacting policies that encourage innovation and entrepreneurship. The European Commission's "Digital Single Market" strategy, for instance, aims to open up digital opportunities for people and businesses and enhance Europe's position as a world leader in the digital economy, introducing migration friendly policies to attract global talent (European Commission 2013). Such policies create an environment conducive to the growth of tech hubs by providing the necessary support structures for startups and established tech companies alike. Government regulations, digital literacy campaigns, and infrastructure investments have been instrumental in establishing conditions conducive to technological innovation. Further insights are provided by the EU Innovation Scoreboard's 2021 "Annual Report on European Tech Hubs," particularly regarding the post-COVID-19 landscape. The report emphasizes innovation's role in tech hub evolution and notes how the European Commission's updated industrial strategy, focusing on SMEs and startups, is reshaping the technological landscape.

Hence, availability of venture capital and other forms of investment is another key driver for the growth of tech hubs. This influx of capital acts as a magnet for these startups, which are key to drawing in a pool of skilled programmers. Notably, hubs with dense concentrations of investors, especially those with a keen interest in cutting-edge technologies, often coincide with areas rich in research and development institutions. These include universities, such as the ETH in Zurich, tech incubators, and specialized medical facilities. (Davis et al. 2023) These factors combined are prevalently given in urban regions of economically developed countries, such as Berlin, London, or Stockholm. Finally, soft factors, as quality of life and cultural diversity

additionally impact a regions attractiveness to developers, enriching its location with programming expertise (Varley 2023).

3 Methodology

In the forthcoming analysis, we will conduct an empirical examination of programming activity across six European countries, applying quantitative analyses. By integrating economic indicators with programming literacy data, we aim to explore the extent to which a country's economic development serves as a predictive measure for its programming rates, thus validating theoretical assertions and shedding light on the interplay between economic vitality and a population's programming skills. Rigorous data collection from sources such as Stack Overflow and Eurostats, analyzed using Python, will help address potential biases and ensure a robust exploration of the relationship between programming literacy and economic factors in European regions.

3.1 Main Dataset

The main dataset contains information on programming activity and economic factors over the years by NUTS-3 region. Regarding programming activity, Stack Overflow metrics *question, answer, comment, upvote,* and *downvote counts* were used. Economic indicators include *employment, gross value added (GVA), gross domestic product (GDP)* and *population*, covering the countries Spain, France, Italy, Turkey, Sweden, and Switzerland.

3.2 Data Sources

The data on Stack Overflow activity by NUTS-3 regions is provided by Stack Overflow. The economic data is curated from reliable online sources, such as Eurostats and national statistics databases. Merging additional data requires an identifier for accurate integration. NUTS-3

names from the main dataset are used for this purpose. Discrepancies with Eurostats data, which uses NUTS-3 identification codes, require matching the regions with corresponding codes. An official Excel sheet from Eurostats that maps the regions to their codes facilitates this process. Differences in region names between the main and Eurostats datasets are solved by creating a mapping file, ensuring no data loss during linkage. Economic factors are chosen according to the ones that are defined as regional economic accounts by the European Commission (ESA 2010), namely employment, gross domestic product (GDP) at current market prices, gross value added (GVA) at basic prices and the average annual population to calculate regional GDP data. After merging this data with the main dataset, a subsequent search for missing values reveals significant gaps in economic data for Switzerland and Turkey. Missing data for Switzerland is collected from the Federal Statistics Office (FSA) and for Turkey from the Turkish Statistical Institute. The further handling of missing data is discussed in the following chapter.

3.3 Data Preprocessing

Data collected from various sources, especially in a multifaceted field that combines economic factors and programming activities across NUTS-3 regions, is often raw, unstructured, and may contain inconsistencies or errors. Data cleaning is an essential step to transforming this raw data into a clean, structured format suitable for analysis.

Given the temporal and regional diversity of the dataset, ensuring comparability across different time periods and regions is crucial. Preprocessing standardizes the data, allowing for meaningful comparisons and trend analysis. Moreover, in real-world datasets, the presence of missing values is a common challenge. It is necessary to implement strategies to handle missing data, either through imputation or removal, depending on their impact on the overall analysis.

In particular, tackling the challenge of rows with missing years, the most relevant time period for analysis is investigated, considering the focus on programming activity. This approach allows for the possibility of excluding certain years where data is unavailable.

Figure 2 Stack Overflow Activity across the Years based on ES, IT, FR, SE & CH data

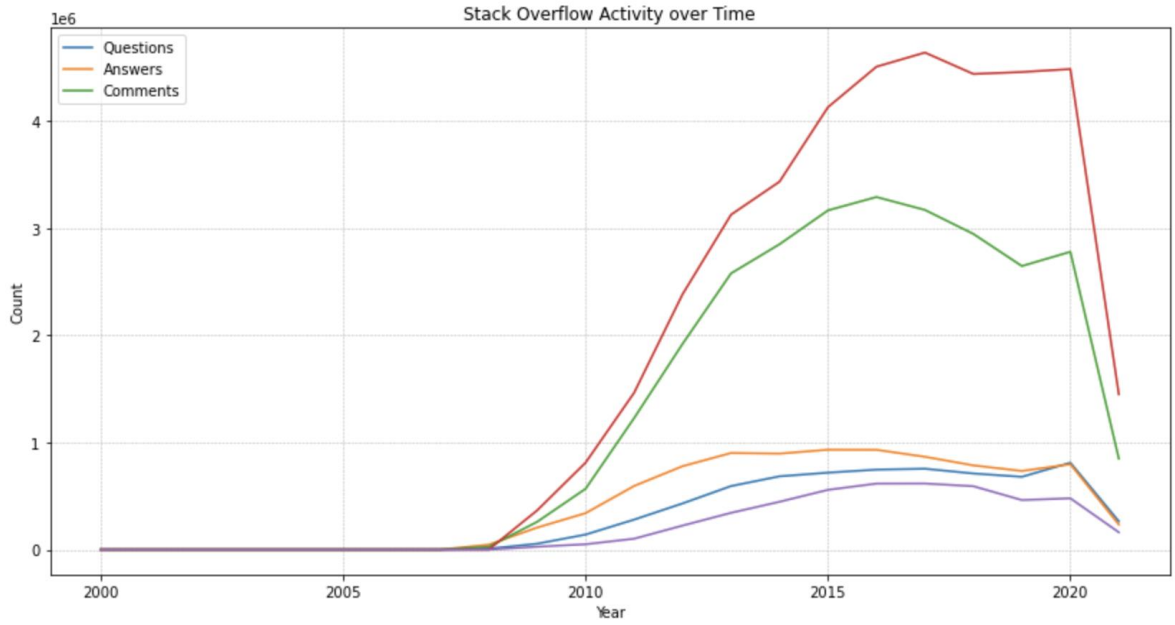


Figure 2 shows different Stack Overflow activities plotted over time. The Figure reveals no data before 2008 and a notable decrease in 2021, suggesting the data collection occurred within that year, omitting complete data for 2021. Consequently, the analysis period is set from 2008 to 2020. This adjustment significantly reduces the number of missing values for the economic data. However, gaps remain, notably in employment data for Switzerland and Turkey, and in GDP, GVA, and population data for Switzerland.

Handling missing data for Switzerland involves solving several different problems. First, to align with Eurostats' average annual population data for other countries, it is necessary to calculate the average yearly population for each Swiss canton. These calculations match the values for Switzerland in the Eurostats data. Second, the Swiss data for GDP and GVA, presented in millions of Swiss francs, differs from the other countries' data in millions of euros. To resolve this, the average CHF/EUR exchange rate per year, provided by the European

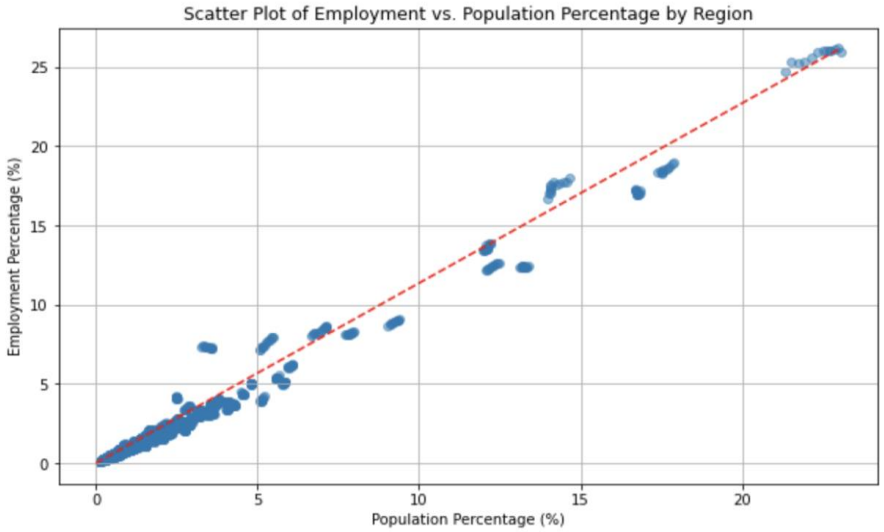
Central Bank Data Portal, is used to convert Swiss values into euros. Last, employment data is only available from 2010 onwards, resulting in a lack of information for the initial two years (2008 and 2009) of the defined analysis period. These values are backwards filled with the respective observations of the year 2010.

Regarding employment data for Turkey, it is not possible to find any reliable data that is grouped by NUTS-3 regions. The granularity of the data provided by the Turkish Statistical Institute is monthly and on a national level. The monthly data can easily be transformed by calculating the yearly average, but the fact that the data is only on a national level imposes a more difficult problem.

It is hypothesized that the distribution of employed persons per year per NUTS-3 region relative to the national level is similar to the distribution of the population. Thus, missing data can be calculated by using this distribution. To test this hypothesis, it is necessary to show that this holds true for the other five countries in the dataset. A statistical analysis is performed, and the distributions are plotted against each other.

In Figure 3 the percentages of employment and population by regions relative to the countries are plotted against each other. By observing the relationship between the distribution of employment and population across the different regions one can infer that there is a high positive correlation which suggests that regions with a high population density tend to have high employment rates. The red dashed line represents the line of equality, which means that it's the point where both distributions are equal. Most of the points lay very close to the line. Additionally, a Pearson Correlation Coefficient of 0,989 is calculated. This further suggests a very highly positive correlation between both indicators. In other ways, as the population increases in a NUTS-3 region, the employment rate tends to increase linearly and vice versa.

Figure 3 The Relation between Employment and Population in Turkey



However, there are concerns as to whether it makes sense to fill the values in this way. Firstly, the underlying assumption of a uniform relationship between population and employment across all regions could be problematic because variations in employment patterns due to economic, social, or geographical factors might not be captured. This means future models might oversimplify the employment landscape. Secondly, outliers in the data are important because they can represent regions with unique employment characteristics. For instance, a region with a small population but high employment due to a major industrial center would be an outlier. By aligning employment strictly with population, future models might miss these nuances. The outliers, rather than being anomalies to be smoothed over, could be essential for understanding regional economic dynamics. Lastly, the relationship between population and employment is not static. It can change over time due to various factors like economic policies, technological advancements, and social changes. The assumption that past relationships will hold in the future can be problematic, especially in regions undergoing rapid change. Generally, if population and employment are highly correlated, using both as separate predictors in a regression model could lead to multicollinearity, where it becomes difficult to separate the

effect of one predictor from the other. This can inflate the variance of the coefficient estimates and make the model less reliable.

To address these concerns, it is decided to drop the Turkey data as of chapter 4.1.3 for the prediction analysis, because missing values cannot be properly handled by the models. A description of the final dataset and its columns is provided in Table 2.

Table 2 Main Dataset columns and their description

Column	**Description**
NUTS-3_code	NUTS-3 region code
year	Time period of the data
NUTS-3_name	NUTS-3 region name
country	NUTS-3 region country
questioncount	Number of questions posted from the region on Stack Overflow
answercount	Number of answers posted from the region on Stack Overflow
commentcount	Number of comments posted from the region on Stack Overflow
upvotecount	Number of upvotes for the region on Stack Overflow
downvotecount	Number of downvotes for the region on Stack Overflow
EMP (THS)	Employment in thousands
GDP (MIO_EUR)	Gross Domestic Product in million euros
GVA (MIO_EUR)	Gross Value Added in million euros
POP (THS)	Population in thousands

3.4 Quantifying Programming Literacy

For all the further modeling and analysis, a singular metric for programming literacy needs to be established. Weighting the different forms of Stack Overflow activities could be a meaningful approach in quantifying programming literacy, as these activities directly reflect the engagement levels of programmers in each region. By weighting these metrics, it is possible to differentiate between passive and active forms of engagement (like voting versus posting or

answering questions), thus identifying regions with not just high traffic but also high participation and expertise in programming.

Questions, answers, and comments on Stack Overflow are tangible indicators of an active programming community. They represent not just the problems and challenges faced by programmers but also their willingness to share knowledge and collaborate on solutions. Posting questions is a primary indicator of active problem-solving and learning in the programming community. It shows a need for knowledge and a willingness to seek assistance, which is fundamental in a vibrant and growing programming hub. A high volume of questions suggests a significant level of engagement in programming activities. Therefore, a weight of **40 percent** is assigned to Stack Overflow questions.

Providing answers is equally crucial, as it demonstrates expertise and a community's capacity to support its members. A region with a high number of answers indicates not only active participation but also a certain level of proficiency and knowledge-sharing culture in the programming community. Thus, answers are given a weight of **40 percent**.

Comments, while important for clarifying questions and answers, are generally less indicative of programming activity than the questions and answers themselves. Thus, comments are given a weight of **10 percent**. Upvotes and downvotes are indicators of community engagement and the quality of content being produced. However, they are more passive forms of participation compared to asking questions, providing answers, or commenting. Therefore, they are assigned a lower weight of **5 percent** each. To summarize, we define programming activity as a weighted summation of questions answered (40%), answers given (40%), number of comments (5%) as well as number of up- and downvotes (5%).

It is important to note that these weights are solely based on a subjective assessment of the importance of the different Stack Overflow activities. Although there is no scientific research that could justify these choices, we think that this methodology is reasonable.

4 Geographical Analysis

In this chapter of the thesis, we delve into a geographical analysis to explore how programming activity is distributed and how it correlates with economic factors. The focus lays on identifying the most important programming hubs and understanding the density of programmers in relation to population and GDP across various regions. This analysis also lays the groundwork for the subsequent chapter, where we will probe deeper into predictive variables for high programming activity in urban areas. Understanding these predictors is crucial for identifying the characteristics that make a region conducive to a thriving programming community.

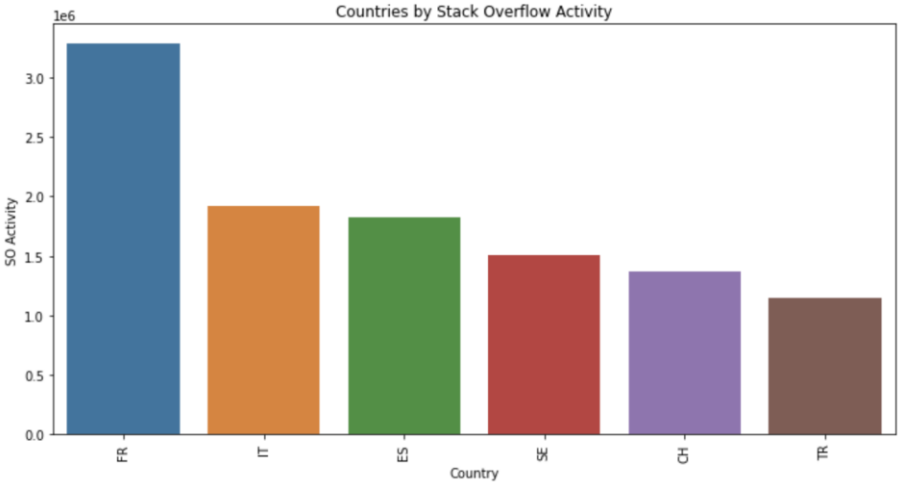
4.1 Identification of Programming Hubs

To identify the most important programming hubs by NUTS-3 regions it is crucial to establish certain criteria that define what constitutes as *important* in this context. Given the information that is included in our dataset, there are several approaches and steps to consider. First, the weighted programming activity is investigated on both country- and region-level. Then, the growth of this activity over time is analyzed to see if it influences the results.

4.1.1 Country-Level Analysis

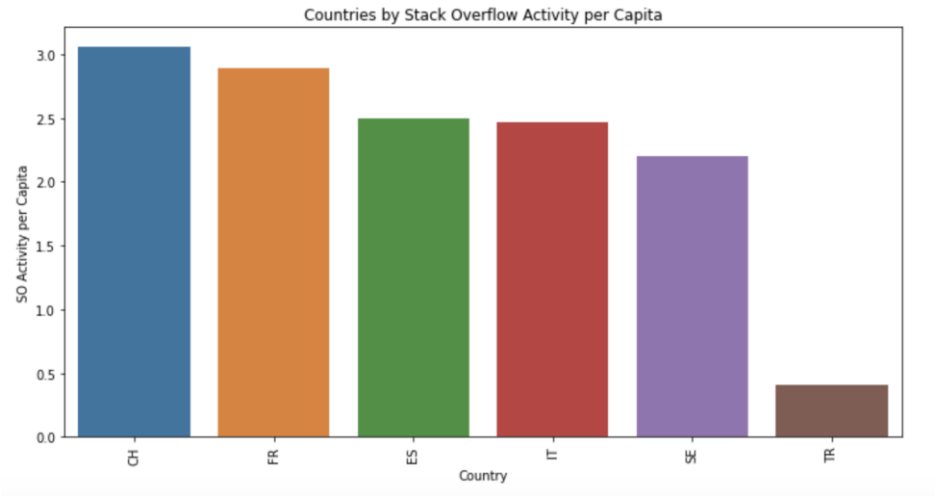
We start by investigating how the programming activity differs on a country level. When plotting the activity and activity per capita for the six countries, significant differences in their ranking can be observed.

Figure 4 Amount of Stack Overflow's Activity across the analyzed countries.



In Figure 4, France (FR) appears to be the leading country in terms of total programming activity, reaching more than three million. Italy (IT) has the next highest level of activity, followed by Spain (ES), Sweden (SE) and Switzerland (CH). All four have lower activity levels than France but are quite close to each other. Turkey (TR) has the least activity among the presented countries, with a value just under one and a half million.

Figure 5 Size of Stack Overflow's Activity per Capita across analyzed countries.



In Figure 5 the y-axis represents Stack Overflow activity per capita, adjusting the activity levels in proportion to the population size of each country, which is a more accurate depiction of the programming literacy across NUTS-3 regions. Switzerland now appears to have the highest level of activity per capita, which suggests that while the overall activity in Switzerland might

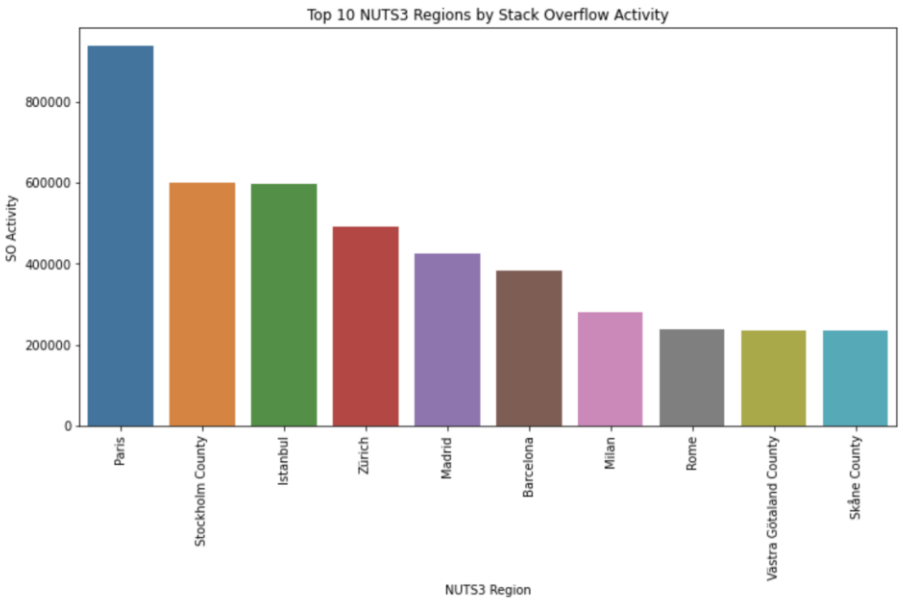
be lower, the engagement per person is higher compared to other countries. France has the second highest score, thus performing very well in both total activity and activity per capita. Spain, Italy, and Sweden still show significant activity per capita. Turkey not only remains the country with the lowest activity per capita, but also shows a huge discrepancy towards the other countries with less than 20% of Sweden, which has the second lowest activity per capita.

Comparing the two Figures, it's evident that this adjustment provides a more nuanced understanding of Stack Overflow activity, suggesting that while some countries may not have the highest overall numbers, their smaller populations are very active on the platform.

4.1.2 NUTS-3-Level Analysis

When performing the same analysis on a NUTS-3 level, there are similar differences that can be observed.

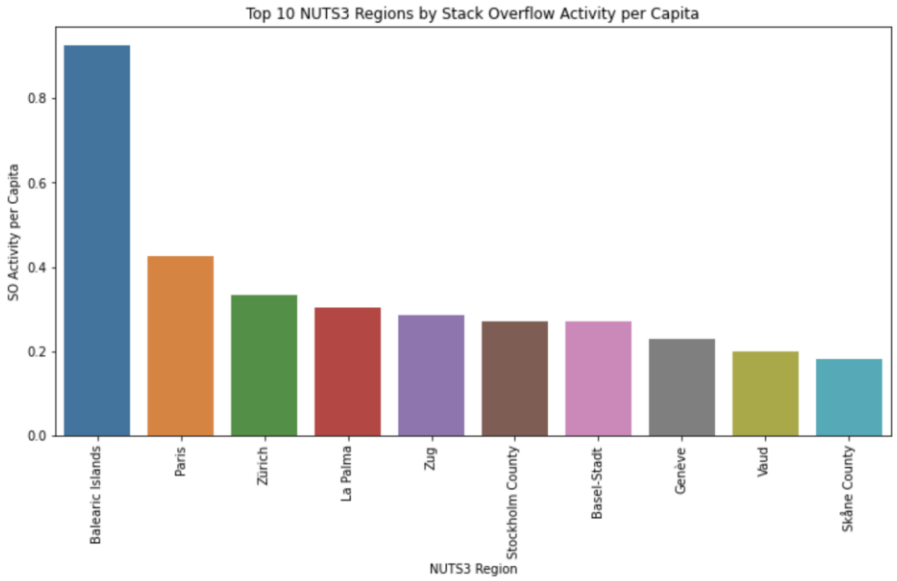
Figure 6 Top 10 NUTS-3 Regions based on the Stack Overflow's Activity Size



With Figure 6 visualizing total programming activity in absolute numbers, we can observe that Paris has the highest Stack Overflow activity, followed by Stockholm County and Istanbul. Except for Switzerland, all of the capitals of the examined countries are represented. This could

be due to various factors such as a larger population of programmers, a strong tech industry, or a combination of both.

Figure 7 Top 10 NUTS-3 Regions based on the Stack Overflow's Activity Size per Capita



The second graph (Figure 7) changes the perspective again by showing programming activity per capita. This measurement gives us an insight into how active the programming community is relative to the size of the population in each NUTS-3-region. Interestingly, the Balearic Islands lead in per capita activity, despite not being the top in overall activity, indicating a highly active programming community relative to its population size. Paris, while still high in per capita activity, is not as dominant as it was in the total activity graph. With Zürich, Zug, Basel-Stadt, Genève and Vaud, Switzerland seems to have many hubs with high engagement per capita. The average programming activity per capita across all regions is 0.035 which highlights the dimension of activity in the displayed programming hubs.

Major cities like Paris, Zurich, and Stockholm County show high activity in both total and per capita terms, indicating both a high number of users and a high engagement relative to population. The presence of smaller regions such as Zug and Basel-Stadt near the top of the per capita graph suggests that smaller populations can have a very active core of Stack Overflow

users. Differences in the rankings between the two graphs highlight the impact of population size on Stack Overflow activity measurements. Some regions may seem less active overall but are very active on a per capita basis.

Overall, the activity per capita graph is crucial for understanding the true level of engagement within smaller populations, while the total activity graph is more reflective of the raw number of interactions on Stack Overflow, which can be influenced by larger populations.

4.1.3 Key Findings - Growth Over Time

So far, only the programming activity in the overall time period from 2008 until 2020 has been investigated. Another approach may be to analyze the growth of activity over time. This is crucial for identifying emerging hubs and understanding the dynamics of programming engagement in different regions. A region experiencing a rapid increase in Stack Overflow activities might indicate a surging interest in programming or a growing IT sector. This time-based analysis would help to distinguish established hubs from fast-growing, emerging ones. The results can be valuable for companies and educators to understand where the tech community is growing and to target these areas for investments, recruitment, or expansion of educational programs.

The Compound Annual Growth Rate (CAGR) is computed as a metric for programming activity over time. It measures average annual growth over a given period.

$$CAGR = \left(\frac{\text{Ending Value}}{\text{Beginning Value}} \right)^{\frac{1}{\text{Number of Years}}} - 1$$

When looking at Figure 8 and Figure 9 can observe that seven out of the ten regions displayed in both graphs are in Italy. All these regions have an average annual growth rate above 80%, which is twice as high as the average growth rate per region of 39%. This indicates that Italian

regions experience a significant increase in Stack Overflow activity over time. When computing the CAGR and the CAGR per capita on the country level, Italy shows the second highest overall growth (41%) just after France (44%) which confirms this indication.

Figure 8 Top 10 NUTS-3 Regions based on the GDP Growth defined as CAGR

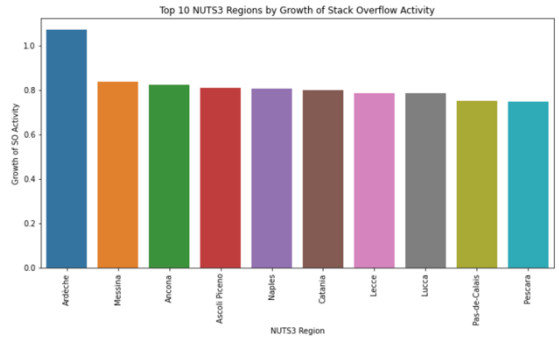
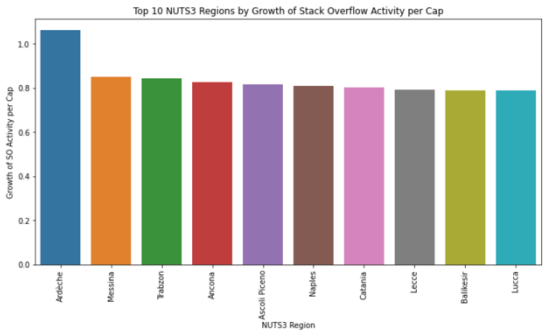


Figure 9 Top 10 NUTS-3 Regions based on the GDP Growth defined as CAGR per Capita



The order of regions by growth rate appears quite similar in both graphs, suggesting that the growth in activity correlates with growth in activity per capita. Regions like Ardèche, Messina and Trabzon lead in both graphs, highlighting that they are not just becoming more active in absolute terms but are also seeing significant engagement growth relative to their population. The presence of the same regions in both graphs suggests that the factors driving the growth of Stack Overflow activity in these areas are affecting the population uniformly, rather than being concentrated in a subset of the population (such as might be the case if growth were driven by migration of tech workers into the region).

What is interesting is that none of the NUTS-3 regions that showed high total programming activity and high activity per capita are represented in these graphs. This could be due to the fact that more established programming hubs don't experience the same level of growth as smaller emerging hubs. All in all, programming activity in absolute terms offers a raw count of how much coding is happening in an area, pointing to regions with the most vibrant tech ecosystems. Per capita Figures adjust this activity for population size, identifying places where programming is not just common, but a central part of the environment. On the other hand, the

growth of programming activity merely shows the increase or decrease over time, which doesn't necessarily reflect the current size or density of programming hubs, making it a less immediate indicator of where the most active communities are.

4.2 Programming Density based on the "Programmer" Definition

Building on the prior chapter's identification of principal programming hubs within Europe, this section delves into a granular analysis of programming density and its influence on Gross Domestic Product (GDP) across regional landscapes. Here, we define programming literacy as the breadth of programmer presence, gauged by Stack Overflow activities. Hence, 'programmers' are those who demonstrate consistent engagement with the platform, indicative of programming literacy.

We derived our definition of a 'programmer' from the analysis of Stack Overflow user activities, such as posting questions, providing answers, and engaging with the content through votes and comments. This data, collated by year and NUTS-3 region, is enriched by insights from the Stack Overflow Developers Survey 2021, which offers detailed accounts of interaction frequency and professional behaviors. This dual-faceted approach enables a nuanced understanding of the term 'programmer.'

Employing clustering techniques, specifically K-modes clustering, we analyzed the 2021 survey data to identify distinct patterns of engagement among users. K-modes, an algorithm particularly adept at handling categorical data, sorts users into clusters based on their activity patterns and professional attributes (Chaturvedi et al. 2001). This method's strength lies in its ability to provide clear interpretability of complex, categorical datasets. Through this process, we are able to delineate a multifaceted definition of a programmer, capturing the essence of programming literacy as represented in diverse user interactions on Stack Overflow.

4.2.1 Definition of a Programmer based on Stack Overflow Activity Data and Stack Overflow Developers Survey

The initial analysis using K-modes clustering unveils three clusters among 36 combinations of *Profession* and *Frequency*. K-modes clustering identified clusters, representing personas based on respondents' professional behaviors and Stack Overflow participation frequency. Based on these clusters, it is possible to define the following personas:

1. Cluster 0: Comprising over 50% of respondents, Cluster 0 signifies the most common behavior – individuals in the early stages of learning programming or pursuing it as a hobby. These users, not yet professionals, interact either daily or monthly, likely reflecting different proficiency levels.
2. Cluster 1: Encompassing around 25% of respondents, Cluster 1 comprises users who write code occasionally as part of their work. Less active on Stack Overflow, they engage in discussions on a monthly basis or even less frequently.
3. Cluster 2: The smallest cluster, Cluster 2, likely represents professionals working daily as programmers. Their frequent engagement in Stack Overflow discussions, a few times per week, marks them as highly involved platform users.

In summary, the clustering results offer valuable segmentation, laying the groundwork for more detailed analysis and targeted strategies based on identified personas. Each of the defined personas are described in the Table below (Table 3). They are segregated based on the clustering results (modeled on the Profession and Frequency variables).

To estimate the number of programmers in NUTS-3 regions, specific activities in the main dataset are considered for each persona. The variables *answercount, questioncount, upvotecount, downvotecount,* and *commentcount* are used to identify and quantify each persona's engagement, with adjustments made for the frequency of their participation.

Table 3 Defined Personas based on the Clustering Modelling of Stack Overflow Survey Answers

Persona	**Expert**	**Enthusiast**	**Late Adopters**
Profession	Programmer	Not a programmer, but can code	Student, learning how to code
Frequency of Participating in SO	Few times per week	Once per month	Almost daily
Type of Interaction	Answers	Questions	Reactions (votes, comments)

New variables are created to estimate the number of total programmers in the region. Each type of interaction is assigned to a specific persona, as presented in Table 3 and then adjusted based on the frequency they present to calculate the final number of Experts, Enthusiasts and Late Adopters in the NUTS-3 region. Then, the total number of programmers is calculated per region and year as a sum of all personas. This comprehensive approach allows for granular interpretation and application of these personas to real-world data. Interestingly, some users are not professional programmers, supporting our definition that programming literacy doesn't have to be tied to being a professional programmer.

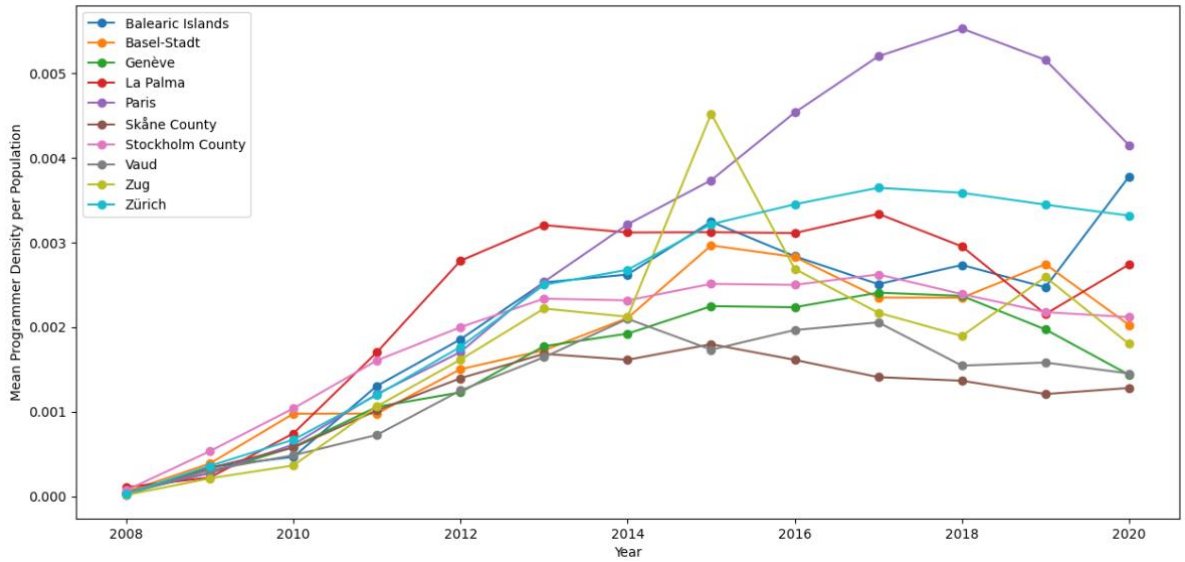
4.2.2 NUTS-3-level Programmers Density Analysis

Based on the established definition on previous stage, the total number of programmers was calculated. It emerges as a significant factor examined in this chapter within the context of key economic indicators such as GDP, Employment, and Population. The objective of this chapter is to enhance comprehension of the correlation between the estimated number of programmers and regional economic development.

The first analysis in Figure 10 reveals that Paris emerges as the leader with the highest estimated programmers count, followed by Stockholm County, Zurich, Madrid and Barcelona, indicating a high level of programming literacy. In 2020 the Balearic Islands stand out as one of the cities with the highest number of programmers per capita, after Paris. This defies expectations for a

region typically recognized for tourism rather than technology (Pons and Rullan 2014). Notably, other high-density locations align with Europe's technological hubs (Paris, Stockholm), attracting programmers, most probably due to the presence of tech companies and well-developed business environment. This proves that most programmers tend to live and work in big technological hubs, however presence of regions like Balearic Islands and La Palma and their rapid growth in 2019-2020, should indicate the existence and growth of the remote work trend across programmers in Europe.

Figure 10 Programmers Density by Population by NUTS-3 Region from 2008-2020



Summarizing findings from Figure 10, while the number of programmers grew over time and therefore programming literacy, the growth concentrated within established tech hubs, indicating their sustained attractiveness to this demographic.

Further analysis of Programmers Density per GDP, presented on Figure 11, has shown that Balearic Islands boast the highest programmer density per GDP. Later analysis has shown that Balearic Islands and La Palma demonstrate high programmer density despite lower GDP per capita, suggesting inefficiency of tech sectors or lack of local companies leading local growth. Another reason could be that their popularity for remote work resulted in the emergence of a new industry, accounting for a considerable proportion of total GDP. Paris and Zurich together

with other NUTS-3 regions have significantly lower density of programmers per GDP, which can be explained by their diverse economies, showcasing their economic strength through various industries. Noteworthy deviations, such as in Spain (Balearic Islands and La Palma), could be traced back to the impact of remote work trends and housing dynamics on programmer distribution (Benitez and Castillo 2023). However, the nature of the data, which presents Stack Overflow activity makes it difficult to understand whether identified programmers are long term habitants of those regions or not, determining their status as digital nomads.

Figure 11 Most Programmer Dense per GDP NUTS-3 Regions across the Years

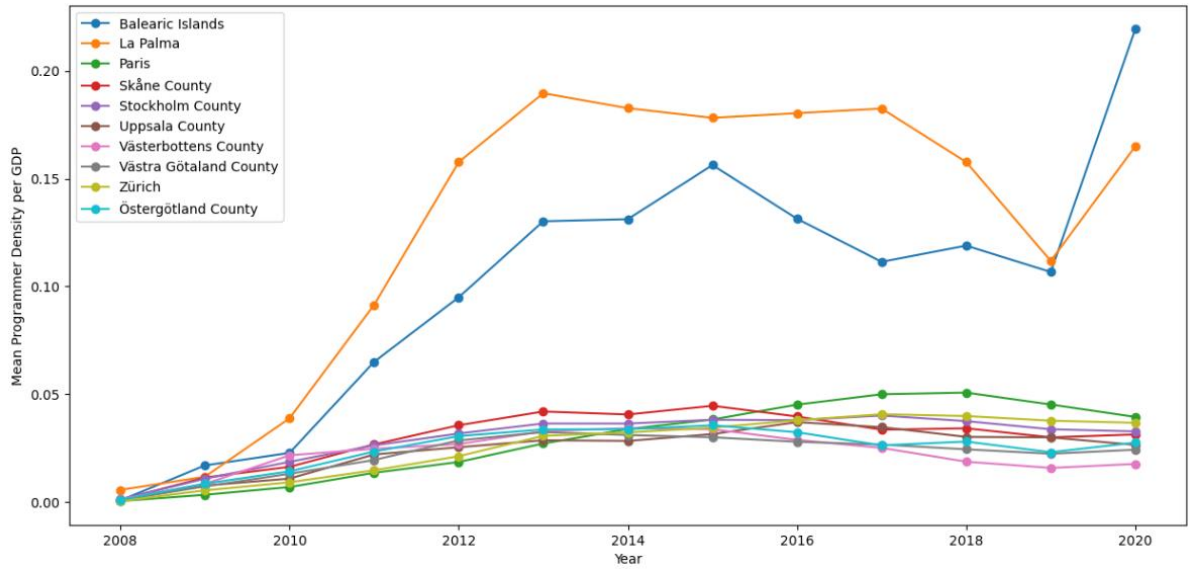
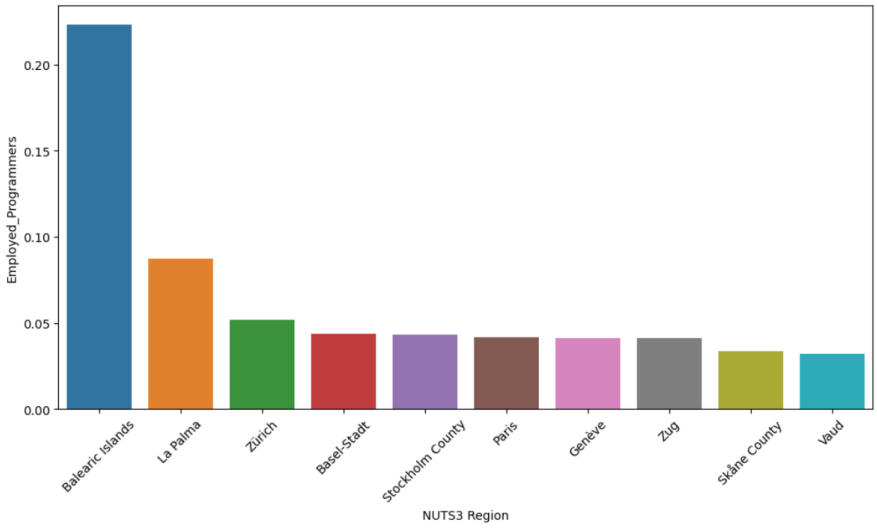


Figure 12 presents the ratio of programmers to the total employed population per region and provides additional insights. Balearic Islands and La Palma continue to showcase a strong tech industry presence. Major economic hubs like Zürich, Basel-Stadt, Stockholm County, Paris, Genève, and Zug exhibit noteworthy ratios, indicating a significant tech workforce contributing to the overall employment landscape. These regions often serve as economic centers with diverse industries. Paris has a high density of programmers per population (Figure 10), but a lower density per employment suggests several conclusions. As a major global city, it likely has a diverse economic landscape with employment opportunities in various sectors. While it attracts programmers, they may not represent the majority of the employed workforce. Other

industries and professions might contribute more significantly to the overall employment makeup. Also, Paris has a much higher employment rate than the rest of the regions analyzed (89% employment rate, with dataset average of 53%), which affect the ratio of programmers when compared to total workforce of Paris, the most employed region across analyzed NUTS-3.

Figure 12 Ratio of Programmers per Employment based on the top NUTS-3 Region



In conclusion, this comprehensive analysis uncovers the dynamics of programming literacy (defined as the number of programmers) and economic development in terms of GDP, employment rates and population size. The cities that observe high programmers' density per GDP (Figure 11), tend to also have higher ratio of programmers in all employed population (Figure 12). This observation suggests that cities with a higher literacy rate are often characterized by more tech-driven economies. These tech-driven economies tend to attract talent in the field of programming, thereby reinforcing this trend and creating a positive feedback loop. They also tend to invest in infrastructure and education. The latter might explain Paris' higher literacy to GDP proportions in comparison to employment, as it is a popular destination for tech-savvy individuals enrolled in quality educational opportunities.

As mentioned previously, disparities in programmer distribution between regions such as the Balearic Islands and La Palma may be significantly influenced by evolving trends in remote work and local housing dynamics, especially since the pandemic. These factors potentially affect the concentration of programmers in various areas, especially in relation to urban centers. This phenomenon presents an intriguing case study, suggesting that these regions could serve as valuable outliers for more in-depth analysis to understand how remote work preferences and residential choices are reshaping the geographic distribution of technology professionals.

5 Predicting Programming Literacy

This chapter is dedicated to establishing the predictive power of economic variables over programming literacy at the NUTS-3 regional level. Informed by our preceding analysis, we posit a correlation between regional economic health and programming literacy, defined here as engagement in Stack Overflow discussions. Such literacy serves as the dependent variable in our predictive models, casting the efficacy of economic indicators as a barometer for forecasting programming engagement.

Tackling the prediction of programmer density within urban areas, we build on the comprehensive preprocessing and methodological foundation set in earlier sections. Four predictive models have been crafted and critically assessed: Ordinary Least Squares (OLS), Ridge, Lasso, and Random Forest. Our mission is to calibrate these models to not only forecast programming activities with high precision but also to ensure an optimal balance between accuracy and generalizability.

5.1 Feature Selection

To develop a strong basis for our models, the process is initiated by finding prospective predictors that have significant correlation with programming density. Correlation analysis

against programming activity indicates that economic indicators such as GDP, employment rates, and Gross Value Added (GVA), as well as demographic variables including population dynamics may be valuable predictors for the model. As a starting point an Ordinary Least Squares (OLS) model is chosen as the baseline for our cross-sectional 2016 analysis. OLS (Ordinary Least Squares) is a statistical method used to find the straight line in linear regression by minimizing the sum of squared differences between observed and predicted values (Lewis-Beck 1980, 9-12) and therefore is most suitable for predicting values that share linear relation.

To determine the most effective set of predictors, we conducted a comparative analysis using various models. The initial model used GDP (in millions of euros) as a standalone measure of economic productivity. Subsequently, we disaggregated GDP into two distinct variables: GDP per capita and Population. This approach allowed us to scrutinize their individual contributions to the predictive model. Since GDP per capita is derived from the total GDP and population size, it was imperative to assess the interplay of these factors within our model. We utilized coefficient and p-value analysis to gauge the impact of each predictor, with coefficients indicating the strength and direction of the variable's effect on programming literacy, as delineated in Tables 4 and 5.

Table 4 Independent Variables Statistics based on the OLS Regression (without GDP per capita)

	coef	**std err**	**P>|z|**
const	-385.8516	653.625	0.555
EMP (THS)	16.4822	26.450	0.533
GDP (MIO_EUR)	0.2331	1.516	0.878
GVA (MIO_EUR)	0.1021	1.608	0.949
POP (THS)	-10.8342	9.807	0.269

GDP (MIO_EUR) has a positive coefficient of 0.2331, suggesting that as GDP increases, programming literacy also increases. POP (Population) has a negative coefficient of -10.8342,

indicating that as the population increases, the dependent variable decreases. The statistical significance (P>|z|) values associated with each coefficient in the models help assess whether the estimated coefficients are significantly different from zero. A lower P-value suggests higher statistical significance.

- For GDP (MIO_EUR), the P>|z| value is 0.878, which is relatively high. This suggests that the coefficient for GDP (MIO_EUR) is not statistically significant at conventional significance levels (e.g., 0.05), indicating that the effect of GDP on the dependent variable may not be reliably different from zero.
- For POP (Population), the P>|z| value is 0.269, also relatively high. This suggests that the coefficient for Population is not statistically significant as well, indicating that the effect of population on the dependent variable may not be reliably different from zero.

Table 5 Independent Variables Statistics based on the OLS Regression (with GDP per capita)

	coef	**std err**	**P>|z|**
const	1209.1301	1427.264	0.397
EMP (THS)	14.7173	26.085	0.573
GDP (MIO_EUR)	-0.0130	1.510	0.993
GVA (MIO_EUR)	0.4254	1.618	0.793
GDP_per_cap	-0.0523	0.039	0.185
POP (THS)	-11.5291	9.839	0.241

The second model consists of an additional variable GDP per Capita (GDP_per_cap). GDP_per_cap has a negative coefficient of -0.0523, indicating that as GDP per capita increases, programming literacy decreases. GDP (MIO_EUR) has a very small negative coefficient of -0.0130, suggesting a minimal effect on the dependent variable. POP (Population) has a negative coefficient of -11.5291, similar to the first model. Comparing p-values it can be concluded that:

- For GDP (MIO_EUR), the P>|z| value is 0.993, which is very high. This further confirms that the coefficient for GDP (MIO_EUR) is not statistically significant, supporting the notion that breaking GDP into components may provide more meaningful insights.
- For GDP_per_cap, the P>|z| value is 0.185. While this is lower than the P-value for GDP in the second model, it is still relatively high, suggesting caution in interpreting the statistical significance of the GDP_per_cap coefficient.
- For POP (Population), the P>|z| value is 0.241, which is again relatively high, however is smaller than in the first performed model.

Comparing the two models, it seems that GDP (MIO_EUR) in the first model has a positive impact, while in the second model, GDP per capita has a negative impact. However, the coefficient for GDP (MIO_EUR) in the second model is very small, suggesting that breaking GDP into GDP per capita and population might be more informative.

The baseline OLS model that includes GDP per capita performs well also across other statistics. The Gross Domestic Product was not included in final model, because as stated previously it is not statistically significant for prediction of programming literacy. The results from the second attempt prove the initial assumptions (refer to Table 6 and Table 7).

Table 6 Independent Variables Statistics based on the OLS Regression (with GDP per capita)

	coef	**std err**	**P>|z|**
const	1201.6333	1514.344	0.427
EMP (THS)	14.6540	24.597	0.551
GVA (MIO_EUR)	0.4113	0.156	0.008
GDP_per_cap	-0.0521	0.037	0.154
POP (THS)	-11.5108	10.069	0.253

The second attempt to define the OLS model reveals that approximately 86.8% of the variability in programming activity is explained by economic factors, as indicated by the adjusted R-squared value. The model results in an F-statistic of 29.68, and the associated probability (Prob (F-statistic)) is very low (8.52e-21).

Table 7 Results of the OLS (Ordinary Least Squares) Regression Model (with GDP per capita)

Statistics:	**Value:**
R-squared:	0.868
Adj. R-squared:	0.866
F-statistic:	29.68
Prob (F-statistic):	8.52e-21

This low p-value suggests that the overall model is statistically significant, and the regression coefficients are not all equal to zero, although a high condition number suggests potential multicollinearity issues. This condition complicates the interpretation of individual predictor effects and can destabilize the regression coefficients, although the final set of independent variables can be defined. For the purpose of this study, statistically significant variables are Employment, Population, Gross Value Added and Gross Domestic Product per Capita.

5.2 Model Selection

To address potential multicollinearity and overfitting issues identified in the OLS model, Ridge and Lasso regressions are employed using independent variables including GDP per Capita. Recognizing the need for a robust evaluation metric, cross-validation scores are chosen to assess the performance of each model. Cross-validation involves partitioning the data into subsets, training the model on one subset, and evaluating it on another, providing a more reliable estimate of generalization performance. Several models are tested to compare performance across cross validation scores, as presented in Table 8.

Ridge and Lasso regressions are models that efficiently address multicollinearity issues, providing a balanced approach to controlling overfitting and identifying key predictors. Both models perform well, with Lasso slightly ahead in cross-validation scores due to its feature selection capability. Ridge does not improve the score achieved for Linear Regression. The Random Forest Regressor is investigated to capture any non-linear relationships. Although it shows a high cross validation score of 0.714 on the training set, the discrepancy between the training dataset cross-validation scores of 0.943 indicates potential overfitting. Based on the comparison of cross validation scores, Lasso regression is chosen as a final predicting model.

Table 8 Comparison of performance of multiple predictive models

Predictive Model	**Cross Validation Score**
Linear Regression	0.7235
Lasso Regression	0.7238
Ridge Regression	0.7235
Random Forest Model	0.7140

As a next step, the growth rate from the last 5 years is calculated and added as an additional variable, in order to calibrate the efficiency of the model. The reason behind this is to learn to model, predict and adjust to long term trends. The best performing model, Lasso Regression, was used to predict programming literacy from 2017-2020 based on the. Results from the Lasso Regression model are presented below in Table 9.

Table 9 Summary of the Lasso Regression Model with growth rate

Statistics:	**Value:**
Mean Cross-Validation Score:	0.650
Standard Deviation of Cross-Validation Score:	0.225
F-statistic:	272,02

5.3 Interpretation of the Results

In summary, the integration of economic indicators with various modeling techniques establishes a robust framework for predicting programming literacy, a crucial factor in guiding strategic urban technological development. Among these techniques, lasso regression emerges as the best-performing method, offering insights into the predictive power of economic factors on programming proficiency. The correlation between the predicted and observed values based on the cross-validation score of 0.65 across all the years validates the model's effectiveness and indicates that economic growth factors can be used for forecasting programming literacy in diverse geographical regions. Overall, the model shows high performance, indicated by consistently high F-statistics scores across all tests, reflecting a statistical strength of predictions.

However, the cross-validation score has not improved much in comparison to previous tested models. To investigate this issue, a brief examination of outliers is warranted. It is important to note that a more thorough analysis and discussion of outliers will be presented in the next chapter. While the economic factors explored demonstrate an ability to explain and forecast programming literacy, critical observation emerges. There is a tendency to overestimate programming literacy in regions with lower GDP per Capita/GVA and, conversely, to underestimate in areas where GDP per Capita/GVA is exceptionally high compared to other locations. This suggests a sensitivity to extreme economic conditions, leading to inaccuracies in predictions. Due to this limitation, regions such as the Balearic Islands, where the model overestimates programming literacy despite lower GDP per Capita, and locations like Paris, where the model underestimates activity due to exceptionally high GDP per Capita, highlight the limitations of relying solely on economic factors (refer to the Table 10). This prompts the realization that programming literacy is influenced by a broader spectrum of variables beyond economic factors, potentially encompassing social dynamics and local business characteristics.

Even though a model can predict programming literacy based solely on economic factors, to enhance the model's stability and reliability, a comprehensive approach should involve integrating indicators from education, social conditions, and industry-specific data, ensuring a more comprehensive and reliable source of information for predicting programming literacy.

Table 10 Analysis of chosen regions based on predicting model of programming literacy

NUTS-3 Name	GDP	GVA	Employment	Activity*	Residual
Balearic Islands	3'106.9	2'821.2	52.9	6'648.4	3'495.3
Paris	234'421.0	208'514.2	2'037.5	114'794.0	-27'229.0

* Variable "activity" is a target variable interpreted as "programming literacy"

5.4 Outliers

5.4.1 Outlier Identification

In predictive modeling, outliers can significantly influence the results and interpretations. In this research, the presence of outliers, especially in large towns with distinctive economic profiles, necessitates a focused analysis. These outliers are not just statistical anomalies; they often represent unique socio-economic environments that can provide valuable insights into the dynamics of programming literacy.

Outliers are detected and identified based on the residual analysis. With residual analysis, it is possible to analyze the differences between the predicted programming literacy and the actual programming literacy recorded in the NUTS-3 region. The results are easy and intuitive to analyze, the higher the difference, either negative or positive, the more outlined the prediction is (Agresti 2002). Our identification of outliers is further supported by the visual analysis of the residuals from our predictive model, as depicted in Appendix 1. The chart displays a distribution with a few extreme values that are quite distant from the rest of the points. Such visual evidence underscores the need to address these outliers comprehensively, as they could

represent regions with anomalous economic growth or programming density that do not align with the general trends.

Based on the residuals analysis, a few major outliers can be defined. To identify what constitutes an outlier, the standard deviation is calculated. Based on that value the regions that record residuals higher than the standard deviation are considered outliers.

Table 11 presents the most deviated outliers from the prediction. The first three rows correspond to the most overestimated regions in terms of programming literacy, while the bottom three rows represent the most underestimated regions in terms of programming literacy. The analysis identifies outliers as particularly large towns with high GDP per Capita and employment rates. These include major cities like Paris, Madrid, Stockholm, and Zurich, which are primarily capital cities and major economic hubs. These cities stand out due to their unique patterns in programming literacy, which deviate from the trends observed in other regions.

Table 11 Outliers based on the Residuals analysis made on the base of OLS Regression

Region Name	**Country**	**EMP (THS)**	**GDP per Capita**	**Activity***	**Prediction**
Rome	Italy	2171	38176	24875	More
Hauts-de-Seine	France	1148	108940	24309	More
Naples	Italy	993	19842	6992	More
Paris	France	2037	107931	114794	Fewer
Zürich	Switzerland	864	90147	55345	Fewer
Stockholm County	Sweden	1319	64446	58633	Fewer

* Variable “activity” is a target variable interpreted as “programming literacy”

The regions where the model overestimates programming activity share some commonalities. Notably, these regions, such as Hauts-de-Seine in France and Barcelona in Spain, often have relatively high values in both Employment (EMP) and Gross Domestic Product per Capita (GDP_per_cap). This suggests that the model might be influenced by regions with higher

economic and employment activities, potentially leading to overestimation of programming activity. On the other hand, regions where the model underestimates activity, like Paris and Zürich, exhibit significant programming activity relative to their economic indicators. This could indicate that the model may not be capturing the nuanced relationship between economic indicators and programming literacy in these specific regions, resulting in underestimation. The differences between overestimated and underestimated regions highlight the importance of considering regional specifics and potentially refining the model to better capture the factors influencing programming activity in diverse geographical contexts. Further investigation into the specific features contributing to these discrepancies and potential interactions between variables could provide insights for model improvement.

5.4.2 Outliers Analysis

The findings presented in Figure 13 and 14, that neither GDP per Capita nor Employment is correlated with the level of residuals in the model, suggests that these significant variables may not be directly influencing the discrepancies between predicted and actual values. In other words, variations in GDP and Employment do not seem to explain the differences in the model's predictions and the observed programming activity levels.

Figure 13 GDP per Capita vs Residuals per NUTS-3 Regions based on the Outliers

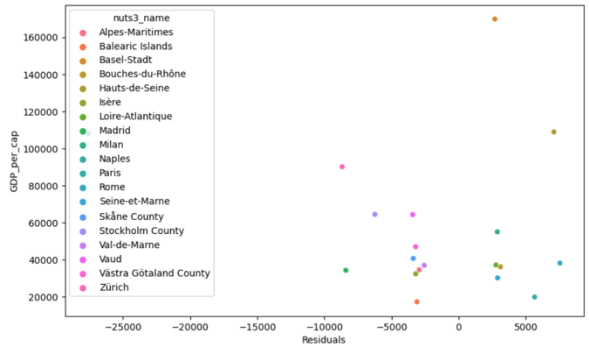
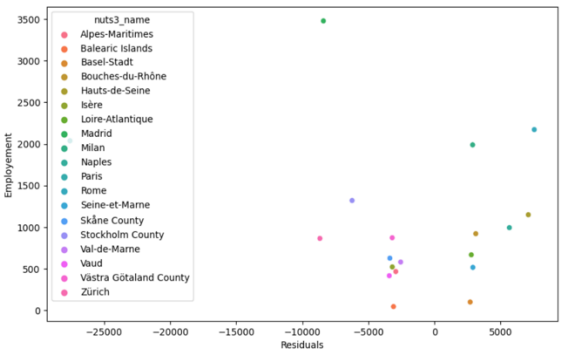


Figure 14 Employment vs Residuals per NUTS-3 Regions based on the Outliers



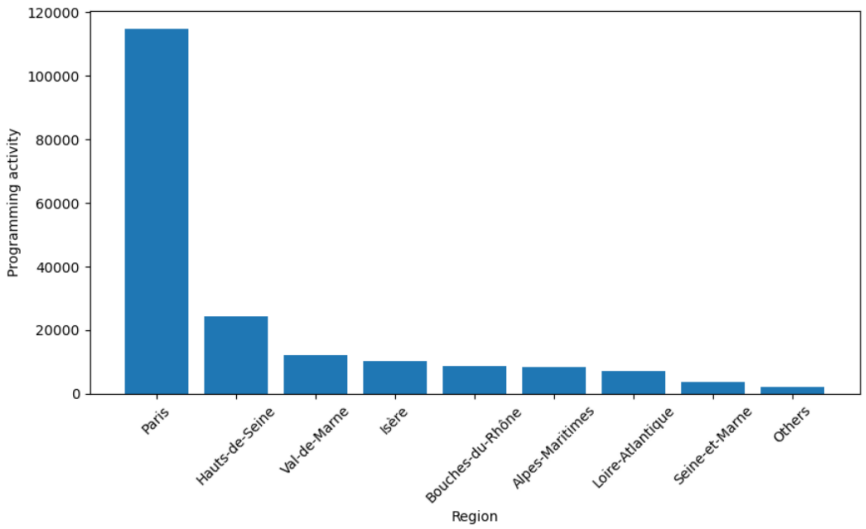
Since GDP per Capita and Employment do not seem to explain the level of residuals, relying solely on these variables for outlier analysis may provide limited insight. Outliers might be

driven by factors not captured by the chosen predictors. Further exploratory analysis is warranted to identify the specific characteristics or regional factors contributing to outliers. This could involve investigating the nature of the residuals, examining potential interactions between variables, and considering region-specific features.

Further investigation shows that France has twice as many regions identified as outliers compared to other countries, with 8 outliers, while the rest of the countries analyzed have an average of 3 outliers. In most cases, top cities such as Barcelona and Madrid in Spain, and Milan, Rome, and Florence in Italy, are identified as outliers. Surprisingly, Switzerland has only 3 outliers, even though the GDP per Capita of the regions in that country is significantly higher than in other countries with a greater number of outliers.

To understand the dominance of France in terms of outliers, a more in-depth analysis is needed. Figure 15 below, shows that Paris dominates in programming activity, exhibiting significantly higher values than other regions, thus indicating a strong presence in programming literacy.

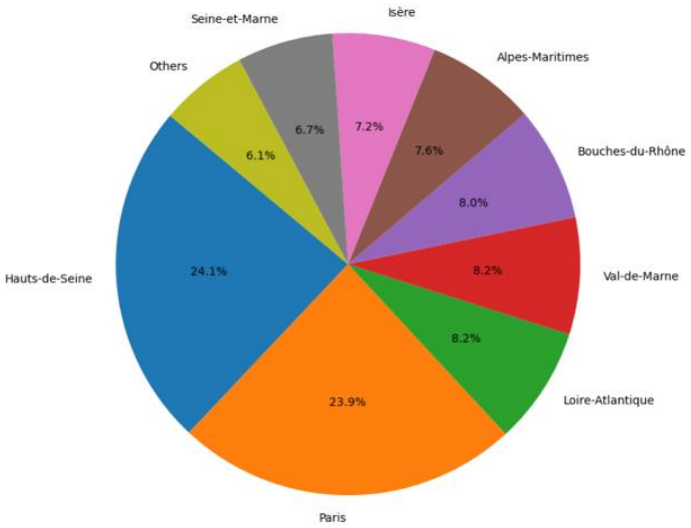
Figure 15 Distribution of Average Programming Literacy across the outlier regions and rest of the French NUTS-3 Regions.



There are substantial differences in programming literacy activity between Paris and other regions, suggesting regional disparities in programming literacy initiatives. The 'Others' category represents the combined programming literacy (defined as “activity”) of multiple

regions, contributing less individually but significantly in aggregate. It shows that all regions that are identified as outliers have higher programming literacy than the remaining NUTS-3 regions identified properly. This proves that the model is not adjusting well to higher programming literacy levels.

Figure 16 Distribution of Average GDP per Capita across the outlier regions and rest of the French NUTS-3 Regions.



As a next step, the GDP per capita distribution is analyzed as shown in Figure 16. Paris and Hauts-de-Seine stand out with high GDP per capita compared to other regions, as well as Loire-Atlantique and Val-de-Marne. The 'Others' category, consisting of the remaining NUTS-3 regions, presents lower mean GDP per capita than the outlier regions. This suggests that the discrepancy in mean GDP per capita in outlier's regions might influence the model's suboptimal predictions, a hypothesis that could be validated by scrutinizing individual locations one by one. France's top region shows the same behavior. Renowned for being one of the wealthiest regions in France, Hauts-de-Seine is a major hub for European commerce. One of Europe's principal commercial hubs, Paris, is primarily responsible for the region's economic prosperity. Examining French administrative departments reveals diverse characteristics. Economic significance in departments like 'Hauts-de-Seine,' near Paris, may be due to major business districts, while 'Alpes-Maritimes' contribute primarily through tourism and services. Variations

in programming literacy and GDP per capita across French regions result from factors such as geographical remoteness, economic focus, and proximity to urban centers. France could be considered an outlier due to its higher-than-average GDP per capita, employment, and economic activity levels compared to the other countries in the dataset. The large standard deviations in these measures also contribute to the perception of France as an outlier, suggesting that there is considerable regional variation within the country, which model could not capture well, without additional dimensions.

In summary, outliers with higher actual programming literacy are predicted with lower, and vice versa, which seems to be an interesting finding. This is probably caused by the lack of additional dimensions; hence the model cannot grasp the more complicated relations between programming literacy and economic growth. Therefore, it is generalizing all regions and averaging the programming literacy across the whole dataset.

5.4.3 ARIMA and non-stationarity in the outliers' regions

Building upon the initial analysis of outliers in predictive modeling, this part progresses towards a nuanced exploration of these regions using ARIMA (AutoRegressive Integrated Moving Average) modeling. This approach focuses particularly on the non-stationary aspects of the identified outlier regions. Non-stationarity, characterized by statistical properties of a time series that change over time, poses a significant challenge to model predictability and stability. It is particularly relevant when exploring socio-economic environments like those represented by outliers, where standard analysis methods may not capture all influencing factors.

In this context, Rome and the Balearic Islands emerge as regions of interest due to their non-stationary behavior. This suggests that these areas may experience economic and social changes not typically captured by more traditional models. By employing ARIMA modeling, a deeper understanding of the time-based dynamics within these regions is sought, aiming to identify

underlying patterns and trends that affect programming literacy. The choice of a p-value threshold of less than 0.05 in the ARIMA model is critical in this regard. It ensures that only the most statistically significant patterns in the time series data are considered for the analysis. This filtering criterion aids in isolating regions where there are meaningful, non-random fluctuations in data over time.

The identification of non-stationarity in regions like Rome and the Balearic Islands through ARIMA modeling offers valuable insights. It underscores the presence of dynamic economic factors that potentially influence programming literacy in ways that deviate from general trends observed in other regions. Figures 17 and 18 depicting the GDP per capita over time for these regions further illustrate this point. Fluctuations in GDP per capita in both Rome and the Balearic Islands indicate economic variability, which might be influencing the programming literacy environment in these regions.

Figure 17 GDP per Capita Over Time for Balearic Islands

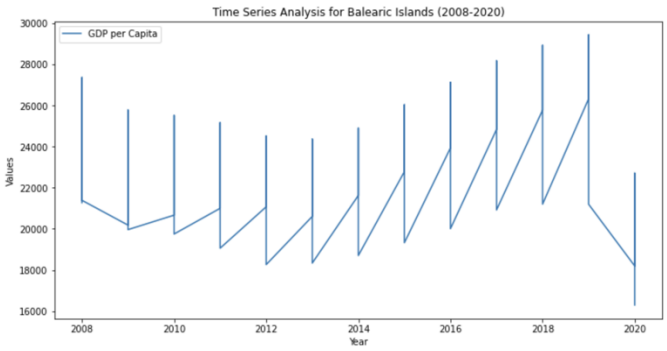
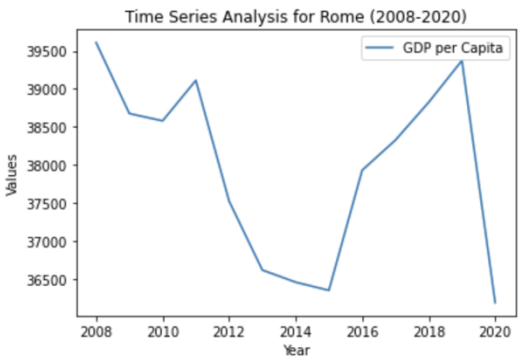


Figure 18 GDP per Capita Over Time for Rome



The ARIMA analysis of non-stationary regions reinforces the complexity of economic patterns and their impact on programming literacy. It suggests that the current model could be augmented to account for fluctuations over time, particularly in dynamic regions like Rome and the Balearic Islands. The findings from this chapter should inform not only the interpretation of the current model's results but also the development of future models.

6 Discussion

In exploring the relationship between programming literacy and economic growth in European regions, the study indeed found a correlation as well as patterns between both dimensions. The study identified major European tech hubs like Paris, Stockholm, and Zurich as central to programming activity, with their status as capital cities or economically dominant regions providing the necessary infrastructure, educational resources, and corporate density essential for thriving programming communities, as theorized in the thesis. Notably, Swiss regions occupied many places in the rankings. An explanation could be Switzerland's strong emphasis on high-quality education, substantial investment in research and development, as well as high living standards attracting global talents, which have collectively fostered a conducive ecosystem for programming excellence and innovation. Unexpectedly, the Balearic Islands and La Palma also demonstrated high programming activity, likely due to their emerging status as remote work havens, while Italian regions, despite initially lagging, have shown significant growth, signaling a broader, policy-influenced tech evolution across Europe.

In addition, the study employed various predictive models to forecast programming literacy using economic indicators. While models like Lasso regression showed potential, the in depth-analysis of outliers revealed limitations, particularly in regions with unique economic profiles or high GDP per Capita. These findings suggest a complex interplay between economic growth and programming literacy, influenced by diverse regional characteristics and evolving trends such as remote work.

As the group study concludes, the forthcoming chapters will delve into a range of focused topics: After conducting panel data analysis, Poland serves as a case study to explore various facets of interest. Subsequently, new variables capturing programming literacy are added and their impact assessed and lastly programming literacy’s effect on economic resilience is investigated.

7 Exploration of New Variables and their Impact

7.1 Introduction

In today's landscape of economic development, programming literacy is not just an insignificant element but an essential component that fosters economic growth. The worlds trend to a digital transformation of the global economy highlights the importance of programming literacy, which translates to a growing demand for programming skills across different sectors. Knowledge of and competence in computer languages is being recognized as a key component of economic development, propelling nations, and industries toward higher levels of productivity and competitiveness.

The purpose of this study is to further develop our analysis to the NUTS-2 level. The approach reduces the amount of detail in terms of region comparison, but it is carried out to find new variables that could improve the performance of our prediction models. Whilst the previous analysis explores the relationship between programming activity and economic indicators such as GDP, GVA, and employment rates, this individual study follows a different course using a different set of variables. It is built on the idea that a less region-specific approach—as demonstrated at the NUTS-2 level—might expose more factors influencing programming activity hidden at higher regional analysis levels.

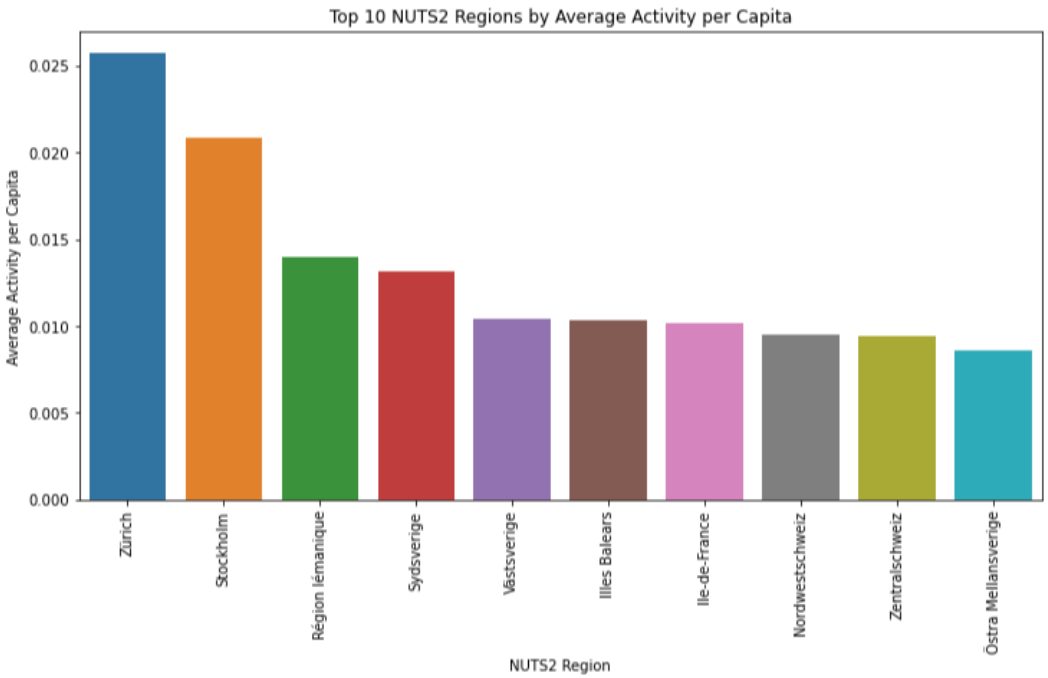
The intention is not to replicate the procedures and findings but rather to strengthen and expand upon them by examining if adding more or different variables would result in a more robust model. By doing this, it hopes to contribute to a more comprehensive understanding of the elements that support or obstruct programming literacy development and its incorporation into strategies for economic success.

7.2 NUTS-3 Analysis at NUTS-2 Level

In this section, we expand our investigation into geographical analysis at the NUTS-2 level in order to acquire a deeper understanding of both the distribution and intensity of programming endeavors across more extensive regions. This allows for insights into the wider economic and demographic contexts that shape engagement in programming activities. Unlike the previous exploration conducted at the NUTS-3 level, which provided a detailed perspective on urban hubs and smaller locales, investigating at the NUTS-2 level encompasses larger territories that often amalgamate various urban and rural areas. Adopting such an approach proves indispensable when endeavoring to comprehend macro-level patterns while simultaneously recognizing disparities in programming activity among different regional settings.

As shown in Figure 29, A notable presence of Swiss regions, such as Zurich, Région lémanique, Nordwestschweiz, and Zentralschweiz can be observed. These regions are characterized by their strong economies and high living standards, which is reflected in their significant engagement in programming activities. The data presented on the Figure reaffirms the idea that economic prosperity and a superior quality of life often go hand-in-hand with increased involvement in technology-related pursuits. Similarly, Swedish regions like Stockholm, Sydsverige,Västsverige,and Östra Mellansverige are prominently represented. This depiction further supports the narrative surrounding Sweden's flourishing technological sector and its dedication to digital innovation. It highlights active participation within these areas when it comes to programming endeavors.

Figure 19 Top 10 NUTS-2 Regions based on the Stack Overflow's Activity Size per Capita



It is worth noting that certain regions like Illes Balears and Ile-de-France maintain their prominence at both NUTS-3 and NUTS-2 levels. Amidst this consistency lies evidence for the enduring influence exerted by these specific locations on the programming landscape.

7.3 Predicting Programming Literacy with the Same Indicators

For this analysis the approach closely mirrors that of our examination on the NUTS-3 level. By employing the Ordinary Least Squares (OLS) model, we endeavor to unravel the intricate connection between programming activity and key economic indicators such as GDP, employment rates, and Gross Value Added (GVA) within these larger geographic areas. This particular methodology enables us to make meaningful comparisons and gain insights into disparities in programming density and economic associations between both NUTS-2 and NUTS-3 levels for the reference year 2016.

Table 12 Comparison of OLS Results in NUTS-2 and NUTS-3 Analysis

Aspect	NUTS-2 Analysis	NUTS-3 Analysis
R-squared	0.940 (High explanatory power)	0.868 (Moderate explanatory power)
F-statistic	41.85 (Statistically significant model)	29.68 (Statistically significant model)
Prob (F-statistic)	7.59e-16 (Highly significant model)	8.52e-21 (Highly significant model)
EMP (THS)	Negative correlation (Significant)	Negative correlation (Insignificant)
GDP (MIO_EUR)	Positive correlation (Insignificant)	Positive correlation (Insignificant)
GVA (MIO_EUR)	Negative correlation (Insignificant)	Negative correlation (Insignificant)
Multicollinearity	High condition number (Potential issue)	High condition number (Potential issue)

The coefficient of determination, R-squared, with a value of 0.940 implies that our model explains approximately 94% of the variability in programming activity. Furthermore, the F-statistic is significant at 41.85 and has a low p-value (7.59e-16), confirming the statistical significance of our model.

Comparing the NUTS-2 and NUTS-3 models, we observe that the former exhibits a higher R-squared value (0.940) than the latter (0.868). This indicates that at the NUTS-2 level, our model possesses greater explanatory power regarding programming activity. Similar to what was found in the NUTS-3 analysis, neither GDP nor GVA emerge as significant predictors at the NUTS-2 level either. This consistent lack of significance across both levels suggests that these economic indicators might not have direct influence on programming activity. Interestingly enough, there exists a negative correlation between programming activity and economic factors at both NUTS-2 and NUTS-3 levels; however, this relationship appears more pronounced within regions defined by correspondingly larger geographical areas. Notably high condition numbers in both models indicate potential issues related to multicollinearity which could be influencing or distorting our results.

By delving into four distinct models, one can gain an all-encompassing comprehension of the correlation between programming activity and economic indicators on various regional levels.

Table 13 succinctly outlines these discoveries using cross validation as metric to evaluate their performance.

Table 13 Comparison of Cross-Validation scores in different models in NUTS-2 and NUTS-3 Analysis

Model	**NUTS-2 - Mean CV Score**	**NUTS-3 - Mean CV Score**
Linear Regression	0.3918	0.6892
Lasso Regression	0.7148	0.7089
Ridge Regression	0.7259	0.6892
Random Forest Model	0.6787	0.7154

The performance of Linear Regression differs significantly between NUTS-2 (mean CV score: 0.3918) and NUTS-3 (score: 0.6892), suggesting that the relationship is more complex at the broader regional level, indicating a need for a more sophisticated model beyond linear regression to capture all nuances accurately. Lasso Regression consistently performs well at both NUTS-2 (score: 0.7148) and NUTS-3 (score: 0.7089) levels, demonstrating its effectiveness in selecting relevant features regardless of regional scale. This consistency underscores certain economic indicators' importance, such as GDP, in influencing programming activity across different regions. Conversely, Ridge Regression shows better performance at the NUTS-2 level (score: 0.7259) compared to NUTS-3 (score: 0.6892). This improvement suggests that Ridge Regression's ability to handle multicollinearity is better suited for capturing the intricate relationships among economic factors in larger geographic areas like those grouped under NUTS-2.

7.4 Data Collection for the New Variables

One single NUTS-2 region can be present in several NUTS-3 regions. This implication of loss of granularity in the region comparison in the model can be explained by the availability of variables at each level, meaning that as the region gets too specific, so does the data. This constraint is important to our study because it limits our capacity to investigate specific

technological and economic factors that may have an impact on programming literacy. Research at the NUTS-2 level proposes us the opportunity to use a larger pool of datasets that can broaden or study. As with the NUTS-3 level analysis, the Eurostat database will be the source of data since it has proven to be a reliable source for obtaining information. Following a thorough examination of the available data, we have chosen to focus on a few key areas:

- *Households with access to the internet* at home are categorized based on individuals who have used the internet at least once in the three months prior to the survey, focusing on private households with at least one member aged between 16 and 74 years to ensure capturing a digitally active population (Eurostat [ISOC_I_ESMS] 2023). The data is presented in terms of percentages of households.

The relevance of this variable to our model lies on the significance of the level of internet connectivity in the digital connection of a given region. We aim to study the existence of the correlation between higher rates of internet access and increased levels of programming literacy based on the assumption that having access to internet is a prerequisite for any individual to develop and refine their digital skills.

- The dataset on *Gross Domestic Expenditure on R&D by Sector of Performance* provides an in-depth analysis of Research and Development (R&D) activities, emphasizing expenses within different sectors. This dataset presents a breakdown of performance, including sectoral distribution, sources of funding, types of costs, research fields, and socio-economic goals (Eurostat [RD_ESMS] 2023).

For the sector of analysis, we will focus on “Total” which is an aggregate of all the sectors which include Business enterprise, Higher education, government, private non-profit sector. For the units of measurement, we will focus on EUR_HAB (Euros per Inhabitant) which

reflects R&D spending per person in a region. Its relevant to our study because GERD (Gross Domestic Expenditure on Research and Development) is a crucial measure of a region's dedication to fostering. This indicator not only signifies the importance placed on research in both public and private sectors, but also plays a pivotal role in cultivating an economy driven by knowledge. To gain valuable insights into the distribution of R&D spending across different regions, analyzing GERD at NUTS-2 regional level proves indispensable. Such analysis aids in comprehending disparities in innovation capacities among regions and provides vital information for devising effective strategies for regional development. Another interesting aspect to consider is the potential correlation between R&D investment levels and programming literacy. Regions that allocate higher expenditures towards research are more likely to possess advanced technological infrastructure, creating a greater demand for skilled professionals proficient in programming languages.

- The dataset on *High-tech Industry and Knowledge-intensive Services - Employment in Technology and Knowledge-intensive Sectors* provides a perspective on employment across industries, categorized by their level of technological intensity. It encompasses various facets such as economics, employment, science, technology, and innovation (STI), offering a multifaceted view of the workforce in these sectors (Eurostat [HTEC_ESMS] 2023).

NACE_r2 codes are used to categorize activities into various sectors such as high-technology manufacturing ("C_HTC") and knowledge-intensive services ("KIS") which allows for a more refined analysis of different industry segments. It also employs distinct units of measurement but for this analysis THS_PER" or Thousands of Persons will be used to gauge workforce size effectively. This variable can be potentially valuable to the analysis because high-tech and knowledge-driven sectors typically exhibit a greater demand for competencies in programming and digital literacy. By examining employment trends within these sectors, we can assess the

robustness and expansion potential of industries that are most likely to leverage programming proficiency to their advantage.

- The '*Human Resources in Science and Technology (HRST)*' dataset shows the labor market dynamics within the science and technology sectors. It offers perspective on 'stocks'—the number of individuals working in these fields at a given time—and 'flows,' which represent the changes occurring over time. This dataset provides a global snapshot that mirrors the workforce composition in science and technology industries, capturing both current states and evolving trends (Eurostat [HRST_ESMS] 2023).

For this analysis, we are placing our research emphasis on the "stocks" component within the HRST dataset with the aim of comprehending the present workforce composition in science and technology sectors. Our examination will entail refining the dataset to examine overall figures for HRST (Human Resources in Science and Technology) as well as individuals possessing tertiary education in science and technology occupations (HRSTC and HRSTE), along with those employed in such roles without formal qualifications (HRSTO). To quantify the labor force within these sectors, we will use 'THS_PER' which stands for Thousands of Persons. By analyzing the size and composition of the workforce in science and technology sectors, we can gauge the scale of employment in industries most likely to utilize and benefit from programming literacy. Which can also provide insights into regional capabilities in innovation and technology-driven economic activities.

- *Structural business statistics SBS data* Eurostat [SBS_ESMS] 2023) examines the structure, behavior, and performance of diverse economic activities. Its main objective is to present a comprehensive overview of the European economy with specific emphasis on three key aspects: Business Demographics, Ouput and Input related variables.

For this research, SBS data is important to understand the economic activities and sectors that most significantly interact with or are influenced by programming literacy. We are particularly interested in sectors like "J62: Computer programming, consultancy and related activities" and "M: Professional, scientific, and technical activities" as they are directly relevant to the technological and programming aspects of our study. Our analysis will concentrate on key variables including the count of local business establishments, compensation and wage levels, employment figures, the rate of employment growth, and the proportion of the workforce engaged in manufacturing. These variables give us insights into the scale and economic health of the sectors most aligned with programming activity. By thoroughly examining these sectors, we can gain a deeper understanding of how they contribute to the economy and their relationship with programming literacy. Analyzing the correlations between economic output data such as turnover and levels of programming literacy at a regional level provides valuable insights into the impact that technical skills have on economies' overall productivity.

7.5 Data Preparation

In analyzing Stack Overflow data, we encountered the significant challenge of aligning the dataset—organized at the NUTS-3 level—with our analytical needs, which demanded a NUTS-2 level perspective. To bridge this gap, we established a mapping between the two levels using Eurostat's regional classification to ensure precision and relevance.

Once the mapping process was successfully implemented, we proceeded to aggregate the original NUTS-3 data to fit within the desired framework of the NUTS-2 level. This aggregation procedure entailed either summing or averaging relevant metrics from the initial dataset based on specific criteria dictated by both data nature and analytical requirements associated with operating at the higher geographical scale represented by NUTS-2.

7.6 Data Cleaning and Preprocessing

The initial step in the data analysis process involved determining the extent of missing data present across all variables. This was accomplished by computing the percentage of missing values for each column. A thorough examination revealed that certain variables, namely 'J62_V11210', 'J62_V16110', 'M_V11210', and 'M_V16110', exhibited a high proportion of missing values exceeding 30%. Based on our established threshold for handling missing data (40%), any columns surpassing this limit were considered unreliable for further analysis and therefore excluded from subsequent calculations. The justification behind this decision stemmed from recognizing that an elevated percentage of absent information within a specific column has the potential to skew outcomes and interpretations derived from comprehensive analyses.

To address gaps in selected variables such as 'internet_access (PC_HH)' and HRST-related factors, we applied linear interpolation techniques. Linear interpolation emerged as an optimal choice due to its effectiveness when dealing with time-series datasets characterized by relatively uniform changes between consecutive time points. In instances where there were blanks within variables like "TOTAL" or "SE," group-based imputation methods were deemed appropriate. This approach entailed deriving mean values for individual groups delineated according to NUTS-2 region and year, subsequently filling in absent values using these calculated means. By adopting this strategy, we ensured that imputed figures maintained contextual relevance while preserving regional disparities and temporal trends inherent within the dataset.

7.7 Analysis with the New Variables

The correlation matrix provides valuable insights into the relationships between various HRST indicators and the programming activity. One noteworthy observation is that there exists a weak to moderate positive correlation (0.27) between 'internet_access (PC_HH)' and 'activity,'

indicating that regions with better internet access tend to exhibit slightly higher programming activity. Furthermore, it is crucial to acknowledge the strong correlations among different HRST-related variables, underscoring their interconnectedness. Moreover, these HRST variables demonstrate moderate to strong positive correlations. Among them, 'HRSTC' stands out as having the highest correlation (0.80). This suggests that regions boasting educated professionals in science and technology witness more significant levels of programming activity. Interestingly, another influential factor is highlighted by the strong correlation between structural business statistics ('SE') and 'activity' (0.81), implying that areas with thriving business statistics tend to foster a greater amount of programming activity. However, when building predictive models or conducting further analysis on this data set, caution must be exercised due to multicollinearity concerns arising from intercorrelations among HRST variables.

7.7.1 OLS Regression

During the OLS regression analysis conducted, it was apparent that this new model possesses a commendable explanatory power with an R-squared value of 0.871, this suggests that the model is able to effectively account for a substantial portion of the variability observed in programming activity. Furthermore, the F-statistic yielded statistical significance, further validating the overall importance and relevance of the model. Nevertheless, upon closer examination of individual variables such as 'SE,' 'TOTAL,' and 'HRSTC,' their level of significance appears to be comparatively weak or inconclusive within this context.

7.7.2 Ridge and Lasso Regression

In our examination of the Ridge and Lasso regression models, we observed that their respective performances were quite comparable. The R-squared values associated with both models yielded moderate results, specifically 0.6120 for Ridge regression and 0.5598 for Lasso

regression. This outcome indicates that when it comes to model performance in relation to this particular dataset, the decision between utilizing either the Ridge or Lasso technique does not appear to have a substantial impact since the penalty types (L2 for Ridge and L1 for Lasso) do not seem to significantly influence overall model efficacy.

7.7.3 Gradient Boosting Model

The Gradient Boosting model exhibits an exceptionally accurate fit to the training data, achieving a remarkable R-squared value of 0.9996. Nevertheless, this high level of accuracy gives rise to concerns about potential overfitting. In contrast, when tested on the test set, its performance is comparable to that of Random Forest models with an R-squared value of 0.7735. However, it is worth noting that the Random Forest model demonstrates greater stability and consistency in its performance across various subsets of data. This assertion can be supported by observing higher mean cross-validation scores as well as lower standard deviation compared to those exhibited by the Gradient Boosting model's results. Consequently, these findings suggest that for this particular dataset, the Random Forest model proves itself more dependable and trustworthy than its counterpart

7.7.4 Random Forest Model

After implementing advanced feature engineering strategies, we observed a remarkable rise in the R-squared value of the test set, reaching 0.8026. Additionally, there was a notable enhancement in the mean cross-validation score, which reached an impressive level of 0.7234. These advancements clearly demonstrate that our most recent Random Forest model surpasses its predecessors by employing non-linear transformations on features to effectively account for variations in 'activity' when dealing with new data instances. Furthermore, this latest model exhibits superior stability and consistency across diverse subsets of data.

7.7.5 Estimated Regression Equations

In the analysis conducted, it was observed that the OLS regression equation with the original indicators demonstrated a noteworthy negative correlation with employment. This suggests that regions characterized by higher levels of employment may exhibit slightly lower programming activity. Consequently, this finding implies that economies which are more diversified and not solely reliant on technology-based sectors might have decreased programming activity.

Table 14 Comparison Table of Estimated Regression Equations

Model Type	NUTS-2 Analysis (Original Indicators)	NUTS-2 Analysis (New Indicators)
OLS	y = -1021.70 - 16.63*X1 + 0.55*X2 - 0.14*X3	y = -3223.50 + 302.16*SE - 7.95*TOTAL + 21.89*HRSTC
Ridge	y = -1115.91 - 99768.90*X1 + 208258.10*X2 + 73310.03*X3	y = 15458.57 + 19773.92*X1 - 6284.29*X2 + 8598.43*X3
Lasso	y = -1002.25 - 101636.06*X1 + 284167.66*X2 + 0.00*X3	y = 15458.57 + 23917.08*X1 - 7342.20*X2 + 5621.12*X3

Note. X1: EMP (THS), X2: GDP (MIO_EUR), X3: GVA (MIO_EUR), SE: Structural business statistics, TOTAL: Total variable, HRSTC: Human Resources in Science and Technology.

Interestingly, when incorporating new variables such as Structural Business Statistics (SE) and Human Resources in Science and Technology (HRSTC), a clearer understanding of direct relationships emerged through the regression equation. Specifically, a strong positive coefficient for SE indicated that regions boasting robust business environments tend to experience heightened levels of programming activity.

Furthermore, employing Ridge and Lasso models proved beneficial in addressing multicollinearity concerns present within these datasets. By implementing regularization techniques into these models, they were able to provide enhanced stability in interpreting the data. Notably, discrepancies between coefficients obtained from Ridge and Lasso highlighted diverse influences exerted by predictors under conditions where multicollinearity is present. Particularly significant is how Lasso effectively reduces certain coefficients to zero while emphasizing those predictors deemed most relevant for accurate prediction purposes.

7.8 Conclusion

The focal point of this research lies in the shift from analyzing programming literacy at the NUTS-3 level to examining it at the NUTS-2 level. The goal is to improve our predictive model by incorporating a wider range of economic and technological indicators. Although this transition results in less detailed data, it opens up new possibilities for understanding how programming literacy relates to economic development across Europe's regions.

The findings reveal that sacrificing granularity for access to more extensive data at the NUTS-2 level contributes significantly to a stronger predictive model. By including variables such as internet access, R&D expenditure, high-tech employment, and HRST indicators, we have enriched our analysis substantially. These additional factors not only provide a broader perspective on programming literacy but also uncover complex correlations that were harder to detect using NUTS-3-level analysis alone.

Through advanced statistical methodologies and exploration conducted at the NUTS-2 level, we have gained valuable insights into the subject matter. The use of techniques like Random Forest modeling combined with advanced feature engineering has proven superior in terms of accuracy and stability when predicting programming literacy levels. This improvement underscores the effectiveness of integrating a broader set of variables that capture different facets of programming proficiency holistically.

In the ever-evolving landscape of Europe's digital transformation, the findings extracted from this research possess profound consequences for policymakers, educators, andof industry leaders who aim to leverage coding proficiency as a driving force behind economic prosperity and scientific progress.

7.9 Discussion and Outlook

This research endeavor was aimed at addressing how programming literacy influences not only economic growth but also resilience. To explore this, (multinomial) logistic regression models were developed, assessing the impact of programming activity—derived from Stack Overflow data—on resilience, quantified through regional employment growth relative to EU-level performance. The primary insight from these regression models is the identification of a positive and statistically significant impact of programming activity and programming density in NUTS-3 regions. Notably, the coefficient values are higher during the post-financial recession recovery phase. This could be attributed to regions with higher digital literacy being more adept at leveraging innovation for growth following crises. Additionally, the influence of digitalization trends such as the smartphone revolution, cloud computing, and big data—which became more pronounced in the late 2000s and early 2010s—may have only manifested significantly during the recovery phase. Moreover, the multinomial regression analysis corroborated that across all four ERI group combinations, increased programming literacy rates bolster resilience in NUTS-3 regions. These insights underscore the need for policymakers to prioritize enhancing programming literacy, a key determinant of ER that bolsters innovation, entrepreneurship, and economic diversification, vital for prospering in the modern, high-value 'New Economy'.

For future research, additional examinations to certify the results from the models are crucial, especially as their mean cross-validation scores are only around 65-70%, giving room for improvement. To do so, developing a more comprehensive variable for programming literacy, encompassing a broader range of data beyond Stack Overflow, might provide a more nuanced

understanding of programming literacy's multifaceted nature. Similarly, enhancing the ERI with additional dimensions beyond employment data could yield a more intricate index. The author recommends attempting the methodology applied by Staníčková and Melecký (2018) Furthermore, the current dataset's timeframe poses certain limitations, as it does not fully encapsulate the financial crisis and lacks data post-2020. The latter hinders us to examine ER in the context of the pandemic, which would allow for a multifaceted analysis across diverse crises and compare them amongst the same regions. Hence, extending the dataset to cover these periods is recommended for a more thorough examination of the entire resilience cycle. However, it's important to note that data availability at the NUTS-3 level is constrained pre-2008 for certain countries, suggesting a potential shift to NUTS-2-level analysis for some variables might be beneficial. Alternatively, expanding the dataset to encompass a broader range of EU countries may enhance the robustness and representativeness of the findings.

8 Conclusions

This research builds on prior work to significantly advance our understanding of how programming literacy contributes to economic growth, a relatively underexplored area. It emphasizes the societal value of programming skills and informs policy strategies to promote such literacy as a key asset for regional development. To improve performance of regression models, predicting programming literacy rates based on economic growth metrics, incorporating comprehensive panel data into economic models significantly enriched the analytical framework. This enhancement has refined the prediction of programming literacy rates from economic growth indicators, crucial for accurate forecasting. Despite the success of the modeling approach, the individual paper acknowledged potential overfitting in machine learning models. Future research suggestions included investigating more advanced ML

algorithms, and linear regression models for panel data that improved accuracy in explaining programming literacy and forecasting economic growth, enhancing our understanding of the impact of programming literacy on economic development.

The next contributions displayed how using data at hand can be applied to conduct in-depth analysis on a single country. They investigated the complexity of the interplay in Poland, emphasizing the need for a tailored approach due to the country's unique development stage in comparison to other European countries. There is a discernible upward trend in programming literacy, particularly in less developed regions, signaling a positive trajectory, driven by economic growth, the availability of technical education, and trending programming languages. The transition from NUTS-3 to NUTS-2 level analysis has enhanced the model's ability to predict regional development by incorporating a broader spectrum of indicators. This enriched model has illuminated several factors that influence the rise of programming hubs, such as economic prosperity, high living standards, and the prevalence of programming activities. Key contributors include robust internet access and significant R&D spending. Despite a negative correlation between programming activity and employment, indicating that the technology sector isn't the only driver of economic health, the presence of strong business environments and high-tech industries is key to the development of programming hubs. GDP and GVA were not reliable predictors at the NUTS-2 level, highlighting the importance of integrating multiple variables, such as R&D and high-tech employment, for more accurate forecasts of programming literacy.

Lastly, the paper delves into the concept of economic resilience, another popular concept next to economic growth, tackling the question how well regions can handle and recover from shocks such as a recession or the pandemic. Just as with growth metrics, the paper coherently showed weak, yet positive correlations between resilience and programming literacy. Nevertheless, the

models' performances could be improved to deliver more accurate and hence reliable results, and for enriched representativeness, more indicators of programming literacy included.

9 Discussion and Outlook

There are a few limitations throughout our study that need to be considered. Firstly, our dataset is restricted to five countries in Europe, all characterized by higher GDPs compared to other European nations. This raises concerns about the generalizability of our findings, particularly in regions with relatively lower GDPs. Another limitation lies in the representativeness of our chosen metric for programming literacy. The reliance on Stack Overflow data, capturing information about comments, answers, and questions, may not offer a fair representation of programming literacy. It is expected that a significant segment of individuals proficient in coding may not actively participate on such platforms. As a result, additional data sources need to be incorporated. Integrating information about projects uploaded on GitHub, a platform facilitating collaboration among programmers, would provide a more comprehensive perspective and generalize our findings. Additionally, exploring models that incorporate geospatial aspects is advised. This approach would contribute to the development of region-focused models, providing a better understanding of the factors influencing programming literacy in specific regions.

It is recommended that future research building on this study address the identified limitations. Upon resolving these, further investigation could meaningfully focus on predicting GDP by leveraging programming literacy as a significant variable, thereby deepening the understanding of the interplay between these two dimensions. Additionally, it is crucial to formulate policy implications drawn from our models' results wherefore here, too, future research is advised.

References

Eurostat. 2022. *Statistical Regions in the European Union and Partner Countries: NUTS and Statistical Regions 2021*. European Union. https://ec.europa.eu/eurostat/documents/3859598/15193590/KS-GQ-22-010-EN-N.pdf/82e738dc-fe63-6594-8b2c-1b131ab3f877?t=1666687530717.

Eurostat. 2023. "Gross Domestic Expenditure on R&D by Sector of Performance, Sex and NUTS-2 Regions." Accessed November 28, 2023. https://ec.europa.eu/eurostat/cache/metadata/en/rd_esms.htm.

Eurostat. 2023. "High-tech Industry and Knowledge-intensive Services - Employment in Technology and Knowledge-intensive Sectors by NUTS-2 Regions and Sex." Accessed November 28, 2023. https://ec.europa.eu/eurostat/cache/metadata/en/htec_esms.htm.

Eurostat. 2023. "Households with Access to the Internet at Home." Accessed November 28, 2023. https://ec.europa.eu/eurostat/cache/metadata/en/isoc_i_esms.htm.

Eurostat. 2023. "Human Resources in Science and Technology (HRST) by Category and NUTS-2 Regions." Accessed November 28, 2023. https://ec.europa.eu/eurostat/cache/metadata/en/hrst_esms.htm.

Eurostat. 2023. "Structural Business Statistics SBS Data by NUTS-2 Regions and NACE Rev. 2." Accessed November 28, 2023. https://ec.europa.eu/eurostat/cache/metadata/en/sbs_esms.htm.