ISEG Lisbon School of Economics & Management
Universidade de Lisboa

MASTER IN

DATA ANALYTICS FOR BUSINESS

MASTER'S FINAL WORK

DISSERTATION

ANALYSIS OF SOCIODEMOGRAPHIC FACTORS INFLUENCING STUDENTS' DATA VISUALIZATION LITERACY

MARIANA ROCHA CRUZ

MARÇO - 2023

Lisbon School of Economics & Management
Universidade de Lisboa

MASTER IN

DATA ANALYTICS FOR BUSINESS

MASTER'S FINAL WORK

DISSERTATION

ANALYSIS OF SOCIODEMOGRAPHIC FACTORS INFLUENCING STUDENTS' DATA VISUALIZATION LITERACY

MARIANA ROCHA CRUZ

ADVISOR:
JOÃO A. BASTOS

MARÇO - 2023

The purpose of visualization is insight,
not pictures.
—Ben Shneiderman

GLOSSARY

CFS Current field of studies. i, ii, v, 8, 20–23, 25, 28, 30

CLE Current level of education. i, ii, v, 8, 17, 19–23, 25, 28, 29

DVL Data Visualization Literacy. i, ii, iv, 1–31, 33–44

FSHS Field of studies in High School. i, ii, v, 8, 19, 20, 22, 23, 25, 28, 30

GVIF Generalized Variance Inflation Factor. i, 17

HL Hosmer and Lemeshow. i, 18, 23–27

LR Likelihood Ratio. i, 18, 23–27

OR Odds Ratio. i, 17

VIF Variance Inflation Factor. i, 17

ABSTRACT

The rapid pace at which data is created has required the creation of new tools to extract information from large amounts of data. Data visualization has proven effective in facilitating access to essential information from a dataset. For this reason, it is critical to examine Data Visualization Literacy (DVL), particularly in the context in which learning occurs, in schools.

Studies related to this topic have been consulted, however, the research done so far to understand the influence of sociodemographic factors on the ability to read, interpret, and draw conclusions from data visualizations has not reached a consensus. Therefore, this study aims to bridge the controversy surrounding the topic by examining whether Age, Sex, Field of studies in High School (FSHS), Current level of education (CLE), and Current field of studies (CFS) predict students' responses to data visualization questions.

In this study, data collection was done through an online survey, which not only contained questions about the sociodemographic characteristics of the students, but also a section intended for data visualization questions. The non-probability convenience sampling technique was used and after processing the collected data, a total of 153 responses were obtained. To analyze the data, 6 binary logistic regressions were developed, each referring to one of the 6 data visualization questions contained in the survey, in order to compare the findings of this study with those previously supported by other authors.

The results suggest that all variables except CLE were important factors in predicting students' ability to answer the data visualization questions correctly.

KEYWORDS: Data Visualization; Data visualization Literacy; Binary Logistic Regression

TABLE OF CONTENTS

List of Figures

LIST OF TABLES

Acknowledgements

I would like to sincerely thank my advisor, Professor João Bastos, for his guidance throughout this project.

I would also like to extend my appreciation to my family and friends, especially my parents, for their unconditional motivation and encouragement in the most challenging moment.

Finally, I am particularly grateful for the love and support given by my boyfriend, Sérgio Nascimento.

1 INTRODUCTION

Data has been a crucial driver of economic and social development due to its increasing presence and relevance (BSA, 2015). According to IDC (International Data Corporation) research, the global data sphere is predicted to reach 175 Zettabytes by 2025 (Rydning et al., 2018). As a result, the term "Big Data" is becoming more prevalent. Big data is characterized by its large volume, high velocity, and great variety and due to these characteristics, traditional data processing tools are not equipped to collect or process the data efficiently. Therefore, dealing with massive volumes of data is not a simple procedure, demanding new and more advanced methods of analysis to derive useful and meaningful information (Hariri et al., 2019; Pilania, 2021).

Displaying the information in a way that enables people to understand the key takeaways and their business value is a challenge. Although individuals can recognize patterns, trends, and relationships in data, they still struggle with massive volumes of data (Manyika et al., 2011). In these circumstances, however, visualization has successfully highlighted crucial information and amplified cognition (Card, 1999; Keim et al., 2013). Instead of data being presented in complex spreadsheets, it can be displayed visually, using visual elements such as maps and graphs. This is referred to as Data Visualization.

Literacy, according to UNESCO (n.d.), is a set of skills such as identifying, interpreting, communicating, and computing using printed and written resources related to various situations. Literacy implies continuous learning, allowing the development of the individual's potential, so that they can achieve their goals and participate more actively in the community and society (UNESCO, n.d.). This study will focus on data visualization literacy (DVL), which Lee et al. (2017) define as an individual's capacity to read, interpret, and draw conclusions from data visualizations.

In a world surrounded by data and given the ease of access to mobile devices and applications, many businesses are betting on simple visualizations to reach audiences (Lee et al., 2020). As a result, users are more likely to interact with data visualizations in various contexts, for example in a financial context, graphs showing their personal finances, in a business context, sales performance dashboards, and in a health context, covid-19 infographic dashboards. Not surprisingly, several researchers clarified how vital it is to correctly interpret data visualizations for informed daily decision making and understanding the world around us (Friel et al., 2001; Börner et al., 2016). According to research by Börner et al. (2016), individuals in general have a low level of data sisualization literacy (DVL). Since literacy entails continuous learning, it is critical to examine the setting in which this learning occurs, namely schools, as well as its intended audience, students.

Several studies have been conducted on students' DVL and although the variables Age, Gender and Education are often present in them, there are conflicting views about their significance (Binali et al., 2022; Börner et al., 2016; Curcio, 1987; Lowrie and Diezmann, 2007, 2011; Ludewig et al., 2020; Wainer, 1980). As well as an overall lack of research about the impact of educational background on students' DVL, so far, previous studies have only focused on differences in students' abilities to read and interpret data visualizations between educational levels (Binali et al., 2022; Lowrie and Diezmann, 2007; Wainer, 1980).

Hence, this study falls within the category of data visualization, especially data visualization literacy (DVL). This research aims to understand whether student's sociodemographic characteristics predict their ability to read, interpret and draw conclusion from data visualizations. In this regard, the study seeks to answer the following question: Are Age, Sex, Field of studies in High School, Current level of education, Current field of studies factors that predict the students' answers to the data visualization questions?

This study will contribute to the body of knowledge on DVL by identifying and assessing which factors influence students' DVL. This will assist in understanding how students can be helped to improve these abilities, in an educational and training context. Consequently, in the medium to long term, it will provide real value to organizations, as individuals will be better prepared to create and make use of data visualizations and the valuable information contained within them.

To address the above-mentioned objective, this dissertation is divided as follows: in the next chapter, which corresponds to the literature review, a framework is given of the topics underlying the research problem, data visualization and DVL, as well as a of the main factors that predict the student's DVL. At the end of this chapter, the conceptual framework is also presented. The third chapter discusses the research methodology, including the research design, source of population, data collection method and data entry and analysis applied. In the fourth chapter, the features of the variables used in this study are explained, the data obtained is analyzed and the results found are discussed. The fifth and last chapter summarizes the main conclusions, contributions, limitations, as well as suggestions for future research.

2 LITERATURE REVIEW

2.1 Data Visualization

Although it is not possible to precisely date the beginning of information visualization, it is thought to have begun with the origin of the human species and their first prehistoric rock paintings (Friendly, 2008b). In his chronological study, Friendly (2008b) concludes that the oldest known map is a Neolithic wall painting dated 6200 BC and represents part of a town in Turkey. According to Few and Edge (2007) the most ancient, preserved table arose later, in the 2nd century in Egypt, when there was a need to organize information about the celestial bodies.

Only centuries later, specifically in the 17th century, René Descartes developed a two-dimensional coordinate system capable of representing data, known as the Cartesian coordinate system (Few and Edge, 2007). The progress in visualization approaches continued, in the 18th century, Playfair (2005) pioneered the creation of statistical graphics, the bar chart, line graph, and pie chart, which today are widely known and used. The development of new technologies and increasingly complex graphical representations, the advances in the study of statistics, and the evolution of visual thinking created the perfect conditions for the exponential growth of statistical graphics (Friendly, 2008a). Resulting in the "Golden Age of Statistical Graphics," the name by which the first half of the 19th century became known (Friendly, 2008a, p. 502). In his paper, Friendly (2008a), mentions some contributions to the history of statistical graphics, which are examples of what was produced in the formerly described golden age. Charles Joseph Minard was one of the authors of these contributions to the field of information graphics (Friendly, 2008a). Although he is responsible for the creation of several notable graphs, the one for which he is most remembered is his flow map portraying the catastrophic march of Napoleon's army in the 1812 Moscow campaign (Friendly, 2002).

However, in the early 20th century, the preference for formal models and their precision over images would give rise to the "modern dark ages" of visualization (Friendly and Denis, 2000, p. 53). Friendly (2008b) believes that John W. Tukey's contributions have helped overcome this decline in popularity. Tukey (1962) introduced, exploratory data analysis (EDA), a new method of exploring and analyzing information. This method tries to find clues and patterns in the information by looking closely at numbers and graphs, using both visual and quantitative approaches (Hoaglin et al., 1985).

Afterward, Tufte (1983) published his book *"The Visual Display of Quantitative Information"*, in which the author aims to convey best practices that should be considered for developing accurate and effective data visualizations. Since then, great attention has been devoted to presentation graphics (Chen et al., 2008).

Another author who impacted the field of data visualization was Card (1999) who published his book, *"Readings in Information Visualization: Using Vision to Think"*, making the practice of representing information in a meaningful and visual way more accessible to the public.

At the present time, the increase in the amount of data boosted the importance of data visualizations and tools used for its development and analysis (Womack, 2014). Due to collaboration with a broad range of disciplines, data visualization is evolving at a fast pace (Friendly, 2008b). Resulting, in the current wide range of data visualization tools and software options available, from point-and-click interfaces such as Excel and Tableau to programming languages such as R and Python. These tools are beneficial in a wide range of scenarios, for example, assisting in better-informed decision-making, monitoring operations, and assessing the market position of firms (Aparicio and Costa, 2015).

2.2 Data Visualization Literacy

Friel et al. (2001) and Börner et al. (2016) recognize that understanding and interpreting data visualizations is necessary for daily life. Börner et al. (2016) go so far as to emphasize the importance of being able to read and create data visualizations, equating it to the importance of being able to read and write text. Womack (2014) states that improving the understanding of data visualization, beyond a simplistic understanding, would bring benefits to the individuals' analytical skills and therefore improve their knowledge about the world around them. Alper et al. (2017) further state that limitations in skills related to data visualization literacy (DVL) can represent a barrier to accessing information and, as a result, making educated judgments.

DVL has been differently defined by researchers. While Börner et al. (2016) define it as an individual's ability to interpret and make sense of patterns, trends, and correlations in data visualizations. Lee et al. (2017) explained it as an individual's ability to read, interpret, and extract information from data visualizations.

There have been several efforts to investigate the differences between beginners and experts in data visualization (Elias and Bezerianos, 2011; Maltese et al., 2015). Maltese et al. (2015) conducted a study to understand how students, throughout their academic path, developed DVL skills. Elias and Bezerianos (2011) through the development of a prototype system for the creation and customization of visualization dashboards, sought to understand the differences between a novice and an expert when interacting with the system.

Aspects influencing people's interactions with visualizations were also explored (Binali et al., 2022; Friel et al., 2001; Kennedy and Hill, 2018; Peck et al., 2019; Peppler et al., 2021). Peck et al. (2019) conducted a study in which they attempt to understand what factors influence people's perceptions of data visualizations. This study was motivated by the under-representation of the rural population in the data visualization literature. Friel et al. (2001) attempted to find out what factors seem to impact the interpretation of statistical graphs. Binali et al. (2022) investigated if the interpretation of graphs by students in scientific and daily scenarios varied between educational levels. Kennedy and Hill (2018) studied the impact of emotions on users' interaction with data and their visualizations. Peppler et al. (2021) evaluated the impact of museum display design on visitors' engagement with data visualization literacy (DVL).

Studies about visualizations whose users were unfamiliar with, also contributed to the literature (Lee et al., 2015; Ruchikachorn and Mueller, 2015). Whereas Lee et al. (2015) showed a special interest in the cognitive activities of users when trying to understand types of visualizations with which they had no previous interaction, Ruchikachorn and Mueller (2015) proposed learning unfamiliar visualizations by visualization morphing.

Other studies have focused on the memorability of visualizations (Bateman et al., 2010; Borkin et al., 2013). Borkin et al. (2013) carried out a large-scale analysis to identify the features that contribute to the memorability of visualizations. As a result, their research helps third parties to create visualizations effectively. Bateman et al. (2010) investigated charts' visual embellishment and memorability, questioning the use of minimalist approaches when creating data visualizations.

Some researchers have contributed to the progress of data visualization literacy (DVL) assessment (Boy et al., 2014; Börner et al., 2016; Lee et al., 2017). Börner et al. (2016) sought to determine the general public's familiarity with data visualizations, by using 20 visualizations possible to encounter in everyday activities. Boy et al. (2014) proposed a set of tests, applying item response theory (IRT), for line graphs, bar charts, and scatterplots and how to conduct them to obtain an individual's visualization literacy level. Lee et al. (2017) have developed a test, specialized for non-expert users in the data visualization field, that captures various areas and typologies of visualizations to assess DVL.

2.3 Factors Predicting Student's Data Visualization Literacy

This section presents the underlying literature for determining the most relevant variables related to Student's DVL. In the following subsections, the three main variables identified in relevant research are addressed: Sex, Age, and Education.

2.3.1 Age as a Factor Predicting Student's Data Visualization Literacy

According to several researchers, significant limitations in reading, comprehending, and interpreting data visualizations impact both children and adults (Börner et al., 2016; Maltese et al., 2015; Shah and Hoeffner, 2002; Shah et al., 1999).

Despite Ludewig et al. (2020) finding that age is not a significant factor in predicting students' ability in reading and interpreting graphs, some studies disagree and found it to be important (Binali et al., 2022; Börner et al., 2016). Börner et al. (2016) sought to understand whether individuals were familiar with the various types of data visualization provided, having concluded that adults recognized the visuals more frequently than youths. Furthermore, Binali et al. (2022) believe that the different difficulties that college and high school students have when interpreting graphs are related to their age.

In the literature reviewed, there seems to be no agreement on the effect of age on individuals' interpretation of data visualization. For this reason, there is interest in including the variable Age in the study.

2.3.2 Sex as a Factor Predicting Student's Data Visualization Literacy

Analyzing and comprehending information graphics entails interpreting information displayed in a visual-spatial format, hence, it relies on spatial abilities (Lowrie and Diezmann, 2007).Although differences in spatial abilities between males and females are widely acknowledged there is significant debate over the extent, nature, and age at which these differences first appear (Linn and Petersen, 1985).

A series of previous studies have suggested that gender differences in students' ability to read and interpret graphs may exist (Lowrie and Diezmann, 2007, 2011; Ludewig et al., 2020).

Some researchers found that on more difficult mathematical tasks, male students tended to outperform female students (Bielinski and Davison, 1998; Lowrie and Diezmann, 2011; Penner, 2003). Although Bielinski and Davison (1998) have also concluded that female students outperformed male students on easier mathematical tasks, there is an inconsistency with this argument, as Lowrie and Diezmann (2011) findings contradict it for graphics tasks.

However, there are also studies that do not recognize gender differences in students' ability to read and interpret graphs (Binali et al., 2022; Curcio, 1987; Ludewig et al., 2020).

Given the literature reviewed, the controversy of the issue is clear. Therefore, the variable Sex will be included in this study. Blakeman (2020) draws attention to the interchangeable use of the variables Sex and Gender. In this study, the variable Sex will be used since the goal is to examine differences in the target population, students, based on their biological differences.

2.3.3 Education as a Factor Predicting Student's Data Visualization Literacy

According to Shreiner (2018) findings, students' encounters with data visualizations will become more frequent as they move through the academic levels. To ensure that students do not struggle and are not hampered in extracting the necessary information from new data visualizations, it is important that they receive adequate instruction to develop basic data literacy skills (Shreiner, 2018)

A considerable amount of literature has highlighted the importance of students being exposed to the basics of data visualization in their education (Binali et al., 2022; Börner et al., 2016; Lowrie and Diezmann, 2007; Ludewig et al., 2020; Maltese et al., 2015; Shah and Hoeffner, 2002; Womack, 2014). Nevertheless, some researchers suggest that students may lack adequate instruction in schools (Coleman, 2010; McTigue and Flowers, 2011). Coleman (2010) studied elementary school teachers' instructional practices involving graphs and concluded that not only were verbal texts given more attention than graphical representations but also, that the most reported practice regarding graphs was to point at the visualizations, often without providing instruction on how to interpret them. In aiming to explore the significance students placed on graphical information in science texts, and what factors made them more complex to understand, McTigue and Flowers (2011) discovered that teachers did not teach graphs directly.

Several researchers have shown concern about the teaching methods of data visualization and have pointed out the need for improvements in teaching (Ludewig et al., 2020; Maltese et al., 2015; McTigue and Flowers, 2011). Although their studies focused on different populations, Ludewig et al. (2020), investigated high school students, and Maltese et al. (2015), studied college students, both reached similar findings regarding how educators should behave while teaching graphical representations. Educators should not assume that students have sufficient prior knowledge to be able to comprehend and interpret data visualizations provided in class (Ludewig et al., 2020; Maltese et al., 2015). Maltese et al. (2015) further state that if educators aim to improve their students' data visualization skills, they should consider what materials to use when planning lessons. In the field of science, McTigue and Flowers (2011) explain that although students are currently more exposed to visualizations, it's not guaranteed that they comprehend them. The researcher also argues that the idea that visualizations are intuitive to interpret is incorrect and encourages teachers to pay greater attention to this issue (McTigue and Flowers, 2011).

Regarding the influence of education on DVL, some researchers have concluded that, when students move on to higher levels of education, their ability to read and interpret graphs improves (Binali et al., 2022; Lowrie and Diezmann, 2007; Wainer, 1980). Wainer (1980) found improvements in the ability to read and interpret graphs from third to fourth grade. However, the findings differ from fourth to fifth grade, Lowrie and Diezmann (2007) noticed significant improvements in these abilities, while Wainer (1980) found little difference. In the context of higher education levels, Binali et al. (2022), discovered that college students outperformed high school students in graph interpretation in both scientific and daily scenarios.

Given the literature review, there is motivation to further explore the influence of educational background on students' data visualization literacy (DVL). With the purpose of making a deeper analysis, three variables will be included in the study: Field of Studies in High School (FSHS), Current Level of Education (CLE), and Current Field of Studies (CFS).

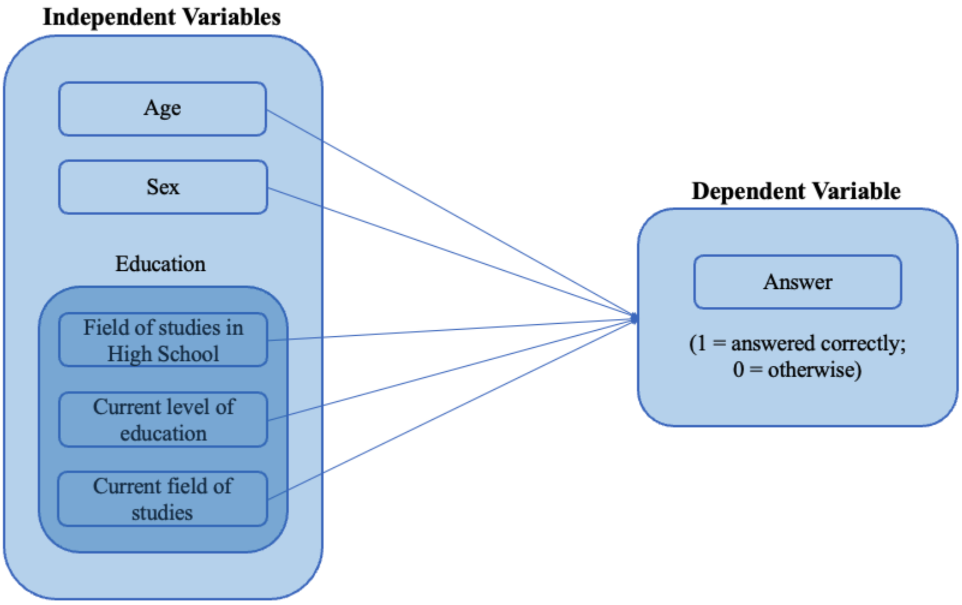


FIGURE 2.1: Conceptual Framework

3 METHODOLOGY

3.1 Research Design

The present research is guided by a positivist philosophy, which means that a social reality that can be witnessed and is independent of the researcher is employed to obtain a final research result that is comparable to ideas formerly developed (Saunders et al., 2019).

A deductive approach is used, since the previously examined literature serves as the foundation of explanation, allowing for the anticipation of events (Saunders et al., 2019).

The research method adopted is the quantitative mono-method and the research horizon is cross-sectional since the goal is to collect, economically and at a single point in time, a large amount of data from a sample of the population using a survey strategy (Saunders et al., 2019).

3.2 Source of Population

The target population for data collection consists of individuals of both sexes, over the age of 18 who are students. Individuals who were not students were excluded.

An attempt was made to reach a heterogeneous population in age, sex, level of education, and field of studies.

Considering financial and time constraints, only a subset of the target population was considered, the target sample, which was obtained through a non-probabilistic convenience sampling technique (Saunders et al., 2019).

3.3 Data Collection Method

For data collection, a structured survey, developed in Qualtrics software and both available in Portuguese and English, was adopted and distributed through a web link, to reach a large number of individuals at a low cost (Saunders et al., 2019). This link was shared on Instagram, LinkedIn, and Whatsapp, as well as by email, between December 16, 2022 and January 24, 2023.

The survey is divided into 2 sections, the first section aims to collect student's sociodemographic information, through multiple choice questions with a mandatory response. The following section aims to assess the students' data visualization literacy (DVL), and consists of six multiple-choice questions, with a non-mandatory but time-limited response. The first two questions were limited to two minutes and the remaining questions to two minutes and thirty seconds.

The questions contained in the survey's DVL section were designed so that the degree of difficulty increased from the first question to the second and so on. Since the difficulty of each question is subjective and each student perceives it differently, the rationale behind the order chosen for the questions in the survey is explained as follows.

Prior knowledge, according to Freedman and Shah (2002), is triggered during the early processing of graphics. The researchers recognize that comprehension is effortlessly done when the information is clearly represented by the visualization and can be easily related to prior knowledge. However, when the information is not clearly represented by the visualization or the individual lacks the necessary prior knowledge, comprehension becomes challenging.

Crato (2006), in turn, points out that there are evident precedents in mathematics, thus the most elementary concepts must be properly understood before moving on to more complex concepts.

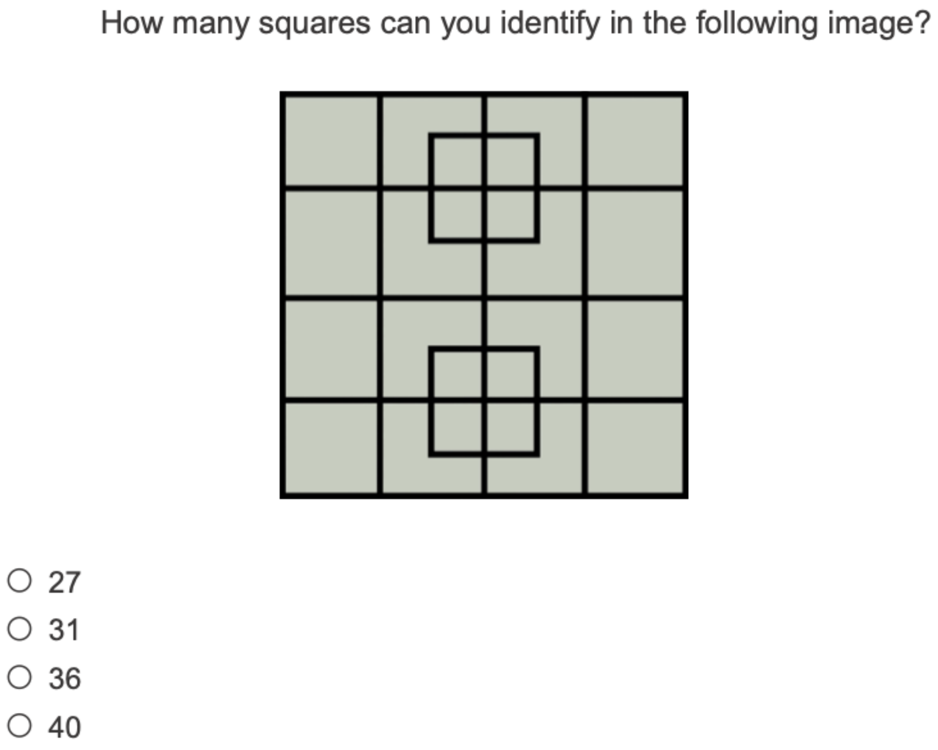
The assumption is made that a question involving more complex concepts requires prior knowledge of a greater amount of elementary concepts. Therefore, this question will be more difficult, since it is more likely that some of the elementary concepts have not been properly consolidated.

Relevant documents on the Mathematics curriculum in Portugal were consulted, mainly those concerning the essential learning of students from 1st grade to 9th grade. As a result, it was possible to identify the order in which students learn the concepts explored in the survey, and consequently those that require more prior knowledge.

According to Educação (2018a), in the second grade, students begin to make use of bar charts to solve problems. Later, in the fifth grade, continuous variables and line charts are introduced (Educação, 2018b). However, only in the sixth grade, students are able to collect, organize and represent information using line charts (Educação, 2018c). As stated in Educação (2018d), in the 7th grade, students develop the ability to understand statistical measures, such as the median, quartiles, and interquartile range. Nevertheless, only in the 8th grade are they able to analyze and interpret the information contained in a data set using these measures (Educação, 2018e).

In light of the above, the image of the first question in the survey's data visualization literacy (DVL) section, shown in Figure 3.1, was selected based on psychotechnical tests, commonly used in academic contexts to assist students in recognizing their areas of interest, since the objective was to choose an image that was accessible to everyone and did not require prior knowledge in any field.

The second question in the survey's DVL section, shown in Figure 3.2, comprises a line chart, followed by the third question, shown in Figure 3.3, which contains a combination of line and bar charts. Although the prior knowledge requirements for the second and third questions are equal, the third question requires more information to be interpreted since it involves more than one type of chart, hence being more demanding.



How many squares can you identify in the following image?

○ 27

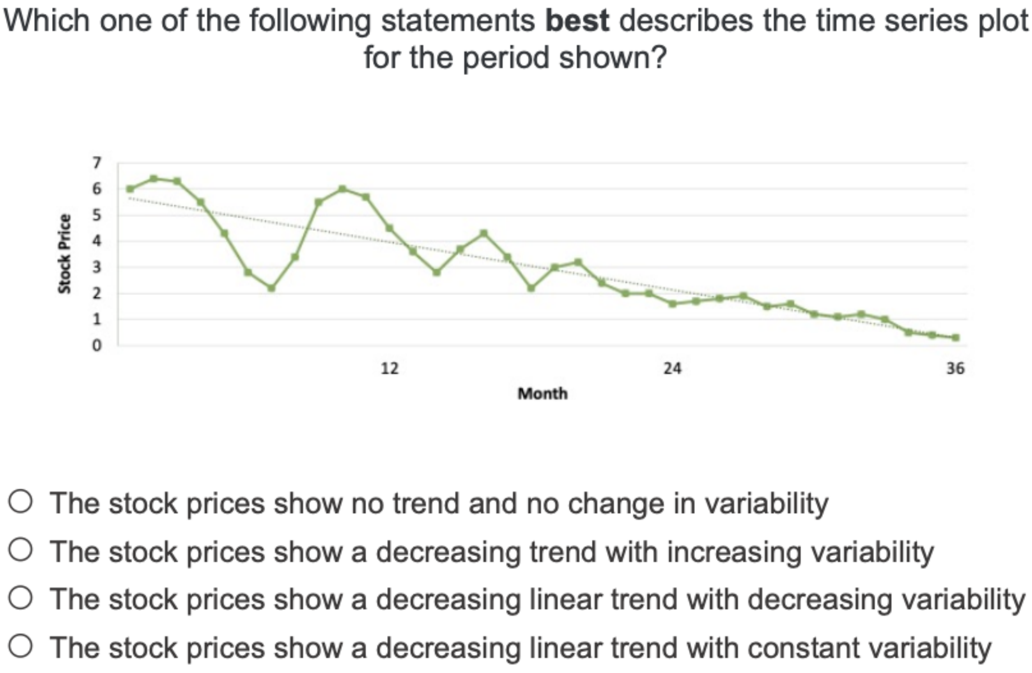
○ 31

○ 36

○ 40

FIGURE 3.1: First Data Visualization Question

Source of the image used in the question: (Nelson and Jones, 2023)



Which one of the following statements **best** describes the time series plot for the period shown?

○ The stock prices show no trend and no change in variability

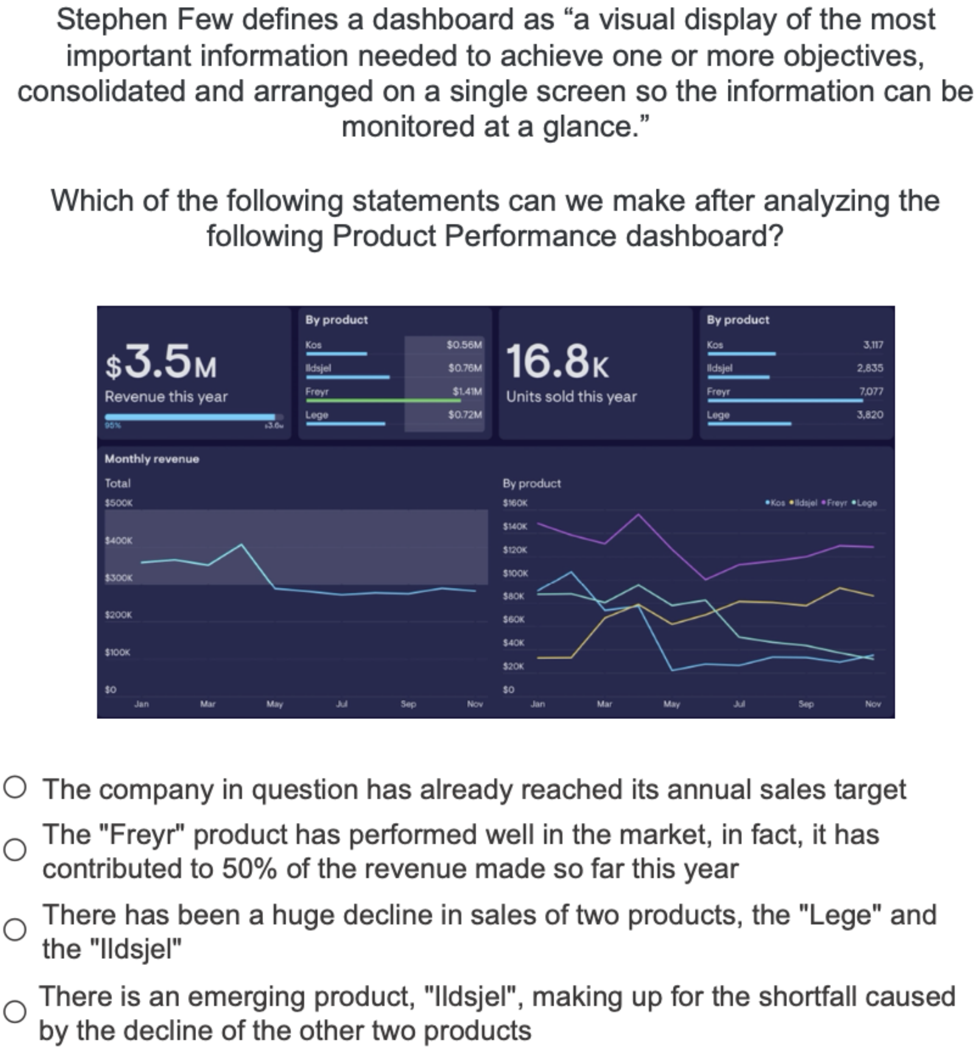
○ The stock prices show a decreasing trend with increasing variability

○ The stock prices show a decreasing linear trend with decreasing variability

○ The stock prices show a decreasing linear trend with constant variability

FIGURE 3.2: Second Data Visualization Question

Source of the image used in the question: (Own work)



Stephen Few defines a dashboard as “a visual display of the most important information needed to achieve one or more objectives, consolidated and arranged on a single screen so the information can be monitored at a glance.”

Which of the following statements can we make after analyzing the following Product Performance dashboard?

○ The company in question has already reached its annual sales target

○ The "Freyr" product has performed well in the market, in fact, it has contributed to 50% of the revenue made so far this year

○ There has been a huge decline in sales of two products, the "Lege" and the "Ildsjel"

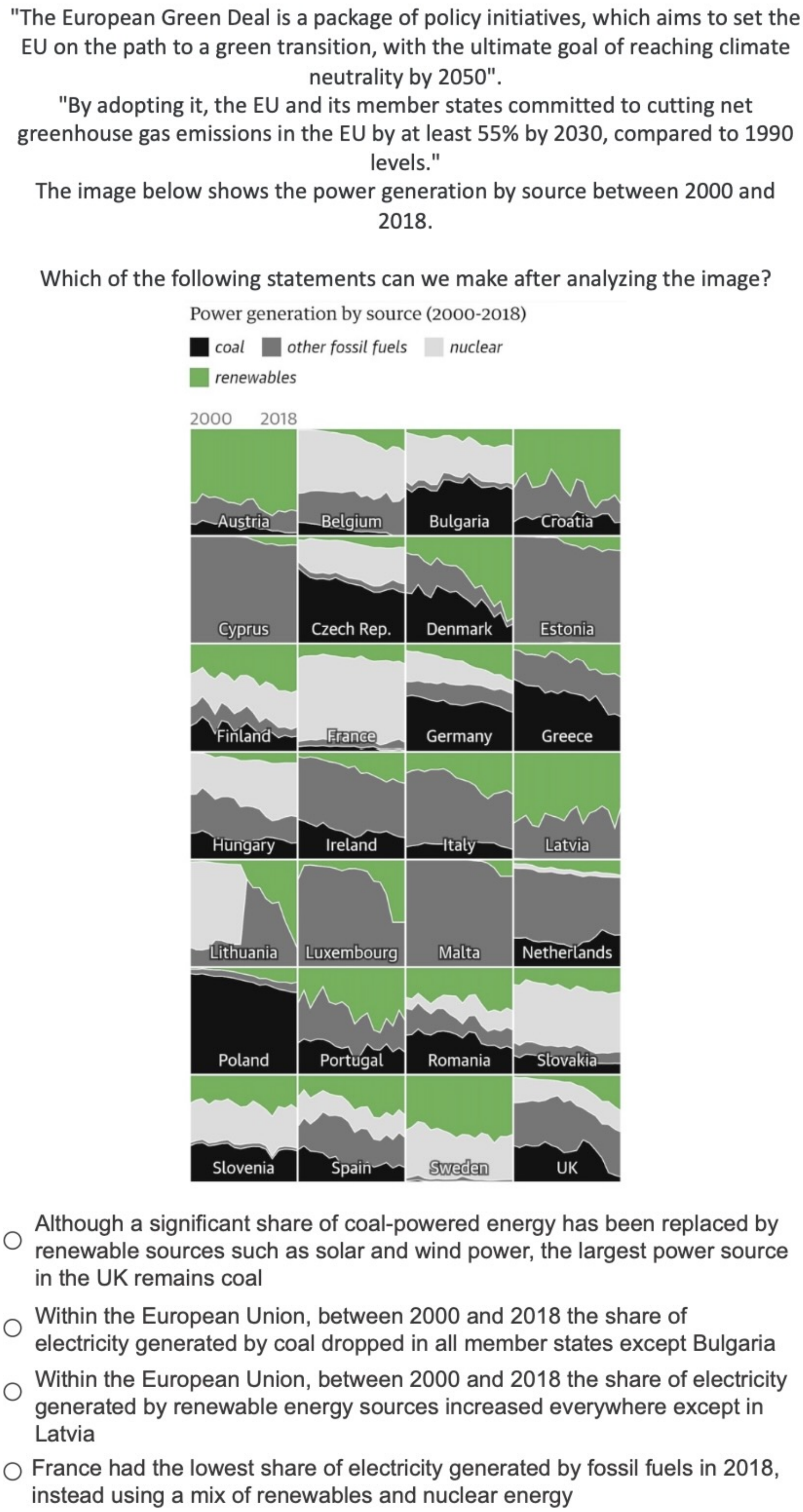
○ There is an emerging product, "Ildsjel", making up for the shortfall caused by the decline of the other two products

FIGURE 3.3: Third Data Visualization Question

Source of the image used in the question: (Geckoboard, n.d.)

The fourth question in the survey's data visualization literacy (DVL) section, shown in Figure 3.4, covers a number of area charts, a combination of line chart and bar chart. The fifth question, shown in Figure 3.5, contains a set of box plots, a subject that is not taught until after the bar charts and line charts. At last, in the sixth question, shown in Figure 3.6, a single filled map, also known as cloropleth, is used to relate two sets of data. This type of visualization is less likely to be encountered by students, therefore the sixth question was considered the most difficult.

Before sharing the survey, a pre-test was conducted by convenience on a sample of 17 students to confirm the heterogeneity of the students and the suitability of the questions (Saunders et al., 2019). Slight changes were made, and the final version of the survey is presented in Appendix A.1.



"The European Green Deal is a package of policy initiatives, which aims to set the EU on the path to a green transition, with the ultimate goal of reaching climate neutrality by 2050".
"By adopting it, the EU and its member states committed to cutting net greenhouse gas emissions in the EU by at least 55% by 2030, compared to 1990 levels."
The image below shows the power generation by source between 2000 and 2018.

Which of the following statements can we make after analyzing the image?

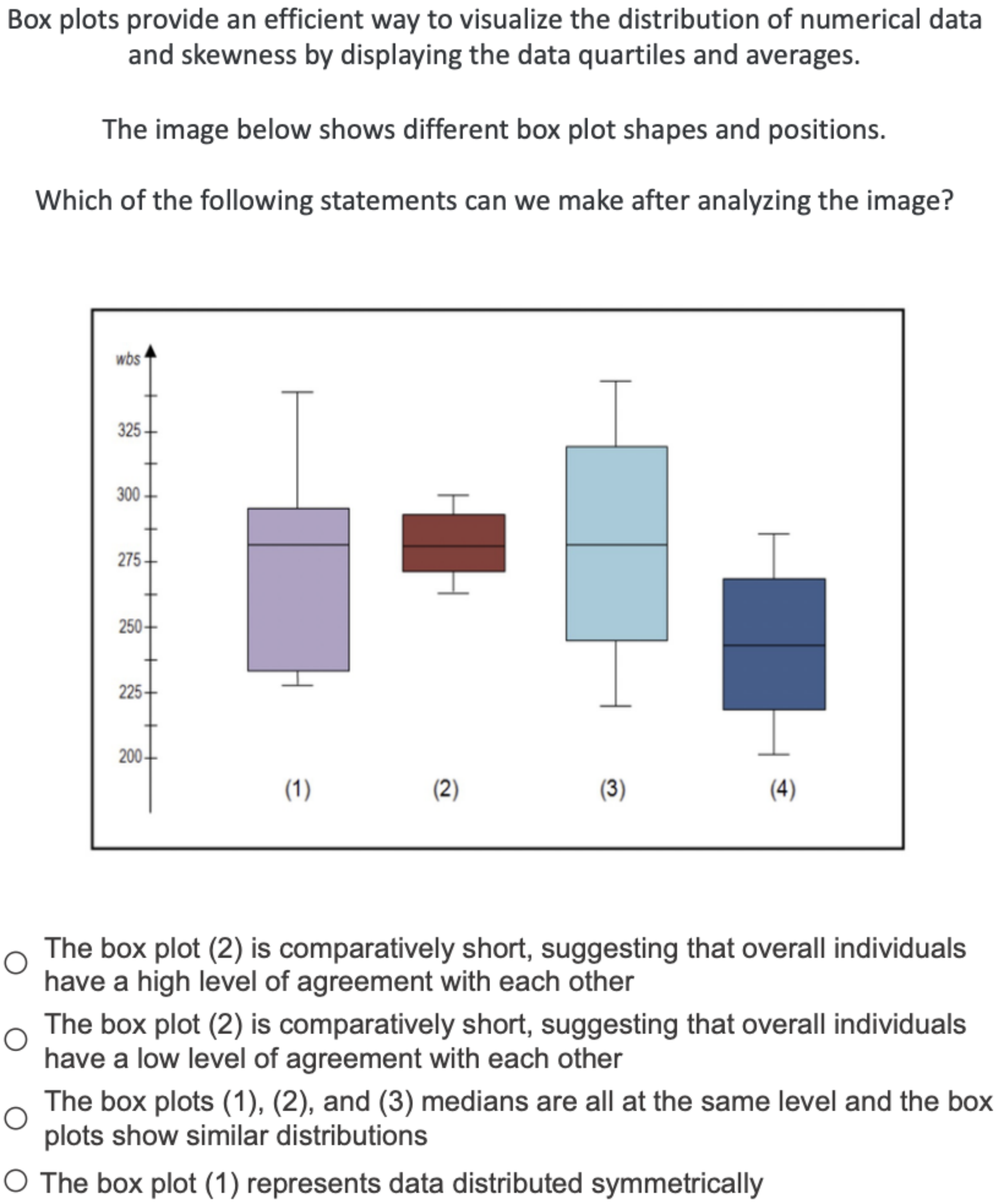
○ Although a significant share of coal-powered energy has been replaced by renewable sources such as solar and wind power, the largest power source in the UK remains coal

○ Within the European Union, between 2000 and 2018 the share of electricity generated by coal dropped in all member states except Bulgaria

○ Within the European Union, between 2000 and 2018 the share of electricity generated by renewable energy sources increased everywhere except in Latvia

○ France had the lowest share of electricity generated by fossil fuels in 2018, instead using a mix of renewables and nuclear energy

FIGURE 3.4: Fourth Data Visualization Question
Source of the image used in the question: (Sandbag, n.d.)



Box plots provide an efficient way to visualize the distribution of numerical data and skewness by displaying the data quartiles and averages.

The image below shows different box plot shapes and positions.

Which of the following statements can we make after analyzing the image?

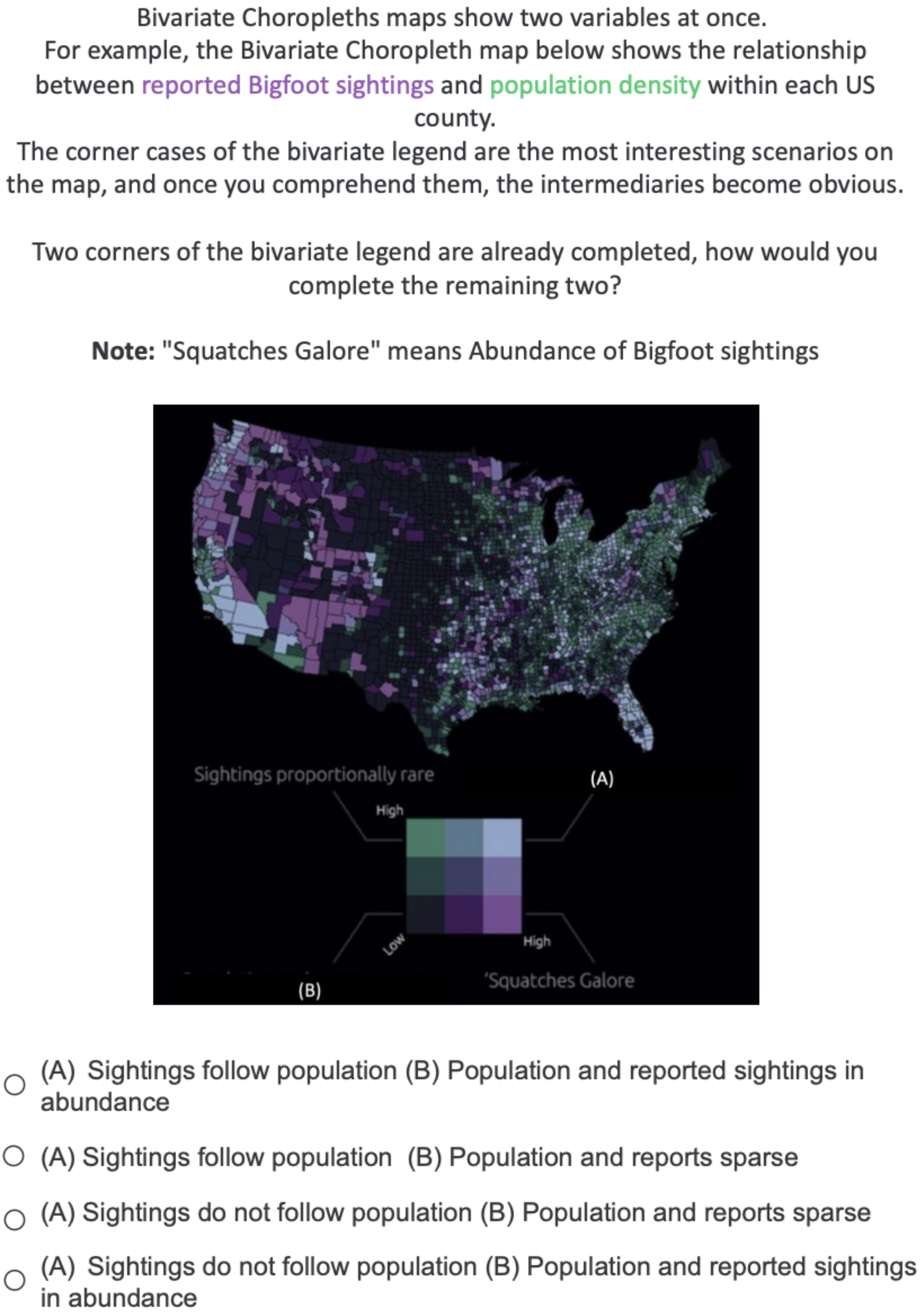
○ The box plot (2) is comparatively short, suggesting that overall individuals have a high level of agreement with each other

○ The box plot (2) is comparatively short, suggesting that overall individuals have a low level of agreement with each other

○ The box plots (1), (2), and (3) medians are all at the same level and the box plots show similar distributions

○ The box plot (1) represents data distributed symmetrically

FIGURE 3.5: Fifth Data Visualization Question
Source of the image used in the question: (Wellbeing@School, n.d.)



Bivariate Choropleths maps show two variables at once.
For example, the Bivariate Choropleth map below shows the relationship between reported Bigfoot sightings and population density within each US county.
The corner cases of the bivariate legend are the most interesting scenarios on the map, and once you comprehend them, the intermediaries become obvious.

Two corners of the bivariate legend are already completed, how would you complete the remaining two?

Note: "Squatches Galore" means Abundance of Bigfoot sightings

○ (A) Sightings follow population (B) Population and reported sightings in abundance

○ (A) Sightings follow population (B) Population and reports sparse

○ (A) Sightings do not follow population (B) Population and reports sparse

○ (A) Sightings do not follow population (B) Population and reported sightings in abundance

FIGURE 3.6: Sixth Data Visualization Question
Source of the image used in the question: (Stevens, 2013)

3.4 Data Entry and Analysis

A total of 348 responses were obtained from the survey, and the data were analyzed using R, a statistical and graphical computation software. The data were cleaned, and the results were filtered to obtain only the students who agreed to answer the survey and completed it in its entirety, thus obtaining a total of 153 valid answers.

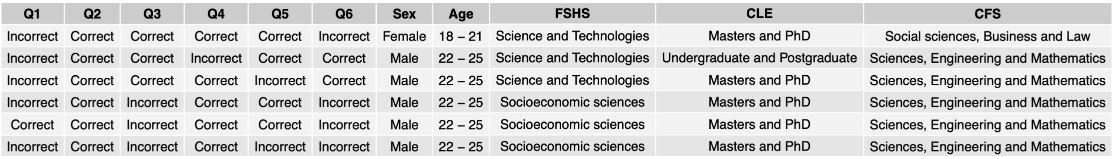
In relation to missing values, since the values of the independent variables come from questions whose answer is mandatory, no missing values were found. However, since the answers to the questions in the survey's data visualization literacy (DVL) section had a time limit and were non-mandatory, missing values were found for the dependent variables. Given that the dependent variables are binary, taking the value 1 if the student answers the question correctly and 0 otherwise, the missing values were recoded with the value 0.

Afterward, exploratory data analysis was performed, in which it was found that some levels of the categorical explanatory variables had few observations, thus there was an opportunity to combine similar levels of the variables. Details regarding the categorical variables and their levels are explored in the sub-chapter 4.1.

The present study used the analysis of 6 binary logistic regressions to investigate the influence of the explanatory variables on the response to each DVL question in the survey.

Although logistic regression ignores the linear relationship between the dependent and independent variables, it must nevertheless follow some assumptions. These are, the dependent variable is binary or dichotomous, the observations are independent, there is the absence of multicollinearity among the independent variables, there are no outliers, and at last, the continuous independent variables are linearly related to the log-odds (Menard, 2009).

Regarding this latter assumption, there is no relevance in testing it, since the independent variables used in this study are categorical. Assuming that each student answered the survey only once, and therefore appears only once in the sample, Figure 3.7 further supports the assumption that each observation is independent.



Q1	Q2	Q3	Q4	Q5	Q6	Sex	Age	FSHS	CLE	CFS
Incorrect	Correct	Correct	Correct	Correct	Incorrect	Female	18 – 21	Science and Technologies	Masters and PhD	Social sciences, Business and Law
Incorrect	Correct	Correct	Incorrect	Correct	Correct	Male	22 – 25	Science and Technologies	Undergraduate and Postgraduate	Sciences, Engineering and Mathematics
Incorrect	Correct	Correct	Correct	Incorrect	Correct	Male	22 – 25	Science and Technologies	Masters and PhD	Sciences, Engineering and Mathematics
Incorrect	Correct	Incorrect	Correct	Correct	Incorrect	Male	22 – 25	Socioeconomic sciences	Masters and PhD	Sciences, Engineering and Mathematics
Correct	Correct	Incorrect	Correct	Correct	Incorrect	Male	22 – 25	Socioeconomic sciences	Masters and PhD	Sciences, Engineering and Mathematics
Incorrect	Correct	Incorrect	Correct	Incorrect	Incorrect	Male	22 – 25	Socioeconomic sciences	Masters and PhD	Sciences, Engineering and Mathematics

FIGURE 3.7: First 6 rows of the data frame

Since the categorical independent variables used in this study have more than one degree of freedom, except for the variables Sex and Current level of education (CLE), the assumption that there is no multicollinearity between the independent variables, cannot be assessed with the Variance Inflation Factor (VIF). However, Fox and Monette (1992) introduced the Generalized Variance Inflation Factor (GVIF) and suggested that it should be used in the form of $(GVIF^{(1/2 \cdot Df)})$, where Df (degrees of freedom) is the number of levels of the categorical dependent variable minus one. The GVIF must be squared in order to apply the standard VIF criteria.

Although there is debate about the appropriate thresholds for assessing multicollinearity using the VIF, in this paper, Menard's (2002) suggestion will be accepted, that VIF values greater than 5 are cause for alarm, while VIF values greater than 10 indicate a serious case of multicollinearity. In summary, we can conclude that when $(GVIF^{(1/2 \cdot Df)})^2 > 5$, there is reason to suspect multicollinearity between the variables. Table 3.1 demonstrates that for all regressions, the $(GVIF^{(1/2 \cdot Df)})^2$ of each predictor variable is less than 5, as a result, the assumption that there is no multicollinearity between the independent variables is assured for all regressions.

	$(GVIF^{(1/2 \cdot Df)})^2$					
	1^{st} Data Visual	2^{nd} Data Visual	3^{rd} Data Visual	4^{th} Data Visual	5^{th} Data Visual	6^{th} Data Visual
Sex	1.172	1.216	1.240	1.186	1.183	1.274
Age	1.285	1.347	1.426	1.347	1.355	1.357
FSHS	1.115	1.124	1.115	1.084	1.137	1.184
CLE	1.512	1.628	1.670	1.566	1.720	1.638
CFS	1.079	1.099	1.081	1.055	1.105	1.077

TABLE 3.1: GVIF values for each variable in the Binary Logistic Regressions

The analysis of the standardized residuals displayed in Figure 3.8, verified that there were no observations whose absolute standardized residuals exceeded the value of 3.0, for any of the regressions, thus the assumption of no outliers in the regressions is confirmed (Anderson et al., 2016).

The results of the regressions are reported in tables, which first provide the coefficients (β), estimated using the maximum likelihood method, their standard errors (S.E.), Wald test values, p-values, Odds Ratio (OR), and 95% confidence intervals for the OR.

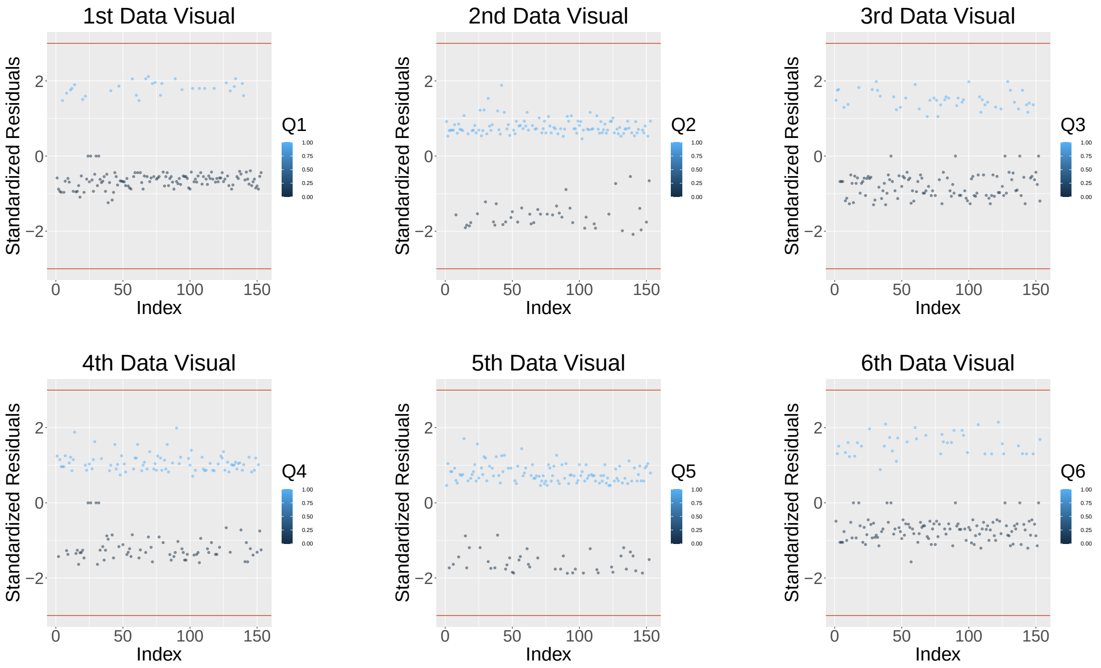


FIGURE 3.8: Standardized Residuals Plots of each Binary Logistic Regression

In this study, the Wald test is used to test the individual significance of the explanatory variables, that is, the following set of hypotheses:

$$H_0 : \beta_j = 0 \;\; vs \;\; H_1 : \beta_j \neq 0 \qquad j = 1, ..., p \tag{1}$$

The Null Hypothesis is rejected if the p-value is less than the significance level, thus existing strong evidence that the variable being tested is important to include in the model, given the other explanatory variables (Kleinbaum and Klein, 2010).

The tables also provide the results of the Likelihood Ratio (LR) and Hosmer and Lemeshow (HL) tests. The LR test is used to measure the improvement in the fit that the explanatory variables make relative to the null model. The significance of this test, in other words, the rejection of the null hypothesis, that the inclusion of all explanatory variables in the model does not add to predicting the dependent variable, occurs when the p-value is less than the significance level. Thus indicating that the full model is a significant improvement in fit over the null model (Kleinbaum and Klein, 2010). To test how well the data fit the model, the HL test, a goodness of fit test, is used. The non-significance of this test, that is, the failure to reject the null hypothesis, that there are no significant differences between the values predicted by the model and the observed values, occurs when the p-value is greater than the significance level. Thus suggesting that the model is adequate to fit the data (Kleinbaum and Klein, 2010).

4 RESULTS ANALYSIS

4.1 Exploratory Data Analysis

The categorical explanatory variables used in this study were carefully explored after the data were cleaned. It was observed that some levels of these variables had few observations, that is, they occurred infrequently. As a result, similar levels of the variables were combined, to obtain less disparate frequencies between levels.

Regarding the variable Age, only 10 observations were found for students aged between 26 and 29 years, and 11 observations for students aged 30 years or above. Therefore, these two levels were combined into one, encompassing students aged 26 or above. The distribution of the variable Age prior to and after this procedure is illustrated in Figure 4.1.

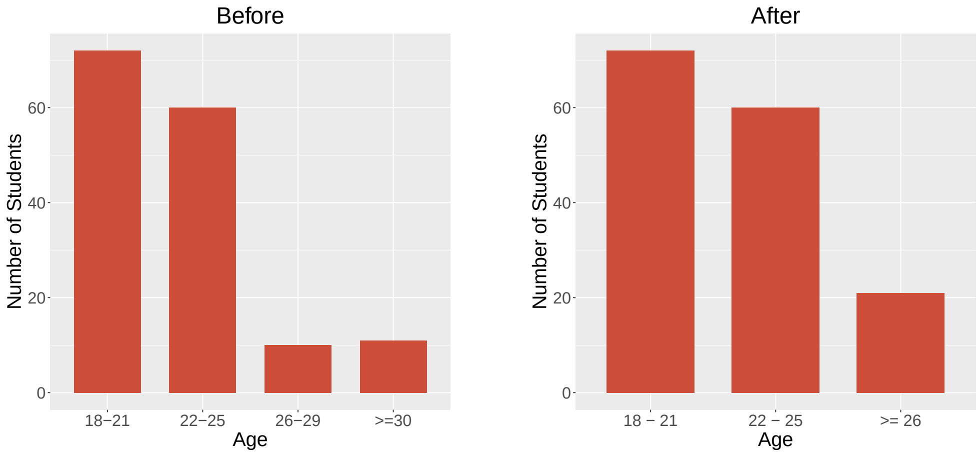


FIGURE 4.1: Variable Age before and after combining levels

Concerning the variable Field of studies in High School (FSHS), only 16 and 3 observations were found for students who attended either Languages and Humanities or Visual arts fields in high school, respectively. For this reason, these two levels were combined into one, Languages and Arts, which covers students who attended one of the two fields mentioned above. Figure 4.2 portrays the distribution of the variable FSHS before and after this process.

With respect to the variable Current level of education (CLE), only 4 observations were found for students currently pursuing a PhD and 3 observations for students currently pursuing a Postgraduate degree. As a result, the Masters and PhD levels were merged into one, Masters and PhD, which includes students from the two indicated educational levels. The Undergraduate and Postgraduate levels were also combined into one, Undergraduate and Postgraduate, which comprise students from both educational levels mentioned. The distribution of the variable CLE before and after this procedure is illustrated in Figure 4.3

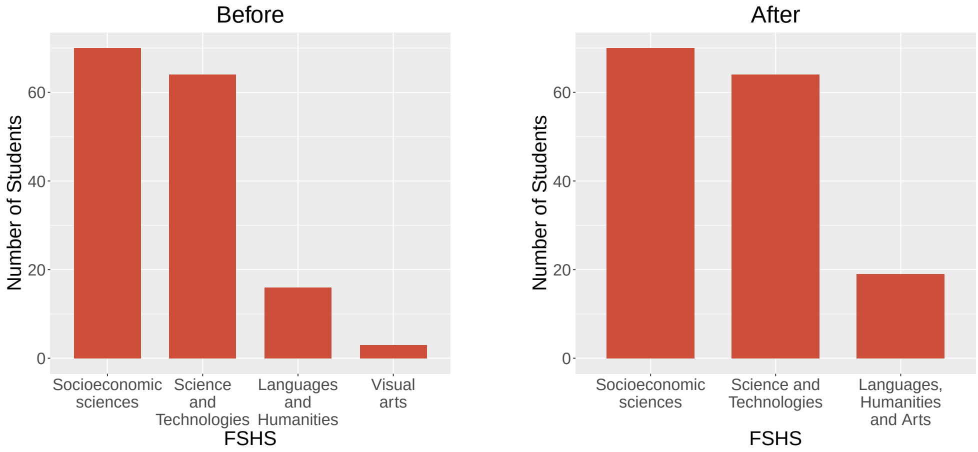


FIGURE 4.2: Variable Field of studies in High School (FSHS) before and after combining levels

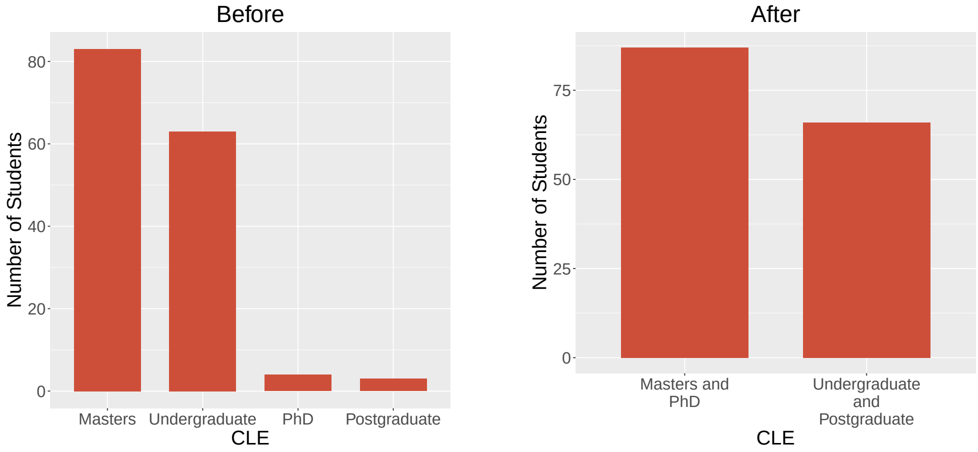


FIGURE 4.3: Variable Current level of education (CLE) before and after combining levels

Finally, in relation to the variable Current field of studies (CFS), only 4 observations were found for students currently attending the Engineering, Manufacturing, and Construction fields and 1 observation for students currently attending the General programs field. Thus, the General programs level and Arts and Humanities level were combined into one, Humanities and Arts, which includes students attending, at least, one of the previously mentioned fields. The Engineering, Manufacturing, and Construction level and the Science, Mathematics, and computing level were also combined into one, Sciences, Engineering, and Mathematics which encompasses students attending, at least, one of these fields. The distribution of the variable CFS before and after this procedure is portrayed in Figure 4.4

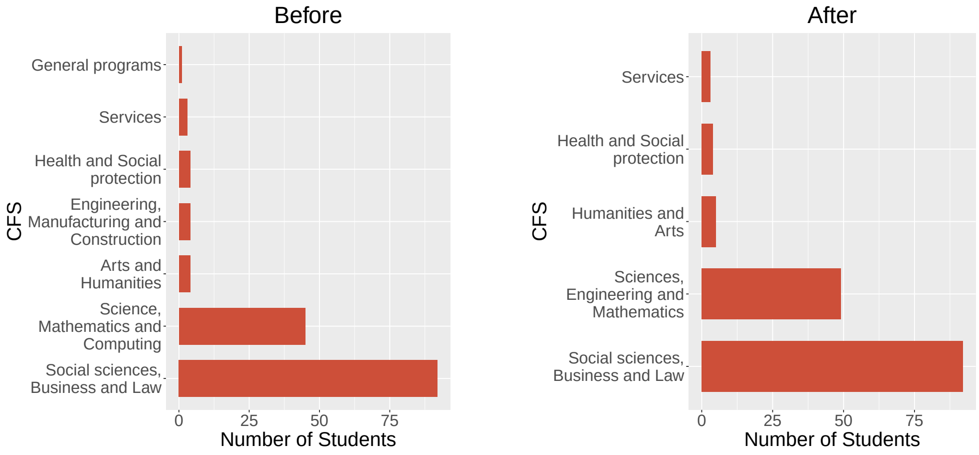


FIGURE 4.4: Variable Current field of studies (CFS) before and after combining levels

4.1.1 Dependent and Independent Variables

The purpose of this study is to determine what factors predict student responses to data visualization questions, hence, the dependent variables are binary (Q_i), each representing a question, that takes the value 1 if the student answers correctly and 0 otherwise:

$$Q_i = \begin{cases} 1 & , \ answered\ correctly \\ 0 & , \ otherwise \end{cases} \qquad i = 1, ..., 6 \tag{2}$$

Regarding the explanatory variables, Sex and Current level of education (CLE) are dichotomous categorical variables, meaning that they only have two categories, which take the values 0 and 1. The remaining explanatory variables are nominal categorical variables, with more than two levels. Considering k, the number of levels of the explanatory variable, as a rule, for these variables to be included in the model it would be necessary to create $k - 1$ dummy variables, corresponding to the $k - 1$ levels of the explanatory variable. The level to which no dummy variable is assigned corresponds to the reference level, which will serve to make comparisons with each of the $k - 1$ levels. However, the R software does not require dummy variables to be created, as long as the categorical variables are properly transformed into factors. Regarding the reference levels, the R software, by default, considers the first level of each variable as the reference level for the regression, thus these were manually changed. Table 4.1 provides information about the frequency and proportion of each level of the categorical variables. It also indicates the reference level chosen for each categorical variable.

Variable	Measurement Levels	Levels	Frequency	%
Sex	Categorical Dichotomous	Female	90	58.8
		Male	63	41.2
Age	Categorical Nominal	18 - 21	72	47.1
		22 - 25	60	39.2
		$\geq$ 26	21	13.7
FSHS	Categorical Nominal	Socioeconomic sciences	70	45.8
		Science and Technologies	64	41.8
		Languages, Humanities and Arts	19	12.4
CLE	Categorical Dichotomous	Masters and PhD	87	56.9
		Undergraduate and Postgraduate	66	43.1
CFS	Categorical Nominal	Social sciences, Business and Law	92	60.1
		Sciences, Engineering and Mathematics	49	32.0
		Humanities and Arts	5	3.3
		Health and Social protection	4	2.6
		Services	3	2.0

□ Reference Level

TABLE 4.1: Categorical Dependent Variables: Frequency and Percentage Distribution by Level

4.1.2 Sample Characterization

The survey sample consists of 153 students, 58.8% of these being female and 41.2% male. Students aged between 18 and 21 are predominant (47.1%), followed by students aged between 22 and 25 (39.2%). With regard to the Field of studies in High School (FSHS), there is a predominance of Socioeconomic sciences (45.8%), followed by Science and Technologies (41.8%). Of the respondents, 56.9% are Master's or PhD students and 43.1% are Undergraduate or Postgraduate students. With regard to the Current field of studies (CFS), Social sciences, Business, and Law are predominant (60.1%), followed by Sciences, Engineering, and Mathematics (32.0%). Further demographic information is available in the Table 4.1.

4.2 Regressions Analysis

Table 4.2 summarizes the results of the first binary logistic regression analysis for determinants of response to the first data visualization literacy (DVL) question in the survey.

The output suggests that none of the explanatory variables are statistically significant, as all p-values associated with the explanatory variables are greater than the conventional significance levels. Therefore, it cannot be concluded that there is a statistically significant association between Sex, Age, Field of studies in High School (FSHS), Current level of education (CLE), Current field of studies (CFS) and the response to the first DVL question of the survey.

Regarding the overall model, the non-significance of the Likelihood Ratio (LR) test suggests that the full model does not represent a signficant improvement in fit over the null model, LR$\chi^2(10)$ = 7.560, p-value = 0.672. In turn, the signficance of the Hosmer and Lemeshow (HL) test indicates that the model is not suitable to fit the data, $\chi^2(8)$ = 19.5, p-value = 0.012.

Variable	$\hat{\beta}$	S.E	Wald	p-value	Odds Ratio	95% CI for Odds Ratio
Constant	-1.362	1.288	-1.057	0.290	0.256	(0.021, 3.199)
Sex(male)	0.239	0.453	0.528	0.597	1.271	(0.522, 3.090)
22 - 25 years	0.359	0.535	0.672	0.502	1.432	(0.502, 4.084)
$\geq$ 26 years	-0.906	0.904	-1.002	0.316	0.404	(0.069, 2.378)
Science and Technologies [a]	-0.296	0.717	-0.413	0.680	0.744	(0.182, 3.034)
Socioeconomic sciences [a]	-0.226	0.697	-0.325	0.745	0.797	(0.203, 3.125)
Masters and PhD	0.180	0.526	0.342	0.732	1.198	(0.427, 3.360)
Health and Social protection [b]	-15.096	1195.237	-0.013	0.990	0.000	(0.000, ∞)
Sciences, Engineering and Mathematics [b]	0.220	1.275	0.172	0.863	1.246	(0.102, 15.173)
Services [b]	0.301	1.762	0.171	0.864	1.352	(0.043, 42.689)
Social sciences, Business and Law [b]	-0.265	1.247	-0.212	0.832	0.768	(0.067, 8.838)

	χ^2	df	p-value
Likelihood Ratio test	7.560	10	0.672
Hosmer & Lemeshow test	19.5	8	0.012

$**$ p < 0.05, $*$ p < 0.1
[a] regarding the variable FSHS, [b] regarding the variable CFS

TABLE 4.2: First Binary Logistic Regression results

Table 4.3 summarizes the results of the second binary logistic regression analysis for determinants of response to the second DVL question in the survey.

The output suggests that currently frequenting "Sciences, Engineering and Mathematics" field is significantly, at a 10% level, associated with an approximately 8 times higher probability of answering the second question correctly compared to frequenting the "Humanities and Arts" field (OR = 7.840, 95% CI: 0.728, 84.493). Additionally, currently frequenting "Social sciences, Business and Law" fields is significantly, at a 5% level, associated with an approximately 11 times higher probability of answering the second question correctly compared to frequenting the "Humanities and Arts" field (OR = 10.886, 95% CI: 1.073, 110.443).

Regarding the overall model, the non-significance of the LR test indicates that the full model does not reflect a signficant improvement in fit over the null model, LR$\chi^2(10)$ = 13.29, p-value = 0.208. In turn, the non-signficance of the HL test suggests that the model is adequate to fit the data, $\chi^2(8)$ = 3.347, p-value = 0.911.

Variable	$\hat{\beta}$	S.E	Wald	p-value	Odds Ratio	95% CI for Odds Ratio
Constant	-1.316	1.249	-1.053	0.292	0.268	(0.023, 3.103)
Sex(male)	0.366	0.435	0.841	0.400	1.442	(0.614, 3.386)
22 - 25 years	0.641	0.509	1.258	0.208	1.898	(0.700, 5.151)
$\geq$ 26 years	-0.343	0.657	-0.522	0.602	0.710	(0.196, 2.572)
Science and Technologies [a]	0.195	0.634	0.307	0.759	1.215	(0.351, 4.208)
Socioeconomic sciences [a]	0.159	0.621	0.255	0.798	1.172	(0.347, 3.961)
Masters and PhD	-0.545	0.497	-1.095	0.273	0.580	(0.219, 1.537)
Health and Social protection [b]	0.800	1.628	0.492	0.623	2.226	(0.092, 54.089
Sciences, Engineering and Mathematics [b]	2.059	1.213	1.698	0.090*	7.840	(0.728, 84.493)
Services [b]	1.796	1.724	1.042	0.298	6.024	(0.205, 176.654)
Social sciences, Business and Law [b]	2.387	1.182	2.020	0.043**	10.886	(1.073, 110.443)

	χ^2	df	p-value
Likelihood Ratio test	13.29	10	0.208
Hosmer & Lemeshow test	3.347	8	0.911

$**$ p < 0.05, $*$ p < 0.1

[a] regarding the variable FSHS, [b] regarding the variable CFS

TABLE 4.3: Second Binary Logistic Regression results

Table 4.4 summarizes the results of the third binary logistic regression analysis for determinants of response to the third DVL question in the survey.

The output suggests that none of the explanatory variables are statistically significant, as all p-values associated with the explanatory variables are greater than the conventional significance levels. Therefore, it cannot be concluded that there is a statistically significant association between Sex, Age, Field of studies in High School (FSHS), Current level of education (CLE), Current field of studies (CFS) and the response to the third DVL question of the survey.

Regarding the overall model, the non-significance of the LR test suggests that the full model does not represent a signficant improvement in fit over the null model, LRχ^2(10) = 16.846, p-value = 0.078. In turn, the non-signficance of the HL test indicates that the model is suitable to fit the data, χ^2(8) = 7.449, p-value = 0.489.

Variable	$\hat{\beta}$	S.E	Wald	p-value	Odds Ratio	95% CI for Odds Ratio
Constant	-1.414	1.226	-1.154	0.249	0.243	(0.022, 2.687)
Sex(male)	0.093	0.378	0.245	0.806	1.097	(0.523, 2.300)
22 - 25 years	0.483	0.455	1.061	0.289	1.620	(0.664, 3.951)
$\geq$ 26 years	-0.410	0.626	-0.655	0.513	0.664	(0.194, 2.265)
Science and Technologies [a]	0.209	0.583	0.359	0.720	1.233	(0.393, 3.864)
Socioeconomic sciences [a]	0.666	0.574	1.160	0.246	1.946	(0.632, 5.991)
Masters and PhD	-0.348	0.434	-0.802	0.423	0.706	(0.302, 1.653)
Health and Social protection [b]	-15.610	1186.900	-0.013	0.990	0.000	(0.000, ∞)
Sciences, Engineering and Mathematics [b]	1.106	1.187	0.931	0.352	3.021	(0.295, 30.936)
Services [b]	0.282	1.712	0.165	0.869	1.326	(0.046, 37.996)
Social sciences, Business and Law [b]	1.482	1.158	1.280	0.200	4.402	(0.455, 42.568)

	χ^2	df	p-value
Likelihood Ratio test	16.846	10	0.078
Hosmer & Lemeshow test	7.449	8	0.489

$**$ $p < 0.05$, $*$ $p < 0.1$
[a] regarding the variable FSHS, [b] regarding the variable CFS

TABLE 4.4: Third Binary Logistic Regression results

Table 4.5 summarizes the results of the fourth binary logistic regression analysis for determinants of response to the fourth DVL question in the survey.

The output suggests that being a male is significantly, at a 5% level, associated with an approximately 3 times higher probability of answering the fourth DVL question correctly compared to being a female (OR = 3.030, 95% CI: 1.285, 7.145). It also suggests that students who have pursued "Science and Technologies" in high school are significantly, at a 5% level, less likely to answer the fourth DVL question correctly compared to students who have pursued "Languages, Humanities and Arts" in high school (OR = 0.245, 95% CI: 0.063, 0.954). Additionally, students who have pursued "Socioeconomic sciences" in high school are significantly, at a 10% level, less likely to answer the fourth DVL question correctly compared to students who have pursued "Languages, Humanities and Arts" in high school (OR = 0.328, 95% CI: 0.089, 1.212).

Regarding the overall model, the non-significance of the LR test indicates that the full model does not reflect a signficant improvement in fit over the null model, LR$\chi^2(10)$ = 16.88, p-value = 0.077. In turn, the non-signficance of the HL test indicates that the model is adequate to fit the data, $\chi^2(8)$ = 2.503, p-value = 0.962.

Variable	$\hat{\beta}$	S.E	Wald	p-value	Odds Ratio	95% CI for Odds Ratio
Constant	-17.232	1611.552	–0.011	0.991	0.000	$(0.000, \infty)$
Sex(male)	1.108	0.438	2.532	0.011**	3.030	(1.285, 7.145)
22 - 25 years	-0.158	0.513	-0.308	0.758	0.854	(0.312, 2.334)
$\geq$ 26 years	-0.302	0.739	-0.408	0.683	0.740	(0.174, 3.146)
Science and Technologies [a]	-1.404	0.692	-2.028	0.043**	0.245	(0.063, 0.954)
Socioeconomic sciences [a]	-1.116	0.667	-1.671	0.095*	0.328	(0.089, 1.212)
Masters and PhD	-0.482	0.496	-0.972	0.331	0.617	(0.233, 1.633)
Health and Social protection [b]	17.616	1611.552	0.011	0.991	4.470×10^7	$(0.000, \infty)$
Sciences, Engineering and Mathematics [b]	17.501	1611.552	0.011	0.991	3.985×10^7	$(0.000, \infty)$
Services [b]	1.031	2723.707	0.000	1.000	2.803	$(0.000, \infty)$
Social sciences, Business and Law [b]	17.016	1611.552	0.011	0.992	2.455×10^7	$(0.000, \infty)$

	χ^2	df	p-value
Likelihood Ratio test	16.88	10	0.077
Hosmer & Lemeshow test	2.503	8	0.962

$**$ p < 0.05, $*$ p < 0.1
[a] regarding the variable FSHS, [b] regarding the variable CFS

TABLE 4.5: Fourth Binary Logistic Regression results

Table 4.6 summarizes the results of the fifth binary logistic regression analysis for determinants of response to the fifth DVL question in the survey.

The output suggests that students aged 26 and over are significantly, at a 5% level, less likely to answer the fifth DVL question correctly compared to students aged 18 to 21 (OR = 0.246, 95% CI: 0.065, 0.934).

Regarding the overall model, the non-significance of the LR test indicates that the full model does not reflect a signficant improvement in fit over the null model, LR$\chi^2(10)$ = 13.244, p-value = 0.210. In turn, the signficance of the HL test suggests that the model is not suitable to fit the data, $\chi^2(8)$ = 18.154, p-value = 0.020.

Variable	$\hat{\beta}$	S.E	Wald	p-value	Odds Ratio	95% CI for Odds Ratio
Constant	-0.176	1.088	-0.162	0.871	0.838	(0.099, 7.074)
Sex(male)	0.471	0.423	1.111	0.266	1.601	(0.698, 3.672)
22 - 25 years	-0.754	0.506	-1.490	0.136	0.470	(0.174, 1.269)
$\geq$ 26 years	-1.401	0.680	-2.061	0.039**	0.246	(0.065, 0.934)
Science and Technologies [a]	0.867	0.629	1.379	0.168	2.381	(0.694, 8.170)
Socioeconomic sciences [a]	0.651	0.599	1.087	0.277	1.917	(0.593, 6.205)
Masters and PhD	0.742	0.502	1.478	0.139	2.101	(0.785, 5.624)
Health and Social protection [b]	-0.314	1.490	-0.211	0.833	0.730	(0.039, 13.551)
Sciences, Engineering and Mathematics [b]	0.012	1.021	0.012	0.991	1.012	(0.137, 7.489)
Services [b]	-1.035	1.611	-0.643	0.520	0.355	(0.015, 8.347)
Social sciences, Business and Law [b]	0.784	0.987	0.794	0.427	2.189	(0.316, 15.162)

	χ^2	df	p-value
Likelihood Ratio test	13.244	10	0.210
Hosmer & Lemeshow test	18.154	8	0.020

$**$ p < 0.05, $*$ p < 0.1
[a] regarding the variable FSHS, [b] regarding the variable CFS

TABLE 4.6: Fifth Binary Logistic Regression results

Table 4.7 summarizes the results of the sixth binary logistic regression analysis for determinants of response to the sixth DVL question in the survey.

The output suggests that students aged 22 to 25 are significantly, at a 5% level, less likely to answer the sixth DVL question correctly compared to students aged 18 to 21 (OR = 0.274, 95% CI: 0.097, 0.774).

Regarding the overall model, the non-significance of the LR test suggests that the full model does not represent a signficant improvement in fit over the null model, LR$\chi^2(10)$ = 16.755, p-value = 0.080. In turn, the non-signficance of the HL test suggests that the model is adequate to fit the data, $\chi^2(8)$ = 10.331, p-value = 0.243.

Variable	$\hat{\beta}$	S.E	Wald	p-value	Odds Ratio	95% CI for Odds Ratio
Constant	-16.899	1023.230	-0.017	0.987	0.000	$(0.000, \infty)$
Sex(male)	0.310	0.418	0.743	0.458	1.364	(0.601, 3.093)
22 - 25 years	-1.295	0.530	-2.443	0.015**	0.274	(0.097, 0.774)
$\geq$ 26 years	0.370	0.666	0.555	0.579	1.447	(0.393, 5.334)
Science and Technologies [a]	0.485	0.701	0.692	0.489	1.625	(0.411, 6.426)
Socioeconomic sciences [a]	0.331	0.699	0.474	0.636	1.393	(0.354, 5.478)
Masters and PhD	-0.072	0.480	-0.151	0.880	0.930	(0.363, 2.383)
Health and Social protection [b]	15.857	1023.231	0.015	0.988	7.703×10^6	$(0.000, \infty)$
Sciences, Engineering and Mathematics [b]	16.223	1023.230	0.016	0.987	1.111×10^7	$(0.000, \infty)$
Services [b]	17.294	1023.231	0.017	0.987	3.241×10^7	$(0.000, \infty)$
Social sciences, Business and Law [b]	15.896	1023.230	0.016	0.988	8.010×10^6	$(0.000, \infty)$

	χ^2	df	p-value
Likelihood Ratio test	16.755	10	0.080
Hosmer & Lemeshow test	10.331	8	0.243

$**$ p < 0.05, $*$ p < 0.1
[a] regarding the variable FSHS, [b] regarding the variable CFS

TABLE 4.7: Sixth Binary Logistic Regression results

4.3 Discussion of Results

This research aims to understand whether the explanatory variables Sex, Age, FSHS, CLE, CFS predict students' ability to read, interpret and draw conclusions from data visualizations.

As found by Bielinski and Davison (1998), Penner (2003) and Lowrie and Diezmann (2011), and looking at the results of logistic regression 4.5, there appear to be gender differences, favoring males, in students' ability to perform on more difficult mathematical tasks. However, since gender did not appear to be a statistically significant variable for any other question, it was not possible to draw conclusions about its influence on easier mathematical tasks.

In accordance with the ideas of Binali et al. (2022) and Börner et al. (2016), it can be concluded that age was indeed a significant factor in predicting students' response to some questions. Looking at the results of the logistics regressions 4.6 and 4.7, it is possible to observe that older students are less likely to correctly answer more complex questions. These results contradict what would be anticipated, since, as suggested by Shreiner

(2018), students' encounters with data visualizations are expected to become more frequent throughout their academic progress, which would allow for the accumulation of prior knowledge for later interactions.

With regard to education, it was intended to take a more in-depth look at educational backgrounds as predictors of students' data visualization literacy (DVL). Not only were the students' current education levels explored, but also the areas of study from high school to the present day. Contrary to the findings made by Binali et al. (2022), Lowrie and Diezmann (2007) and Wainer (1980) that there is an improvement in students' DVL when they move on to higher levels of education, the results obtained in this study indicate that Current level of education (CLE) did not reach significance for any DVL question. Therefore, it is concluded that attending a higher level of education did not lead to improved competence in the ability to read, interpret, and draw conclusions from data visualization.

Although there is a propensity to underestimate the mathematical abilities of students from fields such as arts and languages, it is important to statistically investigate this hypothesis. Looking at the results of logistic regression 4.5, we can see that students who attended Science and Technologies or Socioeconomic Sciences in high school are less likely to correctly answer questions that are substantially more difficult than students who attended Languages, Humanities, and Arts. This finding is important since, understanding the impact of the students' high school fields on their DVL, will allow for possible curriculum changes to standardize high school mathematics learning, more specifically topics related to the graphical representations of data.

The results of logistic regression 4.3, indicate that students currently attending Sciences, Engineering and Mathematics or Social sciences, Business and Law are more likely to correctly answer questions that are substantially easier than students who attended Humanities and Arts.

Therefore, the results obtained in this study provide evidence that educational backgrounds have an influence on students' abilities to answer certain data visualization questions.

5 CONCLUSION

In a period when the amount of data and the information extracted from it grows at an exponential rate, contact with data visualizations is inevitable. Examples of this are the adoption of data visualizations by companies, to simplify the display of information, or even the appearance of daily dashboards on television channels, freely accessible to the population, as happened due to the covid-19 pandemic. Thus, the need to understand and interpret data visualizations for informed decision-making both in the work environment and in everyday life is evident.

This study aimed to investigate the predictive capacity of students' sociodemographic characteristics on their ability to read, interpret, and draw conclusions from data visualizations. According to the quantitative analysis of students' data visualization literacy (DVL) performed, it can be concluded that Sex, Age, Field of studies in High School (FSHS), and Current field of studies (CFS) are important factors in students' ability to correctly answer the data visualization questions.

The results indicate that male students are more likely to correctly answer more difficult questions. In addition, older students are less likely to answer these questions correctly. Whereas Wainer (1980), Lowrie and Diezmann (2007) and Binali et al. (2022) limited the study of the impact of education on students' DVL to their level of education, this approach, by including past and current fields of study, provides new insight into the influence of educational background on students' DVL. This research has clearly illustrated the influence of FSHS and CFS on students' ability to answer data visualization questions correctly. Regarding easier questions, students currently attending fields with higher mathematical requirements have higher chances of answering them correctly, while on more difficult questions, students who studied languages and humanities or arts in high school are more likely to answer them correctly. But it also raises the question: What changes need to be implemented in education to ensure that all individuals have a sufficient level of DVL that enables them to access information, resulting in informed decision-making?

This research contributes to the recognition of the relevance of data visualization literacy (DVL), a skill that cuts across several areas. Furthermore, it contributes to the literature as two new factors, FSHS and CFS, were discovered to be influential in students' DVL. This research may also spark a debate about how data visualizations are currently taught in schools, and whether there is a need for educators to adapt their teaching to different groups of students, thus raising the aforementioned question.

On a corporate level, I firmly believe that the results of this research will enable the development of more customized training. This will benefit both the employees, who develop their DVL skills more effectively, and the company, which reduces the time and resources needed to increase its human capital.

However, this study and its results should be viewed in light of some limitations. First, the non-probability convenience sampling technique chosen does not allow for the generalization of the results, since the final sample obtained is not representative of the population. Secondly, only the quantitative research method was used, whereas a qualitative approach, such as student interviews, could result in a deeper understanding of the topic. The data collection method should also be considered. Including different types of visualizations and questions could yield different results. At last, data cleaning substantially reduced the number of useful observations obtained through the survey, thus the sample size might have affected the accuracy of the results.

The findings of this study call for further research. Taking into account that primary school is mandatory and universal in Portugal, thus the influence of the field of studies is controlled, for future research, I suggest targeting the research to this population sample. More concretely, the impact of including a lecture, in which several types of data visualizations are shown in an appealing way, on the students' DVL should be investigated. This would give the students the possibility to be on an equal footing when assessing their abilities.

I further suggest developing mixed research, that is, using both quantitative and qualitative methods, targeting high school students. Since the interviews with the students would allow a deeper understanding of the topic and even an insight into whether the knowledge acquired about specific data visualizations came from primary school, which is universal, or from subjects in their specific field of studies in high school.

References

Alper, B., Riche, N. H., Chevalier, F., Boy, J. and Sezgin, M. (2017), Visualization literacy at elementary school, In Proceedings of the 2017 CHI conference on human factors in computing systems, 5485-5497.

Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D. and Cochran, J. J. (2016), Statistics for business & economics, Cengage Learning.

Aparicio, M. and Costa, C. J. (2015), Data visualization, Communication design quarterly Journal 3(1), 7-11.

Bateman, S., Mandryk, R. L., Gutwin, C., Genest, A., McDine, D. and Brooks, C. (2010), Useful junk? The effects of visual embellishment on comprehension and memorability of charts, In Proceedings of the SIGCHI conference on human factors in computing systems 4, 2573-2582.

Bielinski, J. and Davison, M. L. (1998), Gender differences by item difficulty interactions in multiple-choice mathematics items, American Educational Research Journal 35(3), 455-476.

Binali, T., Chang, C. H., Chang, Y. J. and Chang, H. Y. (2022), High school and college students' graph-interpretation competence in scientific and daily contexts of data visualization, Journal of Science Education, 1-23.

Blakeman, J. R. (2020), Words matter: Sex and gender as unique variables in research, Journal of Advances in Nursing Science 43(3), 214-227.

Borkin, M. A., Vo, A. A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A. and Pfister, H. (2013), What makes a visualization memorable?, IEEE transactions on visualization and computer graphics Journal 19(2), 2306-2315.

Boy, J., Rensink, R. A., Bertini, E. and Fekete, J. D. (2014), A principled way of assessing visualization literacy, IEEE transactions on visualization and computer graphics Journal 20(12), 1963-1972.

BSA. (2015), What's the big deal with data?, Retrieved from `https://www.bsa.org/files/reports/bsadatastuy_en.pdf`

Börner, K., Maltese, A., Balliet, R. N. and Heimlich, J. (2016), Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors, Information Visualization Journal 15(3), 198-213.

Card, S. K. (1999), Readings in information visualization: using vision to think, Morgan Kayfmann.

Chen, C., Härdle, W. and Unwin, A. (2008), Handbook of data visualization, Springer.

Coleman, J. (2010), Elementary teachers' instructional practices involving graphical representations, Journal of Visual Literacy 29(2), 198-222.

Crato, N. (2006), O'eduquês' em discurso directo: uma crítica da pedagogia romântica e construtivista, Gradiva Lisboa.

Curcio, F. R. (1987), Comprehension of mathematical relationships expressed in graphs, Journal for research in mathematics education 18(5), 382-393.

Educação, R. P. (2018a), Aprendizagens essenciais - matemática - 2º ano - 1º ciclo, Retrieved from `http://www.dge.mec.pt/sites/default/files/Curriculo/Aprendizagens_Essenciais/1_ciclo/matematica_1c_2a_ff_18julho_rev.pdf`.

Educação, R. P. (2018b), Aprendizagens essenciais - matemática - 5º ano - 1º ciclo, Retrieved from `http://www.dge.mec.pt/sites/default/files/Curriculo/Aprendizagens_Essenciais/2_ciclo/5_matematica_18julho_rev.pdf`.

Educação, R. P. (2018c), Aprendizagens essenciais - matemática - 6º ano - 1º ciclo, Retrieved from `http://www.dge.mec.pt/sites/default/files/Curriculo/Aprendizagens_Essenciais/2_ciclo/6_matematica_18julho_rev.pdf`.

Educação, R. P. (2018d), Aprendizagens essenciais - matemática - 7º ano - 1º ciclo, Retrieved from `http://www.dge.mec.pt/sites/default/files/Curriculo/Aprendizagens_Essenciais/3_ciclo/matematica_3c_7a_ff_18julho_rev.pdf`.

Educação, R. P. (2018e), Aprendizagens essenciais - matemática - 8º ano - 1º ciclo, Retrieved from `http://www.dge.mec.pt/sites/default/files/Curriculo/Aprendizagens_Essenciais/3_ciclo/matematica_3c_8a_ff_18julho_rev.pdf`.

Elias, M. and Bezerianos, A. (2011), Exploration views: understanding dashboard creation and customization for visualization novices, In In 13th International Conference on Human-Computer Interaction 6949(4), 274-291.

Few, S. and Edge, P. (2007), Data visualization: past, present, and future, IBM Cognos Innovation Center, 1-12.

Fox, J. and Monette, G. (1992), Generalized collinearity diagnostics, Journal of the American Statistical Association 87(417), 178-183.

Freedman, E. G. and Shah, P. (2002), Toward a model of knowledge-based graph comprehension, In Diagrammatic representation and inference, 18-30.

Friel, S. N., Curcio, F. R. and Bright, G. W. (2001), Making sense of graphs: Critical factors influencing comprehension and instructional implications, Journal for Research in Mathematics Education 32(2), 124-158.

Friendly, M. (2002), Visions and re-visions of charles joseph minard, Journal of Educational and Behavioral Statistics 27(1), 31-51.

Friendly, M. (2008a), The golden age of statistical graphics. Journal of Statistical Science 23(4), 502-535.

Friendly, M. (2008b), Milestones in the history of thematic cartography, statistical graphics, and data visualization, 1-79.

Friendly, M. and Denis, D. (2000), Discussion and comments. Approche graphique en analyse des données. The roots and branches of modern statistical graphics, Journal de la société française de statistique 141(4), 51-60.

Geckoboard (n.d.), Sales Product Performance Dashboard Example, Retrieved from https://www.geckoboard.com/dashboard-examples/sales/sales-product-performance-dashboard/.

Hariri, R. H., Fredericks, E. M. and Bowers, K. M. (2019), Uncertainty in big data analytics: survey, opportunities, and challenges, Journal of Big Data 6(1), 1-16.

Hoaglin, D. C., Mosteller, F. and Tukey, J. W. (1985), Exploring data tables, trends, and shapes, John Wiley Sons.

Keim, D., Qu, H. and Ma, K. L. (2013), Big-data visualization, IEEE Computer Graphics and Applications Journal 33(4), 20-21.

Kennedy, H. and Hill, R. L. (2018), The feeling of numbers: Emotions in everyday engagements with data and their visualisation, Journal of Sociology 52(4), 830-848.

Kleinbaum, D. G. and Klein, M. (2010), Logistic regression: A self-learning text, Springer.

Lee, B., Choe, E. K., Isenberg, P., Marriott, K. and Stasko, J. (2020), Reaching broader audiences with data visualization, IEEE Computer Graphics and Applications Journal 40(2), 82-90.

Lee, S., Kim, S. H., Hung, Y. H., Lam, H., Kang, Y. A. and Yi, J. S. (2015), How do people make sense of unfamiliar visualizations?: A grounded model of novice's information visualization sensemaking, IEEE transactions on visualization and computer graphics Journal 22(1), 499-508.

Lee, S., Kim, S. H. and Kwon, B. C. (2017), Vlat: Development of a visualization literacy assessment test, IEEE Transactions on visualization and computer graphics Journal 23(1), 551-560.

Linn, M. C. and Petersen, A. C. (1985), Emergence and characterization of sex differences in spatial ability: A meta-analysis, Child development Journal 56(6), 479–1498.

Lowrie, T. and Diezmann, C. M. (2007), Solving graphics problems: Student performance in junior grades, Journal of Educational Research 100(6), 369-378.

Lowrie, T. and Diezmann, C. M. (2011), Solving graphics tasks: Gender differences in middle-school students, Learning and Instruction Journal 21(1), 109-125.

Ludewig, U., Lambert, K., Dackermann, T., Scheiter, K. and Möller, K. (2020), Influences of basic numerical abilities on graph reading performance, Journal of Psychological Research 84(5), 1198-1210.

Maltese, A., Svetina, D. and Harsh, J (2015), Data visualization literacy: Investigating data interpretation along the novice-expert continuum, Journal of College Science Teaching 45(1), 84-90.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A. H. (2011), Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, 1-156.

McTigue, E. M. and Flowers, A. C. (2011), Science visual literacy: Learners' perceptions and knowledge of diagrams, The Reading Teacher Journal 64(8), 578-589.

Menard, S. (2002), Applied logistic regression analysis 106, Sage.

Menard, S. (2009), Logistic regression: From introductory to advanced concepts and applications, Sage.

Nelson, B. and Jones, H. (2023), Logic Puzzle : How many squares, Reader's Digest magazine, Retrieved from https://www.rd.com/article/how-many-squares-image/.

Peck, E. M., Ayuso, S. E. and El-Etr, O. (2019), Data is personal: Attitudes and perceptions of data visualization in rural pennsylvania, In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1-12.

Penner, A. M. (2003), International gender x item difficulty interactions in mathematics and science achievement tests, Journal of Educational Psychology 95(3), 650-655.

Peppler, K., Keune, A. and Han, A. (2021), Cultivating data visualization literacy in museums, Journal of Information and Learning Sciences 122(1-2), 1-16.

Pilania, G. (2021), Machine learning in materials science: From explainable predictions to autonomous design, Journal of Computational Materials Science 193, 1-13.

Playfair, W. (2005), The commercial and political atlas and statistical breviary, Cambridge University Press.

Ruchikachorn, P. and Mueller, K. (2015), Learning visualizations by analogy: Promoting visual literacy through visualization morphing. IEEE transactions on visualization and computer graphics Journal 21(9), 1028-1044.

Rydning, J., Reinsel, D. and Gantz, J. (2018), The digitization of the world from edge to core, Framingham: International Data Corporation 16, 1-28.

Sandbag (n.d.), Power generation by source (2000-2018), Retrieved from https://ember-climate.org/project/ets-emissions-2018/.

Saunders, M. N., Lewis, P. and Thornhill, A. (2019), Research methods for business students, Pearson education.

Shah, P. and Hoeffner, J. (2002), Review of graph comprehension research: Implications for instruction, Journal of Educational psychology review 14(1), 47-69.

Shah, P., Mayer, R. E. and Hegarty, M. (1999), Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension, Journal of educational psychology 91(4), 690-702.

Shreiner, T. L. (2018), Data literacy for social studies: Examining the role of data visualizations in k–12 textbooks. Theory Research in Social Education Journal 46(2), 194-231.

Stevens, J. (2013), Squatch Watch: 92 Years of Bigfoot Sightings in the US and Canada, Retrieved from `https://www.joshuastevens.net/visualization/squatch-watch-92-years-of-bigfoot-sightings-in-us-and-canada/`.

Tufte, E. R. (1983), The visual display of information, Graphics Press.

Tukey, J. W. (1962), The future of data analysis, The annals of mathematical statistics 33(1), 1-67.

UNESCO. (n.d.), Literacy, Retrieved from `http://uis.unesco.org/node/3079547`.

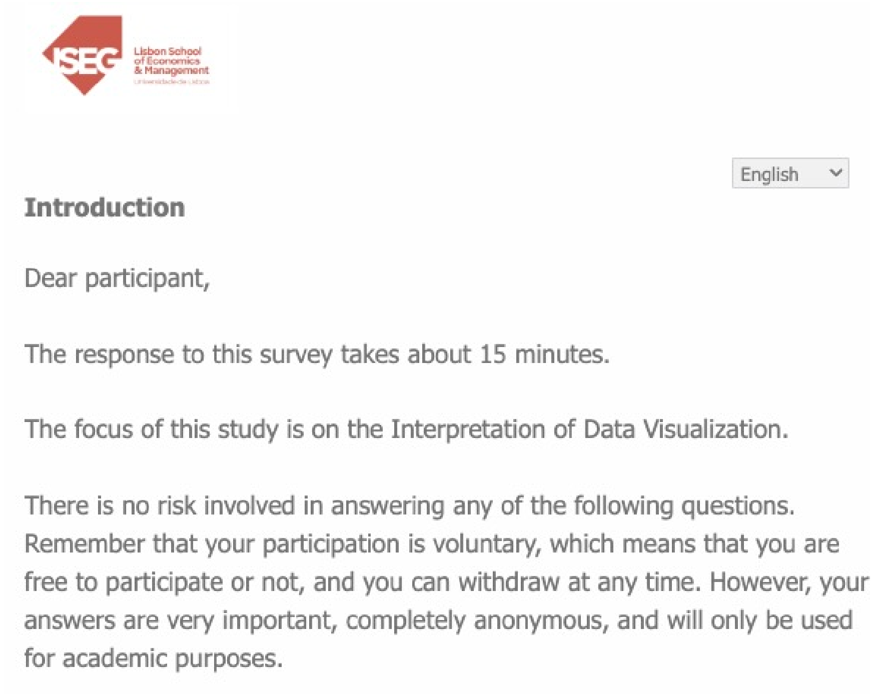
Wainer, H. (1980), A test of graphicacy in children, Applied Psychological Measurement Journal 4(3), 331-340.

Wellbeing@School (n.d.), Understanding and interpreting box plots, Retrieved from `https://www.wellbeingatschool.org.nz/information-sheet/understanding-and-interpreting-box-plots`.

Womack, R. (2014), Data visualization and information literacy, IAssist Quarterly Journal 38(1), 12-17.

A APPENDICES

A.1 Survey



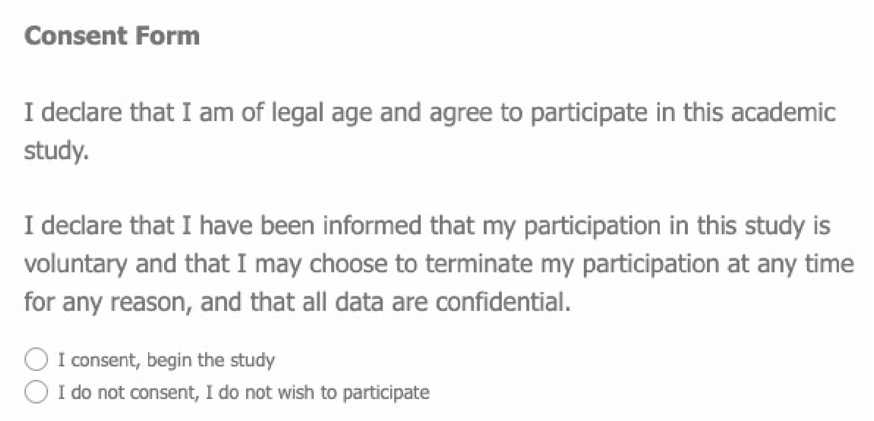
Introduction

Dear participant,

The response to this survey takes about 15 minutes.

The focus of this study is on the Interpretation of Data Visualization.

There is no risk involved in answering any of the following questions. Remember that your participation is voluntary, which means that you are free to participate or not, and you can withdraw at any time. However, your answers are very important, completely anonymous, and will only be used for academic purposes.



Consent Form

I declare that I am of legal age and agree to participate in this academic study.

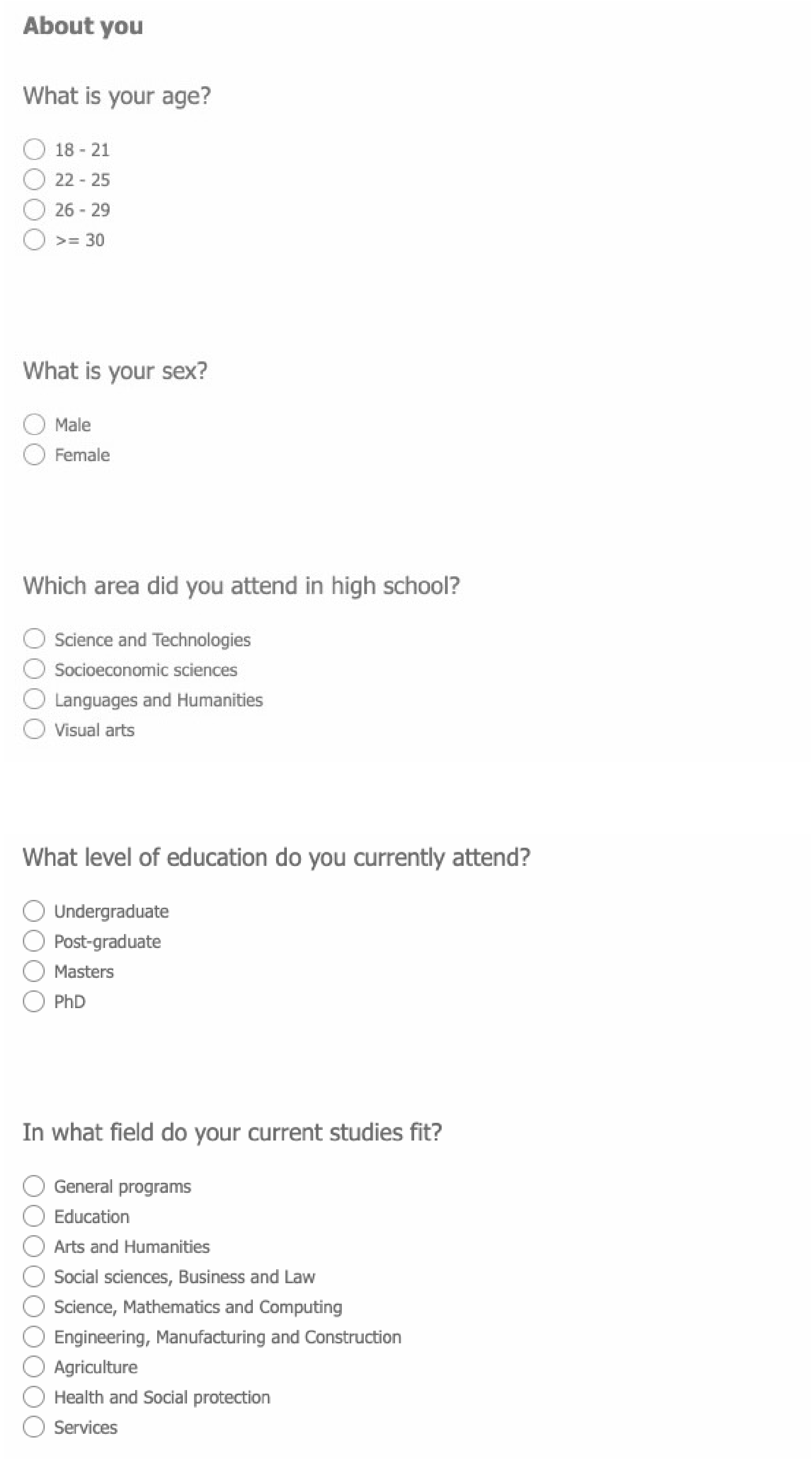
I declare that I have been informed that my participation in this study is voluntary and that I may choose to terminate my participation at any time for any reason, and that all data are confidential.

○ I consent, begin the study
○ I do not consent, I do not wish to participate

Student

Are you a student?

○ Yes
○ No



About you

What is your age?

○ 18 - 21
○ 22 - 25
○ 26 - 29
○ >= 30

What is your sex?

○ Male
○ Female

Which area did you attend in high school?

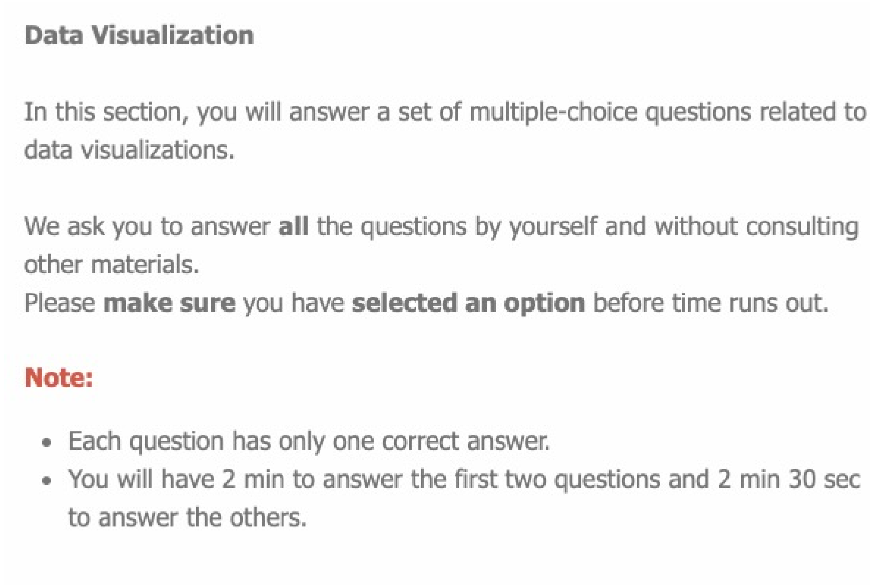
○ Science and Technologies
○ Socioeconomic sciences
○ Languages and Humanities
○ Visual arts

What level of education do you currently attend?

○ Undergraduate
○ Post-graduate
○ Masters
○ PhD

In what field do your current studies fit?

○ General programs
○ Education
○ Arts and Humanities
○ Social sciences, Business and Law
○ Science, Mathematics and Computing
○ Engineering, Manufacturing and Construction
○ Agriculture
○ Health and Social protection
○ Services



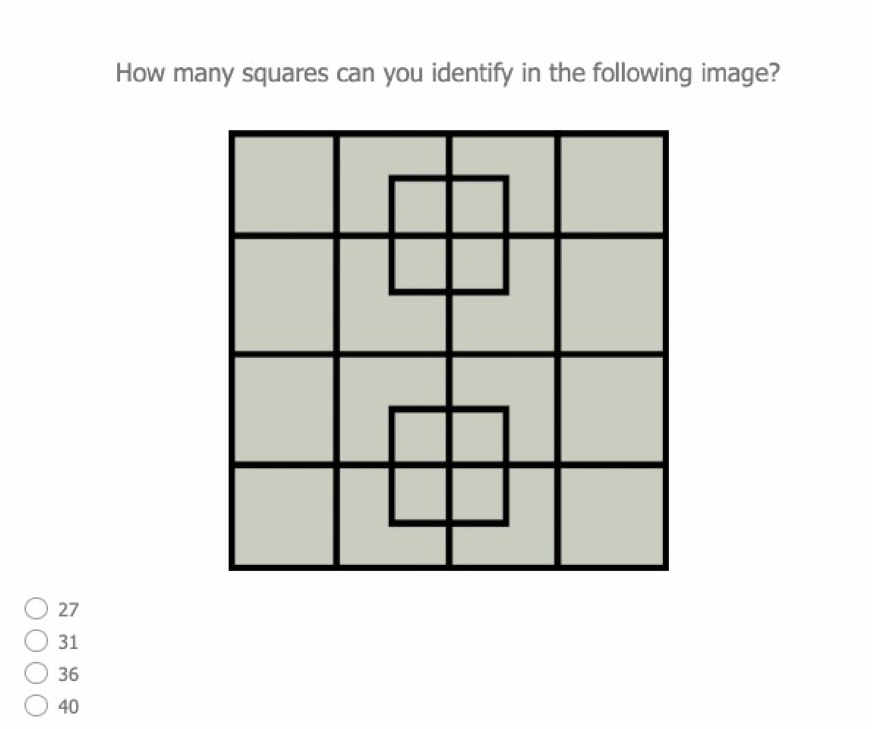
Data Visualization

In this section, you will answer a set of multiple-choice questions related to data visualizations.

We ask you to answer **all** the questions by yourself and without consulting other materials.
Please **make sure** you have **selected an option** before time runs out.

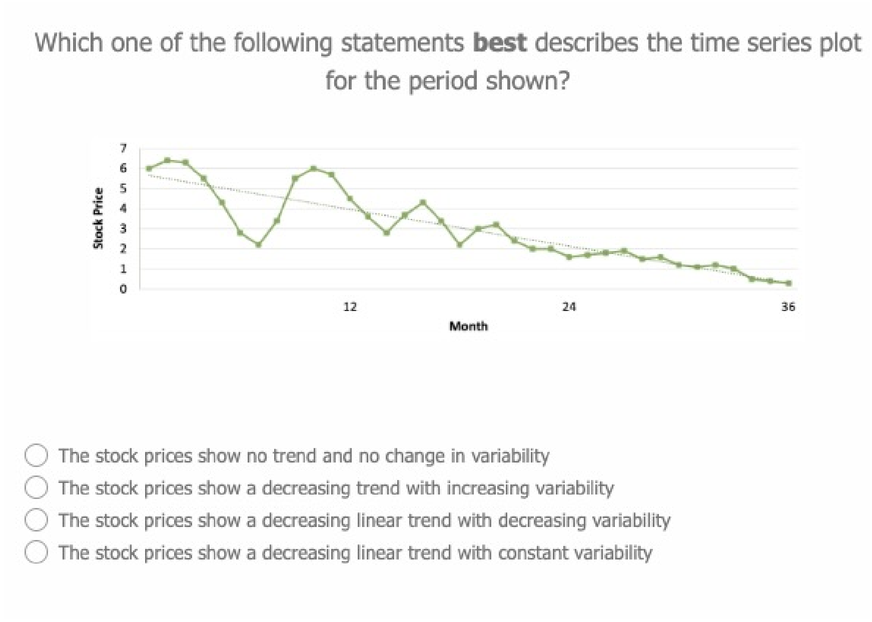
Note:

- Each question has only one correct answer.
- You will have 2 min to answer the first two questions and 2 min 30 sec to answer the others.



How many squares can you identify in the following image?

○ 27
○ 31
○ 36
○ 40



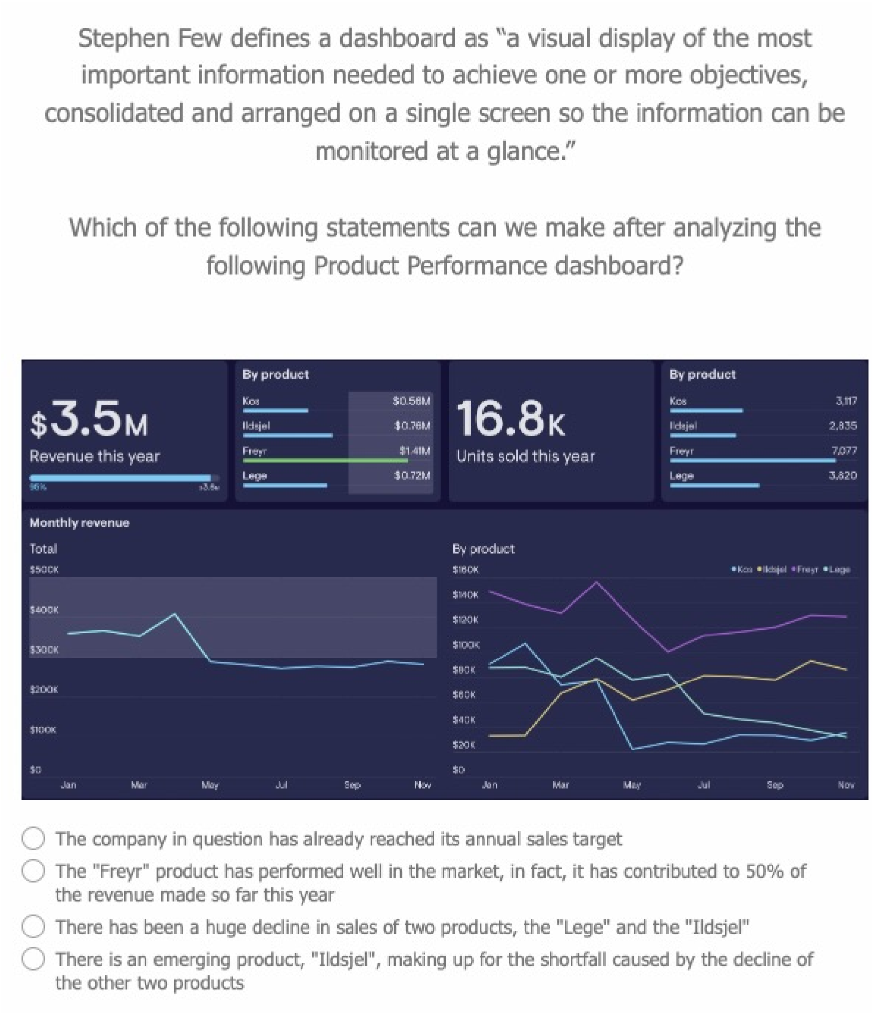
Which one of the following statements **best** describes the time series plot for the period shown?

○ The stock prices show no trend and no change in variability

○ The stock prices show a decreasing trend with increasing variability

○ The stock prices show a decreasing linear trend with decreasing variability

○ The stock prices show a decreasing linear trend with constant variability



Stephen Few defines a dashboard as "a visual display of the most important information needed to achieve one or more objectives, consolidated and arranged on a single screen so the information can be monitored at a glance."

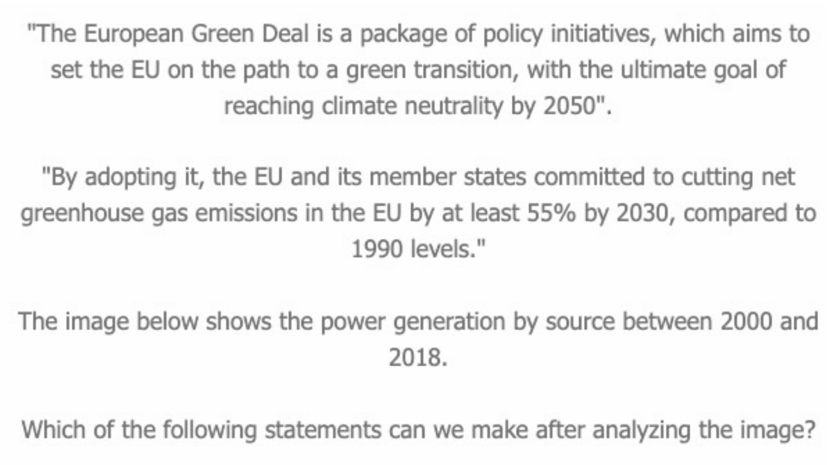
Which of the following statements can we make after analyzing the following Product Performance dashboard?

○ The company in question has already reached its annual sales target

○ The "Freyr" product has performed well in the market, in fact, it has contributed to 50% of the revenue made so far this year

○ There has been a huge decline in sales of two products, the "Lege" and the "Ildsjel"

○ There is an emerging product, "Ildsjel", making up for the shortfall caused by the decline of the other two products

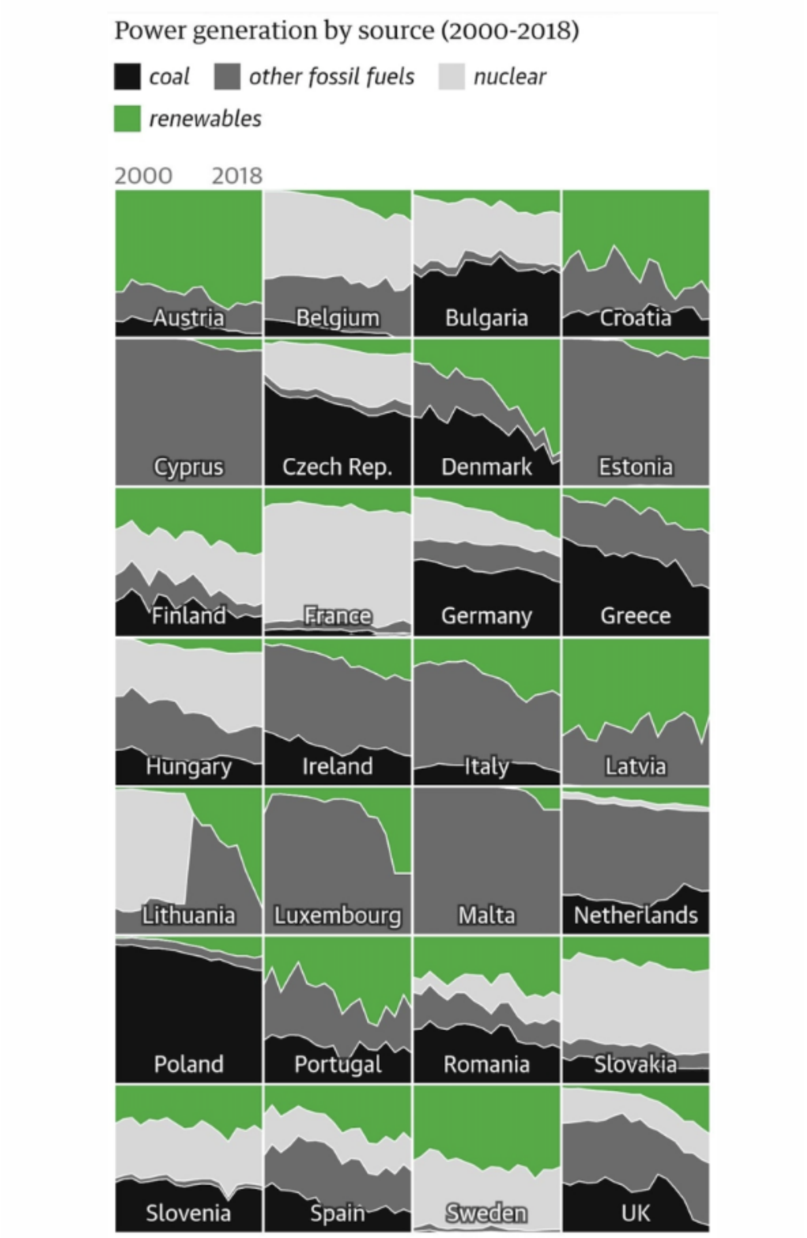


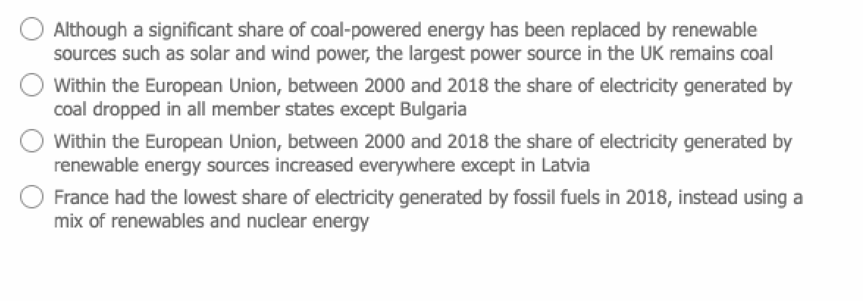
"The European Green Deal is a package of policy initiatives, which aims to set the EU on the path to a green transition, with the ultimate goal of reaching climate neutrality by 2050".

"By adopting it, the EU and its member states committed to cutting net greenhouse gas emissions in the EU by at least 55% by 2030, compared to 1990 levels."

The image below shows the power generation by source between 2000 and 2018.

Which of the following statements can we make after analyzing the image?



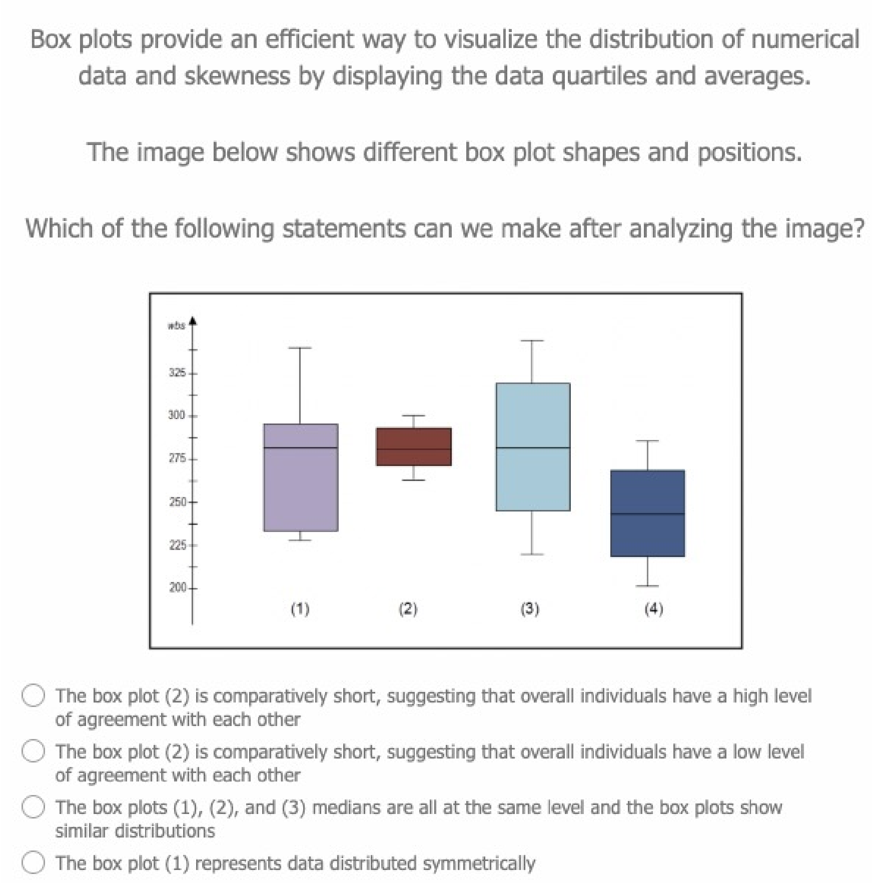


○ Although a significant share of coal-powered energy has been replaced by renewable sources such as solar and wind power, the largest power source in the UK remains coal

○ Within the European Union, between 2000 and 2018 the share of electricity generated by coal dropped in all member states except Bulgaria

○ Within the European Union, between 2000 and 2018 the share of electricity generated by renewable energy sources increased everywhere except in Latvia

○ France had the lowest share of electricity generated by fossil fuels in 2018, instead using a mix of renewables and nuclear energy



Box plots provide an efficient way to visualize the distribution of numerical data and skewness by displaying the data quartiles and averages.

The image below shows different box plot shapes and positions.

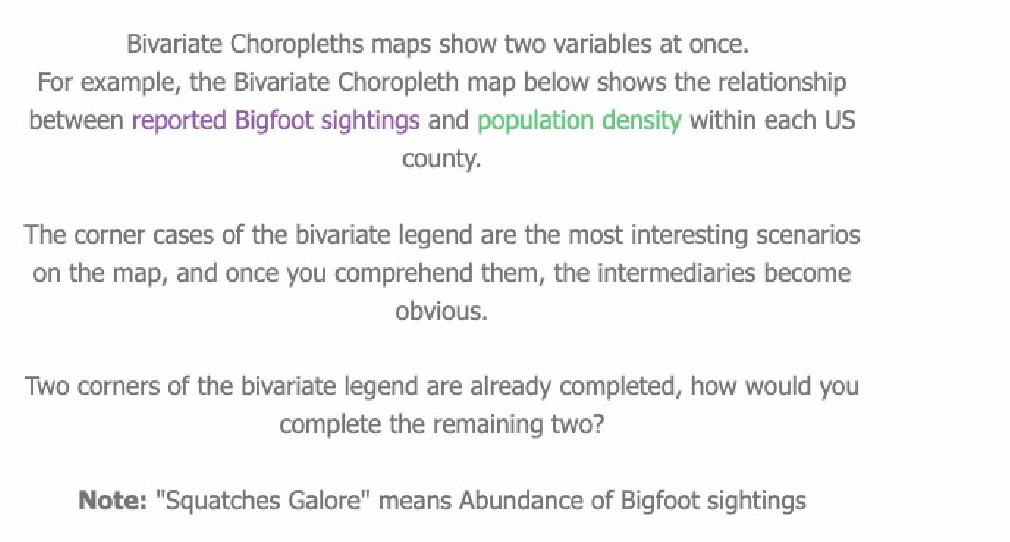
Which of the following statements can we make after analyzing the image?

○ The box plot (2) is comparatively short, suggesting that overall individuals have a high level of agreement with each other

○ The box plot (2) is comparatively short, suggesting that overall individuals have a low level of agreement with each other

○ The box plots (1), (2), and (3) medians are all at the same level and the box plots show similar distributions

○ The box plot (1) represents data distributed symmetrically

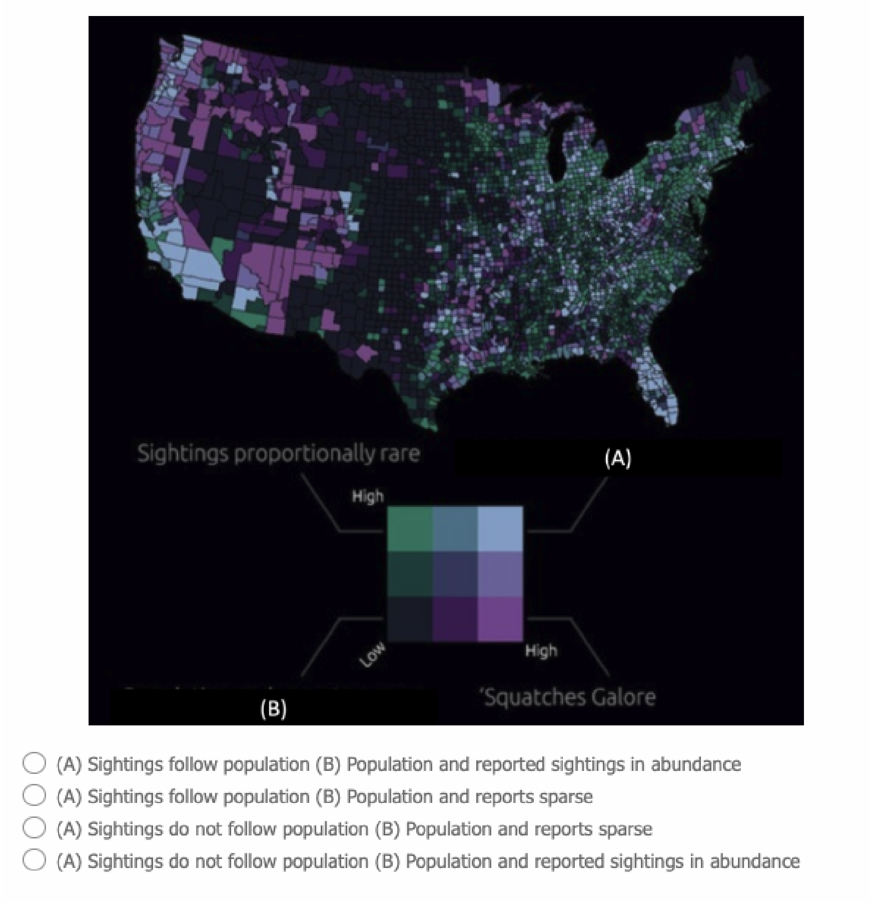


Bivariate Choropleths maps show two variables at once.
For example, the Bivariate Choropleth map below shows the relationship between reported Bigfoot sightings and population density within each US county.

The corner cases of the bivariate legend are the most interesting scenarios on the map, and once you comprehend them, the intermediaries become obvious.

Two corners of the bivariate legend are already completed, how would you complete the remaining two?

Note: "Squatches Galore" means Abundance of Bigfoot sightings



○ (A) Sightings follow population (B) Population and reported sightings in abundance
○ (A) Sightings follow population (B) Population and reports sparse
○ (A) Sightings do not follow population (B) Population and reports sparse
○ (A) Sightings do not follow population (B) Population and reported sightings in abundance