Exploratory analysis of social and economic factors on the National Basic Education Assessment System in Brazil

Erik Nonato Rolin

Bruno Cavalli

Nicholas Farrel Ferraz de Faria

Getúlio Vargas Foundation (FGV)

1. ABSTRACT

This study conducts an exploratory analysis using the 2021 dataset from the National Basic Education Assessment System (SAEB), aiming to investigate correlations between educational variables and socioeconomic factors. Four main hypotheses were analyzed: (1) whether parental encouragement for students to study influences school failure and dropout rates; (2) whether the number of enrolled students and the participation rate during exams affect the school's score on the SAEB; (3) whether the school's location impacts its socioeconomic level and academic performance; and (4) whether students who studied in private schools demonstrate better performance compared to those in public schools. Through data analysis, the study seeks to identify patterns that can inform the development of more effective educational policies, addressing regional disparities and emphasizing the importance of family involvement in academic success. This paper does not aim to provide definitive proof of correlations between these variables; rather, its primary goal is to identify trends and encourage further research and discussion on these crucial topics at hand. To illustrate the correlations, this study will present graphs generated by the Seaborn library, including heatmaps and boxplots, accompanied by a detailed analysis to address the hypotheses. A link to the corresponding GitHub repository will be provided in the references for further exploration of the data and code used in this study.

2. INTRODUCTION

Education is a fundamental pillar for social and economic development, yet inequalities persist in access to quality education across different regions and socioeconomic groups. In Brazil, the National Basic Education Assessment System (SAEB) serves as a comprehensive tool to evaluate educational performance, providing valuable data via a national test given to all students that can shed light on the relationships between various factors influencing student outcomes. These factors include the role of family involvement, the number of enrolled students, geographical location, and the type of

school (public vs. private), all of which contribute to the broader discussion on how to improve educational quality and equity.

While previous studies have explored individual factors like socioeconomic status or regional disparities, there remains a need for a deeper understanding of how these variables interact and impact critical educational outcomes, such as academic performance, school dropout rates, and many others. The 2021 SAEB dataset offers a rich source of data to conduct such an investigation, allowing for the identification of trends and correlations that could provide insights into these challenges.
This paper takes an exploratory approach to examine four key hypotheses. First, (1) Does parental encouragement of a better academic life for children influence the odds of school dropout and/or failure? Second, (2) Does the size of enrolled students impact the school's overall score on the SAEB? Additionally, does the participation rate of the enrolled students influence the school's overall score on the SAEB? Third, (3) Does the location of the school influence its socioeconomic level and its overall score on the SAEB? Fourth, (4) Do students who only received public education have a disadvantage compared to those who had/have access to private education?
This study aims to explore these interconnected themes through an analysis of the 2021 SAEB dataset. By focusing on four key hypotheses, the research seeks to uncover patterns and correlations that can inform educational policy and practice. Rather than providing definitive conclusions, this study intends to stimulate discussion and further investigation into the dynamics at play in the Brazilian educational landscape.
The findings will be presented through visual representations, including heatmaps and boxplots generated using the Seaborn library, to facilitate a clear understanding of the data. A link to the associated GitHub repository will also be provided in the references, allowing for transparency and accessibility in the research process.
By shedding light on these crucial aspects of education, this paper contributes to the ongoing dialogue about improving educational equity in Brazil. It emphasizes the importance of both familial and structural factors in fostering an inclusive and effective educational system, ultimately aiming to guide policymakers, educators, and researchers in their efforts to enhance student outcomes.

3. DEVELOPMENT

3.1 Data cleaning and hypotheses

3.1.1 Understanding the datasets

Our study is based on two main datasets provided by the INEP (National Institute of Educational Studies and Research Anísio Teixeira) from the 2021 research of the SAEB, along with a dictionary for them containing what each variable references and the different types of return for the variable (qualitative or quantitative). The first dataset is "TS_ALUNO_34EM," composed of the data obtained from the answers of all the students participating in this year's SAEB. It encompasses a wide range of questions posed to the students and based on the participants' information, contains data such as the socioeconomic level of the students, their proficiency levels in either mathematics or Portuguese, their responses regarding parental involvement in their academic lives, their

physical information, and many more. The second dataset is “TS_ESCOLA,” which contains information about the schools. This information is collected either separately from the students' overall answers and scores, such as location information and socioeconomic level, or represents its students, with information like the total number of enrolled students, the ratio of participants during the test and the average performance of the school on the SAEB.

3.1.2 Creating the Hypotheses

The main goal of the hypotheses is to cover a satisfactory range of questions that can be derived from both datasets and are valid points in the ongoing discussion. After obtaining a deeper understanding of the different types of variables being analyzed, four main topics of discussion were identified for further analysis:

With the first hypothesis (1), we wanted to observe the correlation between 3 variables, “TX_RESP_Q09c” (corresponds to the student’s answer to “How often do your parents or guardians usually: - Encourage you to study.” Returning either A, B, C, ., *) “TX_RESP_Q18” (corresponds to the student’s answer to “Have you already repeated a year?” Returning either A, B, C, ., *) and “TX_RESP_Q19” (corresponds to the student’s answer to “Have you ever dropped out of school and stopped attending until the end of the school year?” Returning either A, B, C, ., *). The goal with this hypothesis is to show a correlation between a student either failing a year or dropping out all together with how supportive their parents or guardians have been throughout their academic career.

With the second hypothesis (2), we wanted to observe the correlation between 2 variables, “NU_MATRICULADOS_CENSO_EMT” (corresponds to the total number of students on a school) and the average of both “MEDIA_EM_MT” and “MEDIA_EM_LP” (being the average of the school on mathematics and Portuguese, respectively, into one average), additionally, we wanted to observe the correlation between “TAXA_PARTICIPACAO_EM” (being the ratio of present students during the test) and the average of both “MEDIA_EM_MT” and “MEDIA_EM_LP”. The goal of this hypothesis is to find a possible correlation between the number of students (either in total or present on the test) and the average of the school.

With the third hypothesis (3), we wanted to deal with a lot of variables when defining the school location, those variables were: “ID_REGIAO”, “ID_LOCALIZACAO”, those meaning, respectively, the Brazilian’s geographic region that the school is located and either if the school is located on the urban side or the countryside. Those variables would be compared to the average of their schools to determine the correlation between these variables. The goal with this hypothesis is to observe the disparity among all the possible locations a school can have and its average on the SAEB.

With the fourth and final hypothesis (4), we wanted to observe how two variables would interact, the first one was “TX_RESP_Q17” (which refers to the answers from the students to “from the first year of elementary school onwards, what type of school did you study at?”) and their averages on SAEB using both “MEDIA_EM_MT” and “MEDIA_EM_LP”. The goal of this hypothesis is to see if the student that only had

access to a public education will show a disadvantage compared to those who have or had access to a private education in their lives.

With all, these are the hypotheses:

(1) Does parental encouragement of a better academic life for children influence the odds of school dropout and/or failure?
(2) Does the size of enrolled students impact the school's overall score on the SAEB? Additionally, does the participation rate of the enrolled students influence the school's overall score on the SAEB?
(3) Does the location of the school influence its socioeconomic level and its overall score on the SAEB?
(4) Do students who only received public education have a disadvantage compared to those who had/have access to private education?

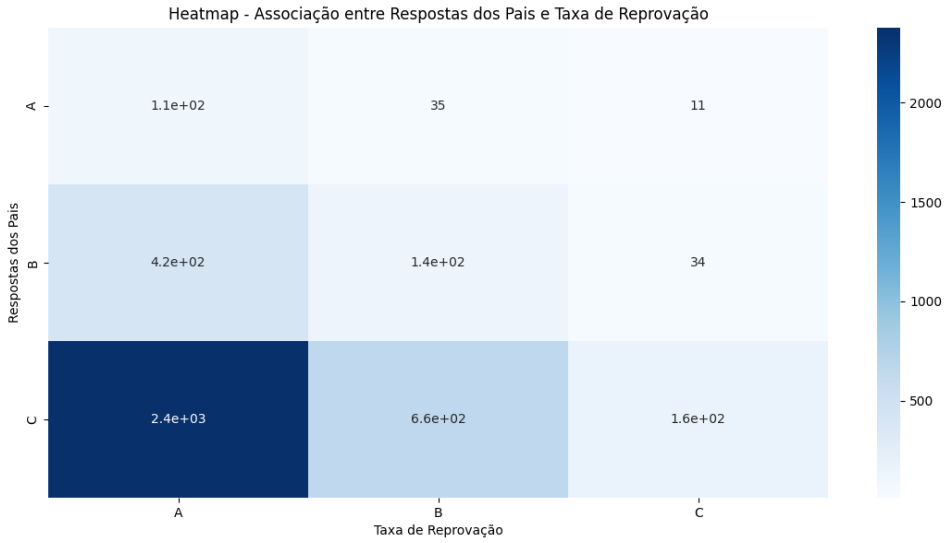
3.1.3 Cleaning the dataset

With the hypothesis at hand and with the variables of interest, the next step on this study was to clean the dataset of any missing or corrupted data and reallocate what was valuable to this research to a cleaner and more simple dataset to start creating the analysis, for this step, we primarily used the pandas library, this tool allowed us to simplify the variables names for clarity and separate them from the rest of the dataset using "pd.Dataframe", after that, using the merge function, we were able to create a new dataset with much less storage and the crucial points for the continuation of this study called "data_saeb.csv"

3.2 Creating the analysis

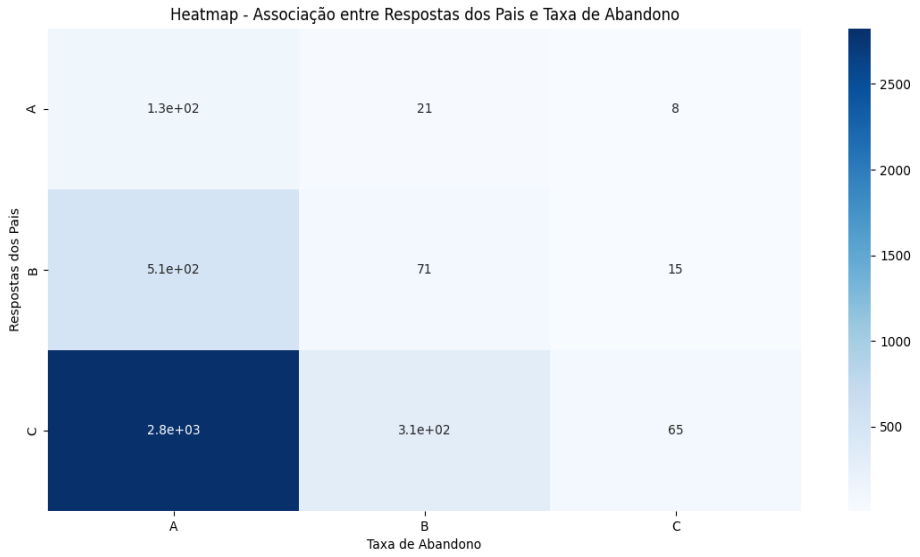
3.2.1 First hypotheses

With the new dataset at disposal, to answer the first hypotheses, it was necessary to generate a heatmap that would correlate two qualitative variables and calculate the x^2(chi2) to find the correlation, this is the result:



Understanding the graph: on the X axis, we have 3 options that refer to the question of “Have you already repeated a year?”, “A” meaning “never”; “B” meaning “just once”; “C” meaning "more than once”, while on the y axis, we also have 3 options that refer to the question of “How often do your parents or guardians usually: - Encourage you to study.”, “A” meaning “Never or almost never”; “B” meaning “sometimes”; “C” meaning "always or almost always”. The number on each box has their decimals cut, but to see the real number on each box, just look at the side bar that, with its color, indicates the scale of the number.

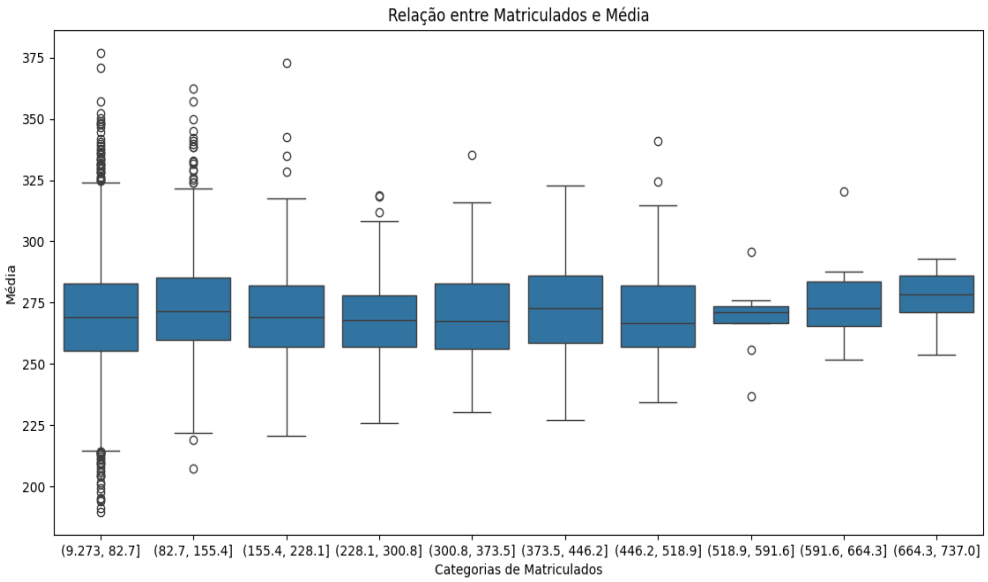
Data analysis: By only looking at the heatmap, it becomes unclear whether the correlation is truly linear, while its clear that most kids answered that they never failed a year and their parents/guardians are always encouraging them to study, the graph also shows a good portion of them saying that they failed more than once and yet their parents are supportive, while the kids that failed more than once and their parents/guardians are not supportive are few and far between. This situation also shows its confusion on the x^2, been a 0.4117928022092565, for context, a 1.0 means maximum correlation and a 0.0 means no correlation. The same conclusion can be drawn when seeing the dropout rate.



Conclusion: Does parental encouragement of a better academic life for children influence the odds of school dropout and/or failure? Yes, but this variable alone does not tell the whole story, while we see a lot of kids never failing a year and having encouraging parents, when we look at the kids that failed or dropped for more than one year and don’t have encouraging parents, we see few and far between, so there must be more than just the parents.

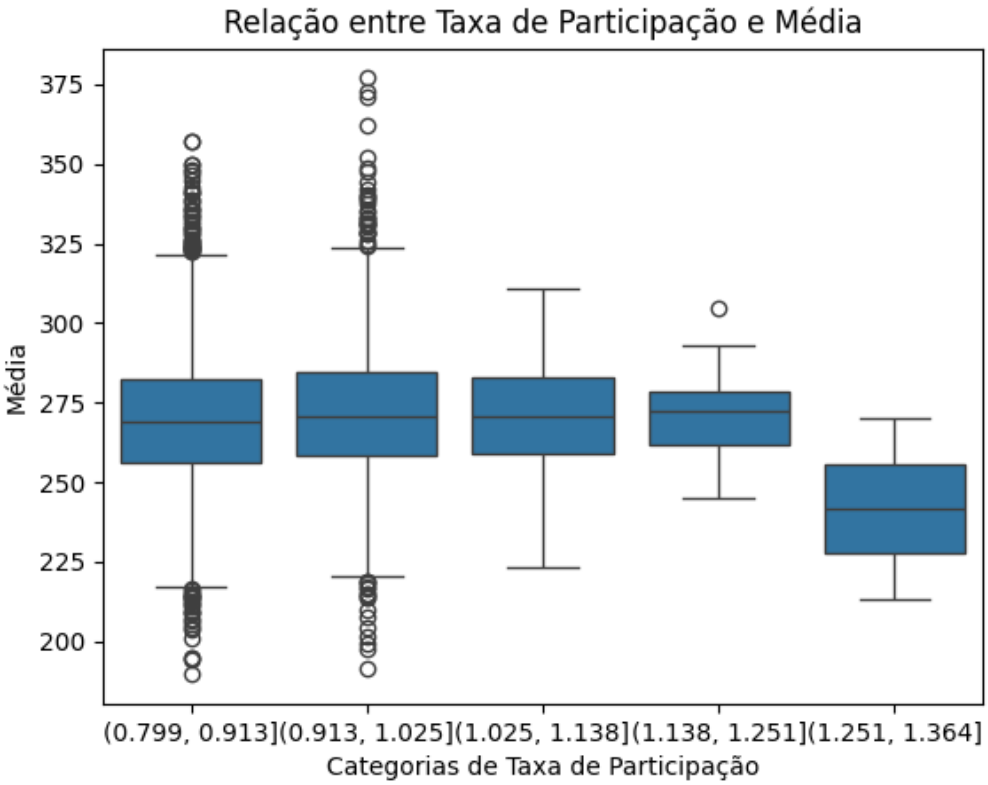
3.2.2 Second hypothesis

With the new dataset at disposal, to answer the second hypotheses, it was necessary to generate a boxplot that would correlate two quantitative variables and calculate “correl” to find the correlation, this is the result:



Understanding the graph: On the X axis, there are 10 intervals that were created to group all the different numbers that all schools have for their total of enrolled students. On the Y axis, there are the average of schools on the SAEB.

Data analysis: On this graph alone, it's clear that there are no significant or concreate correlations between the number of students and the average of the school, with the x^2 being 0.033243515805374435. What can be argued is the number of atypical points on the first interval, the one with the least number of students, although this would require further investigation, one could argue this volatility is a result of different profiles of schools amongst this interval, a school with low number of students can be seeing as either elitist or precarious, alas, further investigation is required.



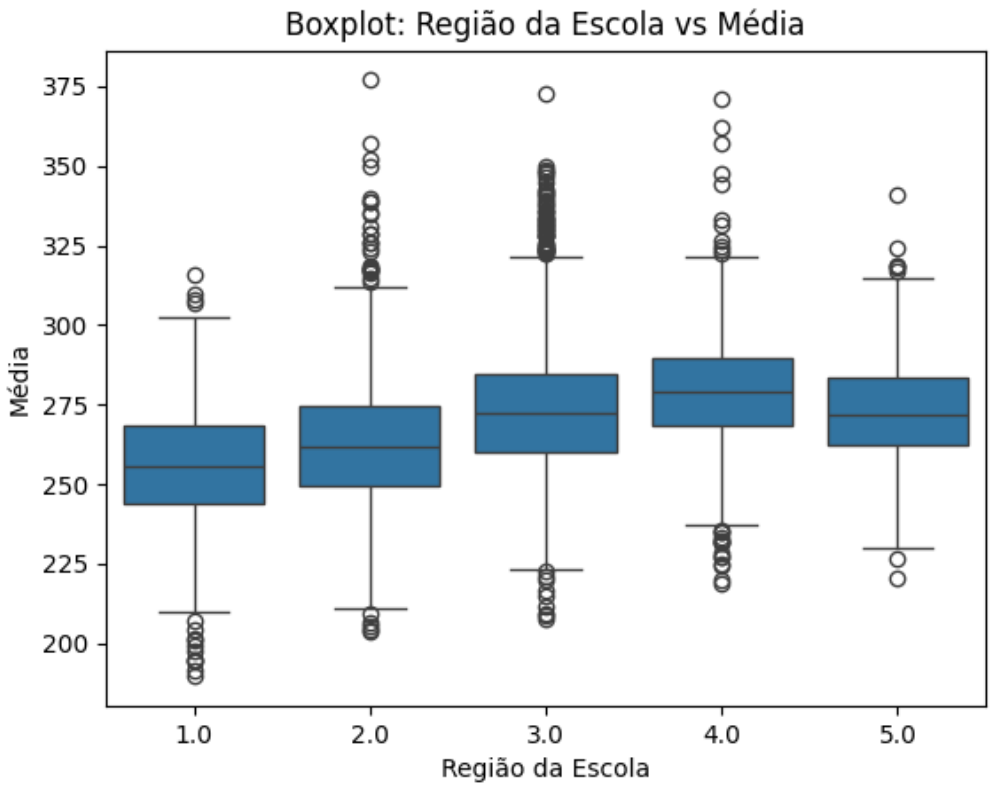
Understanding the graph: On the X axis, there are 5 intervals that were created to group all the different numbers that all schools have for their ratio of present students on the test over all enrolled. On the Y axis, there are the average of schools on the SAEB.

Data analysis: On this graph alone, its still clear that there is no correlation between the ratio of present students on the test and the average of the school, with the x^2 being 0.03833167061829204, an insignificant increase over the last x^2.

Conclusion: Does the size of enrolled students impact the school's overall score on the SAEB? Additionally, does the participation rate of the enrolled students influence the school's overall score on the SAEB? No, neither the number of enrolled students nor the participation rate have shown any correlation with the school's average.

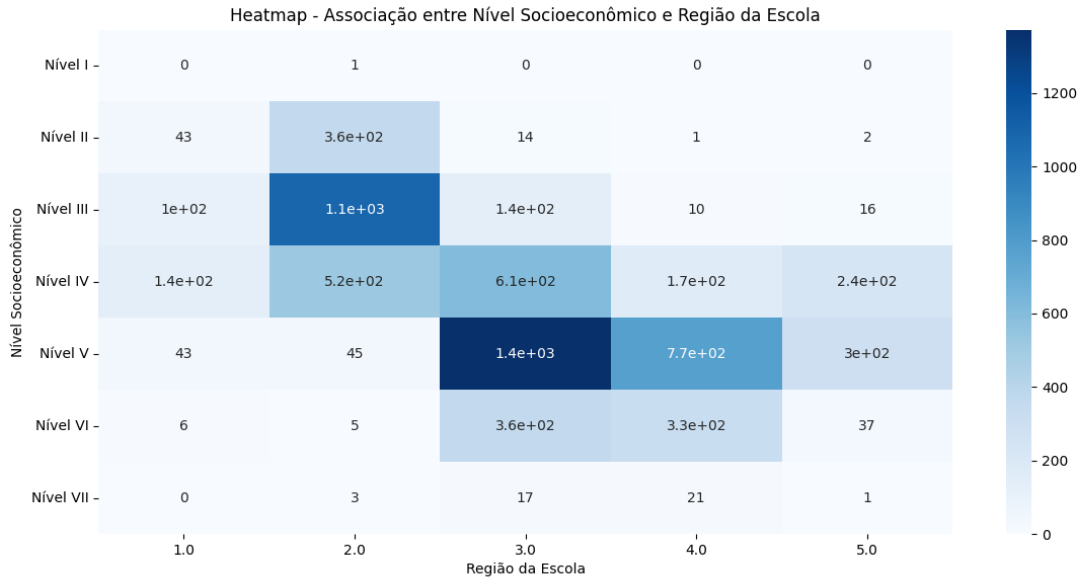
3.2.3 Third hypothesis

With the new dataset at disposal, to answer the third hypotheses, it was necessary to generate a heatmap and three boxplots that would correlate all the variables that will define the location of the school compared to the average and socio-economic levels. These are the results:



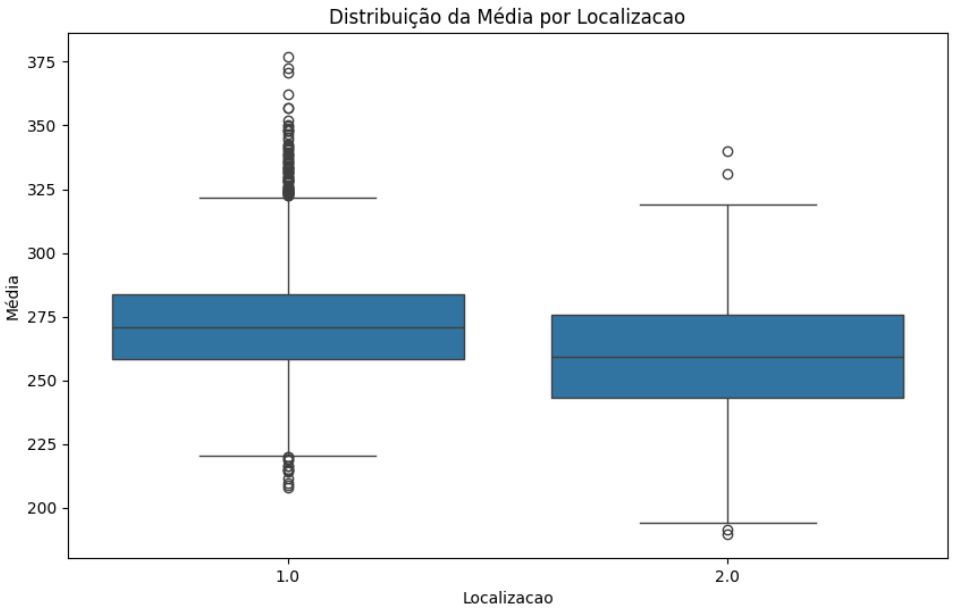
Understanding the graph: On the X axis, it is shown the Brazilian's geographic region, the number 1.0 is North; the number 2.0 is Northeast; the number 3.0 is southeast; 4.0 is South; 5.0 is Midwest. On the Y axis, it is shown the average of the schools.

Data analysis: Just by looking at the graph, its clear to see that there is discrepancy amongst all the 5 regions of Brazil. The one with the lowest average and lowest atypical points is the North, which is expected since it has some of the least developed areas across the Brazilian territory, the one with the highest average across the board is the South, which is not a surprise since it has a fewer population compared to the southeast but still one of the richest regions of Brazil.



Understanding the graph: on this heatmap, the X axis represents the Brazilian's geographic region, the number 1.0 is North; the number 2.0 is Northeast; the number 3.0 is southeast; 4.0 is South; 5.0 is Midwest. On the Y axis, it is shown the socio-economic levels of every school. The number on each box has their decimals cut, but to see the real number on each box, just look at the side bar that, with its color, indicates the scale of the number.

Data analysis: A simple look at the graph indicates a strong concentration of the higher levels of socioeconomics on the most developed regions of Brazil, while regions like North and Northeast can barely find schools that will reach the level V and above and the Northeast is the only region that has an identified Level I school on the socioeconomics scale.



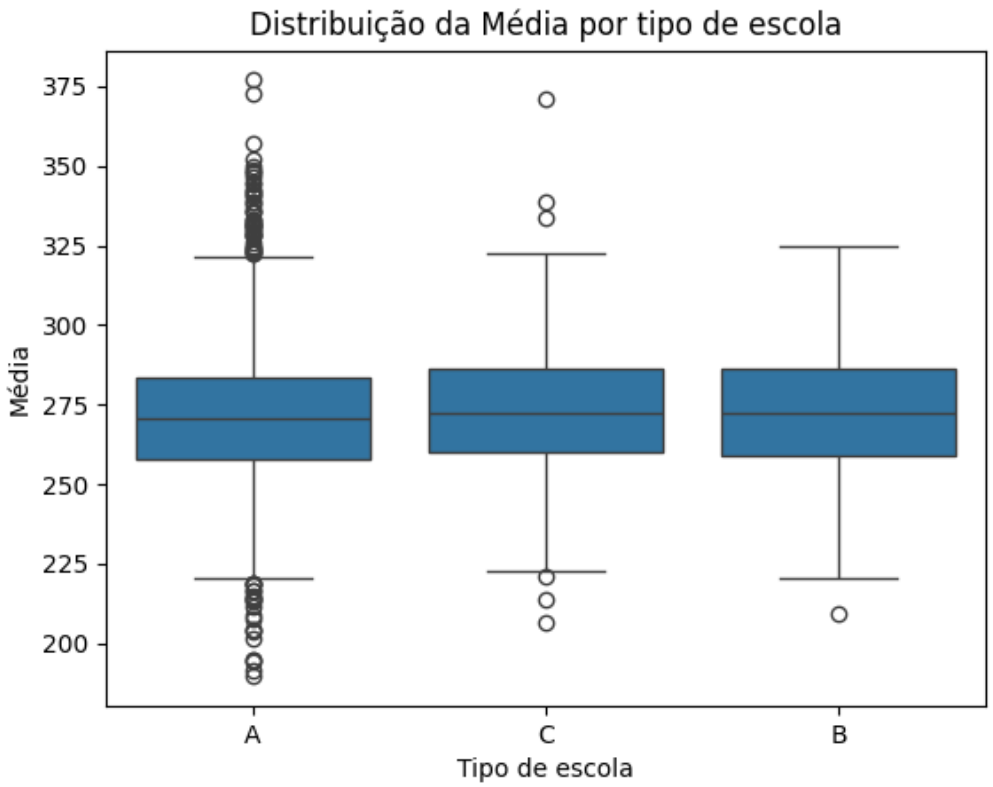
Understanding the graph: on the X axis, it is shown if either the school is located on the urban side or the countryside, on the Y axis, it is shown the average of the schools

Data analysis: It is clear that the average of schools on the urban side is higher than the average of the countryside, although, its not that high of a difference and the urban side has much more atypical points, so not every urban school is treated the same, but to compensate, it also has the highest atypical points and far most than the countryside.

Conclusion: Does the location of the school influence its socioeconomic level and its overall score on the SAEB? Yes, it has consistently shown that the score's average of the school can change significantly depending on its location, not only that, the correlation gets even stronger when comparing the socioeconomic levels.

3.2.4 Fourth hypothesis

With the new dataset at disposal, to answer the fourth hypotheses, it was necessary to generate a boxplot that would correlate the answers that the students gave about their education and their average score. This is the result:



Understanding the graph: on the X axis, its shown the valid answers of the type of education the student has received, being “A” for just public; “C” for a mixed approach; “B” for just private. The Y axis shows the average of the students.

Data analysis: The averages remains almost consistent across all types of education the student has received, although, those who answered “A” have the most amount of atypical points, indicating that the profile of those schools are not the same and can leave students with either amazing averages or the lowest averages. For consistency’s sake, a parent/guardian could see a mixed approach as the most reliable.

Conclusion: Do students who only received public education have a disadvantage compared to those who had/have access to private education? No, the average remains almost consistent across the graph, although, the number of atypical points on the “only public” alerts the parents/guardians to check the conditions of said public school in order to avoid the lower end of the atypical points, not every public school is the same.

4. **Final considerations**

This study aimed to explore the correlations between various educational and socioeconomic factors based on the 2021 SAEB dataset, with a focus on four key hypotheses related to parental involvement, school enrollment, location, and the type of education (public vs. private). Throughout this paper, there was a collection of variables that were used to show whether any significant correlation would appear to instigate further discussions and analyses on these crucial topics for the future and present of society.

The conclusions given by the analysis were as follows: No significant change was found on the comparison between the type of education students received; The location of the school has shown that will be a determining factor for its average and socio-economic levels; neither the number of enrolled students nor the participation rate on the SAEB have shown any correlation with the school's average score; While the encouragement of the parents/guardians for study is not the full picture of why kids will fail or dropout, having the encouragement can lead to better results on their academic progress.

The strongest implication derived from said conclusions were that the Brazilian government must focus its resources on dealing with the disparity shown by comparing the school location and its average performance. Schools in underprivileged or rural areas consistently showed lower academic outcomes compared to those in more affluent regions, suggesting that resource allocation needs to be more equitable. This includes not only financial support but also investments in infrastructure, teacher training, and programs that enhance family involvement in education. Additionally, policies aimed at increasing the accessibility and quality of education in public schools, particularly in disadvantaged regions, could help bridge the performance gap between these locations and promoting greater educational equity across the country.

This research had its limitations, it could not prove any correlation between the topics delt with the hypothesis, only show the interactions or lack thereof. Its an academic challenge to simplify the multiple variables and external factors that comes with sensitive topics such as education, but our goal with this research was always to intrigue fellow researchers to develop and discuss these topics at hand to find a better understanding of the dynamics of education and bring a brighter future for society, just as Nelson Mandela said:

"Education is the most powerful weapon which you can use to change the world"

5. REFERENCES

1. *Wes McKinney - Python for Data Analysis*:

- McKINNEY, Wes. Python for Data Analysis: Data wrangling with pandas, NumPy, and IPython. 2. ed. O'Reilly Media, Inc., 2017.

2. *Hands-On Data Analysis with Pandas*:

- OLIPHANT, Travis E. Hands-On Data Analysis with Pandas. 1. ed. Packt Publishing, 2018.

3. *Basic Statistics - Bussab*:

- BUSSAB, W. O.; MORETTIN, P. A. Estatística Básica. 5. ed. São Paulo: Saraiva, 2006.

4. *Seaborn Documentation*:

- WASKOM, M. L. Seaborn: statistical data visualization. Disponível em: https://seaborn.pydata.org/. Acesso em: 11 out. 2024.

5. *SAEB Dataset*:

- INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Microdados do SAEB 2021. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/saeb. Acesso em: 11 out. 2024

6. Github link for further understanding:

https://github.com/BrunoCavalli/Trabalho_A1_LP