Journal of Computer Science

Original Research Paper

Understanding a Profile of the Participants of the Exame Nacional do Ensino Médio (ENEM), Brazil, in the Year 2019, Through Data Analysis

Thiago Oliveira de Souza, Angélica Félix de Castro and Amanda Gondim de Oliveira

Department of Computer Science, Federal University of the Semi-Arid Region (UFERSA), Mossoró, Brazil

Article history
Received: 17-03-2023
Revised: 30-05-2023
Accepted: 08-06-2023

Corresponding Author:
Angélica Félix de Castro
Department of Computer Science, Federal University of the Semi-Arid Region (UFERSA), Mossoró, Brazil
Email: angelica@ufersa.edu.br

Abstract: In Brazil, there is the Exame Nacional do Ensino Médio (ENEM), which allows people to enter the University to complete their graduation course. It is a selection that takes place throughout Brazil and is the main way for a student to enter the University. Understanding the candidates' profiles is interesting to know if there is a pattern: Which courses are most chosen by women, for example, or if the North Region has a different interest than the South Region. Understanding education throughout the national territory becomes important and necessary. The main objective of this study is to verify if there are relations between some characteristics of the candidates and their performance in the ENEM. With the help of the Python language and its libraries, it was possible to find some factors that influence the performance of the exam participants and categorize some characteristics of their profiles. In general terms, it was noticed that both age and gender of the participants are not deterministic factors for their performance; that the candidates from private schools obtained higher results in all ENEM tests; that the schooling of the parents of the participants tends to influence the result obtained in the grades, among other conclusions.

Keywords: Data Science, ENEM, Python, Pandas, Matplotlib, Seaborn

Introduction

In Brazil, to enter a public or private university, it is necessary to take the Exame Nacional do Ensino Médio (ENEM). The ENEM was established in 1998, with the objective of evaluating the school performance of students at the end of basic education (INEP, 2021a). Since 2009, the exam has also been accepted as a mechanism for access to higher education and maybe a complement to an entrance exam or the entire selection process for college. Thus, the ENEM means for many students a door of access to quality and free higher education that is offered by public institutions by the Sistema de Seleção Unificada (SISU), which is a unified selection system.

It is important for a country to know the profile of people who are interested in joining the university: Which courses are most popular, whether there are more men or women in undergraduate courses, which regions and Brazilian states are the largest number of people and which courses these places offer more, among other issues.

Faced with this need, the present work has as its main objective to analyze the official database of the ministry of education of Brazil, more specifically the year 2019, which contains the data of the candidates of the ENEM 2019. The largest repository of data on education in Brazil is from the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira National (INEP), which has information about basically all levels of education and education in the country. The purpose of this study is to profile the participants of ENEM 2019 and find insights that can relate these profiles to their performance.

The main contribution of this article is to use data analysis techniques to evaluate the official database of the federal government, related to education, to analyze the profiles of candidates who wish to enter the University. Questions such as: Which courses are most popular, which courses are most desired by women, which Brazilian regions are most popular and which courses these regions offer, and what is the average age of candidates, among other questions, can be answered at the end of this study.

Section Background presents the theoretical basis for carrying out the proposed study. Section material and methods presents the materials and methods used in this article. In the section results and discussion there is a discussion of the results obtained. Finally, in the section conclusion, the conclusions are drawn about the study conducted.

Background

This section will be briefly presented: Data science and statistical concepts.

Data Science

At a high level, data science is a set of fundamental principles that support and guide the principled extraction of information and knowledge from data. Possibly the most closely related concept to data science is data mining-the actual extraction of knowledge from data via technologies that incorporate these principles (Provost and Fawcett, 2013).

This study deals with the study and application of data science techniques in the database of ENEM 2019, going through the stages of the data science cycle mentioned above, with greater emphasis on data exploration.

Statistical Concepts

Statistics deals with the collection, treatment, analysis, interpretation, and presentation of numerical data. It is present in most aspects of data science. The field of statistics is very broad, but not all its concepts are mandatory to develop studies with the help of data science (Vickery, 2021).

In this section, some fundamental concepts are presented that help in the analysis and interpretation of the results obtained in this study.

Statistical Sampling

All raw data available for study is called population. It is not always possible to use the entire population to do the desired analysis. Statistics allow it to be possible to perform the desired study based on a sample of the total population and, using probability, it is possible to have a certain degree of certainty about the characteristics of the population in its entirety.

Suppose you want to have an overview of the quality of teaching of high school graduates in Brazil. The desired study population would be all high school graduates in the country. Because it is not possible to obtain all this data, for logistical reasons, a sample representing the entire population can be used. If this sample has a good representativeness of the entire population, inferences can be made about the population in its entirety.

Descriptive Statistics

Descriptive statistics helps to describe the data and understand its characteristics. At this stage, the goal is not to formulate a prediction or inferences, it is where description of the appearance of the sample that one has been presented. Generally, descriptive statistics are obtained from the data, with the central trend means, such as:

- Average-the average value of the data
- Median-if the data is ordered increasingly, this value would be the value of the medium if we divide the set exactly in half
- Fashion-the value with the highest number of occurrences in the entire sample

Distributions

Descriptive statistics, although useful, may mask important information about the sample studied. If a dataset contains values that are much larger than others, the mean is distorted and cannot be considered a trusted representation of the data.

To represent a distribution, a histogram can be used. The histogram is a kind of bar chart that demonstrates a frequency distribution. In this chart, the base of the bars represents a class of values and the height represents the frequency that the value of each class occurs. Figure 1 shows an example of the histogram.

Another chart that can be used to analyze data distribution is a boxplot. To explain what a boxplot is, it is necessary to understand what quartiles are, which are measured in this chart.

Percentiles play an important part in descriptive statistics of continuous data and their use is recommended for reference interval estimation (Horowitz *et al*., 2010). For example, the 25 percentile of an ordered set of ages would result in an age value that would indicate that 25% of the values in that set are equal to or less than the value obtained.

The quartiles presented in the boxplot are equivalent to the 25, 50, and 75 percentiles, which represent, respectively, the first, second, and third quartiles. The second quartile is the 50 percentile, i.e., it represents 50% of the sample. Thus, the second quartile is equivalent to the median of the set. The boxplot also shows discrepant values of the set, called outliers. Figure 2 shows an example of the boxplot.

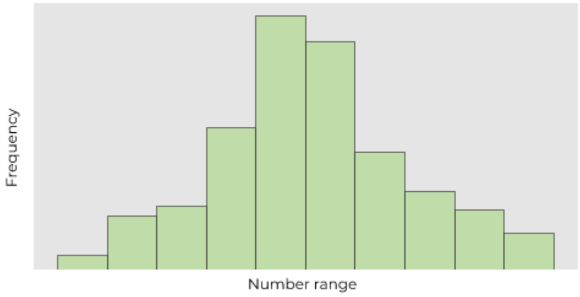

Fig. 1: Histogram example

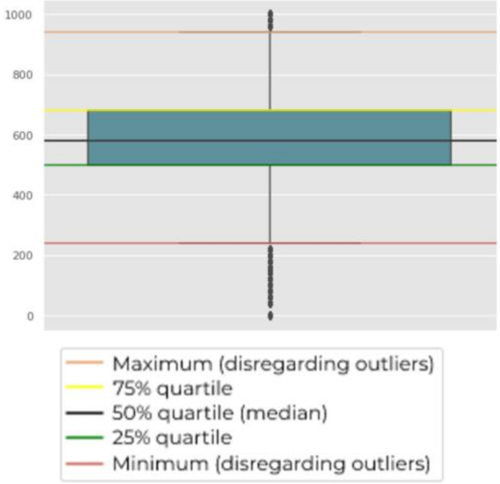

Fig. 2: Example of boxplot

The minimum value considered in the boxplot is represented by the red line, values below this minimum are represented as points and are the atypical values, considered outliers. The green line represents the first quartile, which is the 25% quartile. This means that 25% of the values in this sampling are equal to or less than this green line. The black line is equivalent to the 50% quartile and the median of the values, that is, 50% of the values of the set are equal to or lower than the value of that line. The yellow line refers to the 75% quartile, which indicates that 75% of the values in the set are equal to or less than the value of that line. Finally, the brown line represents the maximum value of the boxplot, values higher than that line are considered outliers. The minimum value and the maximum value are values considered appropriate for these ranges, considering the entire sample set. It does not mean that they are in fact the smallest and largest value of the set itself.

Correlation

Correlation is a statistical technique that measures the relationships between two variables. The correlation can be considered linear and expressed as a number between +1 and -1. This number is called the correlation coefficient. The closer to +1 or -1, the stronger the correlation, and the closer to 0, the weaker the correlation. The value 0 is considered a nonexistent correlation. The correlation coefficient sign indicates how variables correlate. A positive coefficient indicates that variables grow or decrease in the same direction. A negative coefficient indicates that while one variable grows, the other decreases.

The correlation does not imply in question, the fact that there is a correlation between two variables does not mean that one is the reason for the occurrence of the other.

Materials and Methods

This study applied concepts and rules of data analysis and statistics to evaluate the profile of candidates for ENEM 2019. Therefore, the official database of the Brazilian government was downloaded and Python commands were applied to obtain important information.

In Fig. 3, there is a flowchart of how this methodology was done.

In this section are presented: The tools used, the obtaining of ENEM data, and the description and presentation of the data processing used to carry out the proposed studies.

Programming Language

In this study, the Python programming language was used. The choice of this programming language was based on its practicality since Python has very useful libraries for carrying out the proposed study. The following libraries were used to develop the study:

- Pandas-python library used for data analysis. With it, all the reading, treatment, and processing of the data were performed
- Matplotlib-library is used to create various charts for varied data types. Much of the graphics presented in this study were made using this library
- Seaborn-library assists in the creation of graphics. It usually has a more presentable layout than the graphics created by matplotlib
- Numpy-helps to execute numerical calculations. It is mainly used to perform calculations on multidimensional arrays
- Sklearn-this library specifically assists to apply machine learning techniques to the desired dataset

For the execution environment, google collaboratory was used, or simply "Colab". It is a virtual execution environment that can be used through your internet browser, where processing and resources from Google servers are used, which allows machines that do not have such powerful hardware to handle massive databases.

Obtaining Data

On the website of the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), there is a repository with data from all previous editions of the ENEM, which can be accessed through: www.gov.br/inep/pt-br/acesso-a-informacao/dados-open /microdata/enem. For this study, we used the data from the 2019 exam.

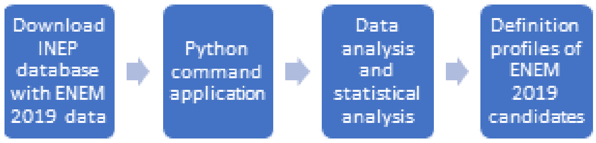
Download INEP database with ENEM 2019 data
Python command application
Data analysis and statistical analysis
Definition profiles of ENEM 2019 candidates

Fig. 3: Flowchart with the methodology used in the work

Data Processing

The .csv file containing the ENEM 2019 data is about 3.2 GB in size. When you download the database from the INEP website, the file is compressed. Because Colab is owned by Google, it has integration with some of google drive's services. Thus, through Python's zipfile library, it was possible to use the compressed database stored in google drive without the need to depress.

The main .csv (microdados_enem_2019) file contains the questionnaires answered by the participants, storing all the information provided by the participants of ENEM 2019 in a single file. This database contains 5.095,270 rows and 136 columns. This table loads a lot of data and execution errors when trying to load this information, even using a robust execution environment like Colab. The information from this main file was uploaded to a DataFrame from the Pandas library.

Pandas DataFrame is a two-dimensional data structure with tabular-aligned data in rows and columns, changeable in size and potentially heterogeneous, similar to an MS-EXCEL workbook. The essential difference is that column names and row numbers are known as column and row index, in the case of the DataFrame. Columns have names (column index) and rows can have column-referring names and rows can have names (textual indexes) or can, by default, be numbered (numeric index).

The main .csv file contains several columns that serve to describe various administrative aspects of the exam, such as the administrative dependence of the school, the color of the tests used, need for adaptations for accessibility, among others. Using descriptive analysis, the most relevant columns of the DataFrame were filtered to study the profile of participants and performance from a socioeconomic and regional perspective. Table 1 shows the columns that were considered for this study.

As the purpose of this study is to investigate relationships between the characteristics of the participants and performance, it was made the removal of empty entries from the Dataframe-that is, of enrollees who did not participate in one or more stages of the exam. After this initial clipping, the main DataFrame was reduced to 3.701,947 rows and 17 columns. This DataFrame will be referenced as base 1 in the next section.

Results

This section describes the studies performed and presents a descriptive analysis of the data obtained. Here, all the analyses made with the database of ENEM 2019 participants are presented.

Age

The exam is allowed the participation of candidates of almost all ages. In Base 1, candidates from 10-92 years old appear. Despite the wide range of ages, there are some more common occurrences, the median age of the participants is 19 years. As shown in the histogram in (Fig. 4), most candidates (89.82%) are between 15 and 29 years old. The most representative interval is that of participants between 15 and 19 years, containing 2.232,119 candidates and representing 60.30% of the whole set.

Figure 5 presents a boxplot with the age distribution of the participants. It is noticeable that the ages are concentrated between 15 and 29 years, as mentioned earlier. Values over 29 years are considered outliers.

Given the wide age range, the subsequent question is whether the age difference affects the test result decisively. For this study, the arithmetic mean of the five participants' scores was made to obtain a notion of their overall performance. The same age range (Fig. 3) was considered to investigate the candidates' performance. For each age interval, the sum of all means was made and then the arithmetic mean of this sum was taken (Fig. 6).

Figure 6 above shows that age does not have much influence on the score obtained in the examination. Although the density of each age range is different, constancy in the values of the average of the notes is maintained for all of them.

Table 1: Columns used in the present study

Name of column	Description
Nu-inscricao	Registration number
Co-municipio-residencia	Residence municipality code
Nu-idade	Age
Tp-sexo	Sex
Tp-cor-raca	Self-declared color/race
Tp-escola	Type of high school
Nu-nota-cn	Nature sciences test note
Nu-nota-ch	Humanities test note
Nu-nota-lc	Language and code proof note
Nu-nota-mt	Math test note
Nu-nota-redacao	Note of the writing test
Q001	Until what grade did your father, or the man responsible for you, study
Q002	Until what grade did your mother, or the woman responsible for you, study
Q006	What is your family's monthly income? (add your income to that of your family members
Q022	Do you have a cell phone at your residence
Q024	Do you have a computer at your residence
Q025	Do you have internet access at your residence

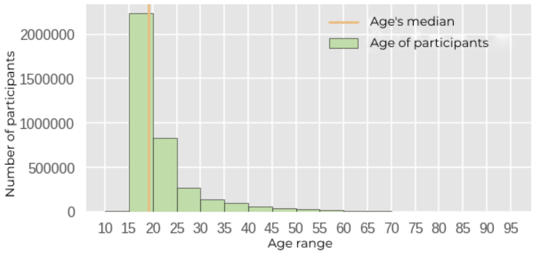

Fig. 4: Histogram of ages of ENEM 2019 participants

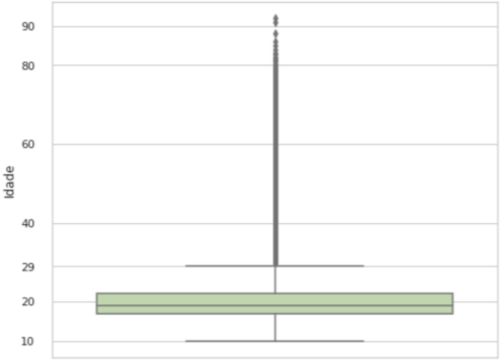

Fig. 5: Boxplot of the ages of ENEM 2019 participants

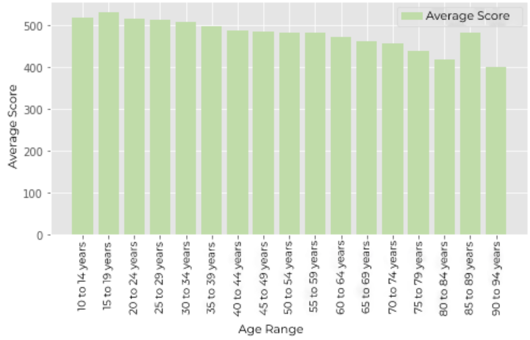

Fig. 6: Average grades by age

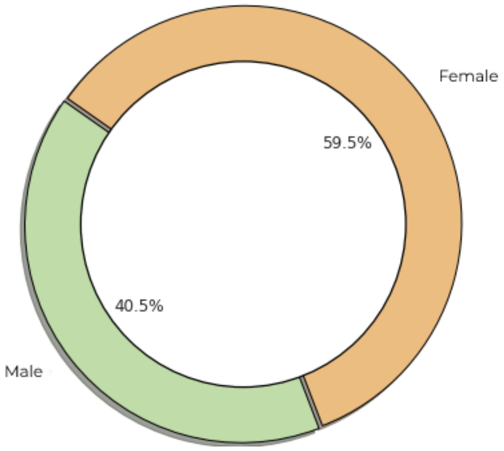

Fig. 7: Distribution of ENEM 2019 participants by gender

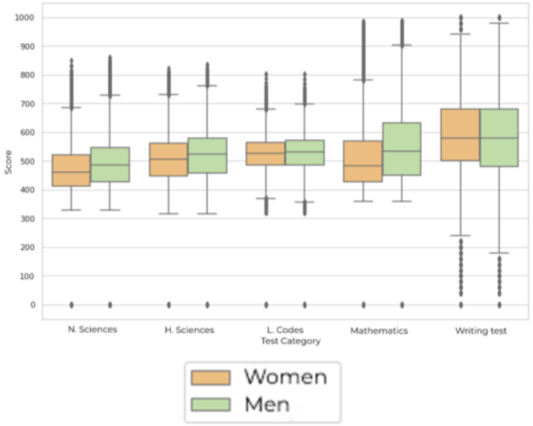

Fig. 8: Performance of participants by sex in the five tests

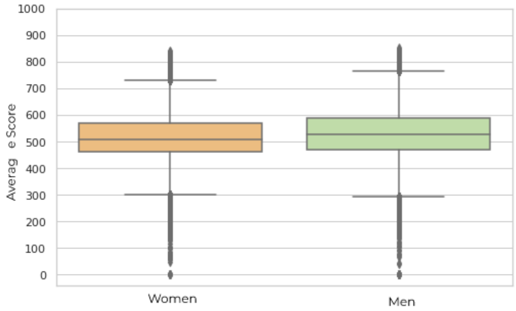

Fig. 9: Average tests by sex

Sex

This section analyzes the performance of participants by gender. In Base 1, there are a total of 3.701,947 entries, where 2.201,184 of these entries correspond to female participants and 1.500,763 correspond to males. In Fig. 7, you can view the distribution of candidates by gender.

The participants were divided into two groups, one corresponding to the female and the other to the male sex. Figure 8 shows a boxplot representation of the grades in the five competencies evaluated in the exam.

It is noticeable that in the competencies of natural sciences, human sciences, and languages and codes; men have performed slightly better overall. To get a more precise notion, the average per sex was calculated in each test. In natural sciences, men have an average of 4.12% higher than women, in human sciences a superiority of 2.71%, and in languages and codes an average of 0.42% higher. In mathematics, it can be observed that the score of male participants obtained a more detachable apex compared to females, with an average of 8.27% higher. In the writing note, women obtained an average of 2.52% higher than men.

For a more direct visualization, we used the calculation of the arithmetic mean of the five disciplines contemplated in the ENEM. Figure 9 shows a boxplot that groups candidates by gender and references the set of means that each of these groups obtained in the exam.

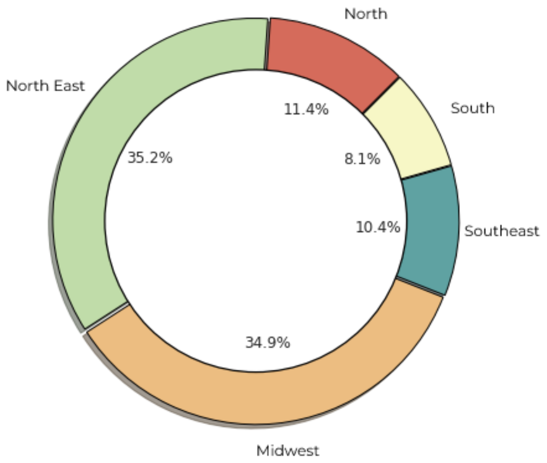

Fig. 10: Distribution of ENEM 2019 participants by region

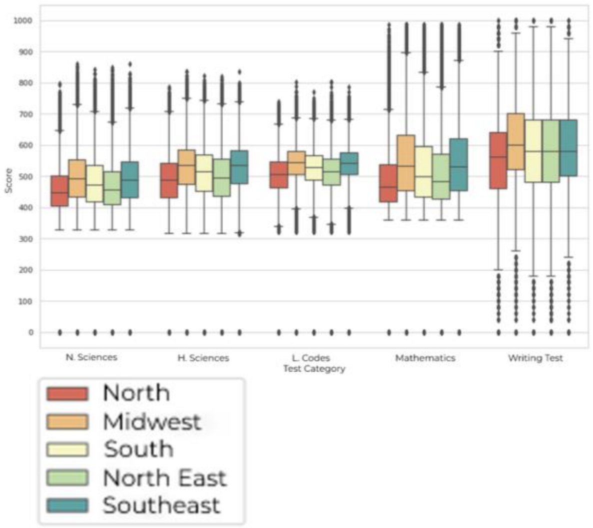

Fig. 11: Distribution of notes by region

Analyzing the boxplot with the average of the five competencies of the exam, it is noticeable that sex is not a determining factor for the overall performance of the candidate.

Region

In this section, the analysis of the distribution and performance of the participants by the region in which they reside is made. Figure 10 shows the distribution of ENEM 2019 participants by region of Brazil.

Some peculiarities can be noted when analyzing this distribution. The Midwest Region is the second most representative in the number of participants, even though the region is less dense among the five regions, according to the last population census conducted by IBGE (2008). This may reflect the ease of access that students in this region have to education and maybe a study tool for improvements in other regions of the country. A similar parallel occurs in the Southeast Region. Even with the largest number of inhabitants in Brazil, it is the second least representative in the number of candidates in ENEM 2019. This also provides a study scenario for understanding the low adherence of these students in relation to the other regions.

Next, the score of the participants of each region was performed in the five competencies of the examination. Figure 11 shows through boxplots the grades of each region by the category of the race.

Analyzing the boxplot in Fig. 10, it can be noted that the Centro Oeste obtains the highest median in all tests, except for humanities, where it is behind the Southeast by only 0.5 points. It is also observed that the North holds the lowest median of the grades among the regions, in all tests. For better visualization, Table 2 shows the median value in each test by region.

The Índice de Desenvolvimento da Educação Básica (IDEB) is a basic education development index and was created in 2007 by INEP with the objective of measuring the quality of learning in the country and setting goals for improvements in education. The IDEB is calculated from the school performance rate (approval) and average performance in the exams applied by INEP. The approval rates are obtained from the School Census, which is done annually (Ministério da Educação, 2018).

On the IDEB website, it is possible to find all the results of the index from 2005-2019, containing the indicators for 4th grade or 5th year, 8th grade or 9th year, and 3rd grade. The site can be accessed at: http://ideb.inep.gov.br. As this study makes a study of ENEM 2019, the data from IDEB 2019 for the 3rd grade of high school were considered. The score of each state was collected and the average IDEB per region was calculated. Table 3 shows the result obtained by each region.

Although the Midwest obtained the best results in the ENEM race, he is third in the IDEB result. On the other, the Northeast and North Regions present a very reliable result to the index. While the Northeast was the penultimate place in the results of the ENEM and IDEB, the North was the last place in both.

As seen in Fig. 11 and Table 2, it is remarkable that the Midwest has the best results, reaching a difference of 68.8 points from the last place in mathematics. In view of the performance of the Midwest Region, another opportunity for analysis is presented. It is possible that public agencies and managers develop studies of the teaching characteristics of the region that justify this superior performance in the examination. The study otherwise is also valid. The search for common characteristics of teaching in the North may present indicators of why it has the worst performance in the tests among the other regions.

Table 2: Median of grades in each test by region

Exam	Midwest	Southeast	South	Northeast	North
Natural sciences	492,2	487,8	470,2	454,9	446,3
Humanities	533,5	534,0	512,8	494,5	486,1
Languages and codes	543,5	541,3	527,5	514,2	505,8
Mathematics	533,0	529,5	499,1	482,4	464,2
Redaction	600,0	580,0	580,0	580,0	560,0

Table 3: Result in IDEB 2019 (3^{rd} grade of high school)

Region	IDEB 2019
Southeast	4,42
South	4,36
Midwest	4,27
Northeast	3,92
North	3,78

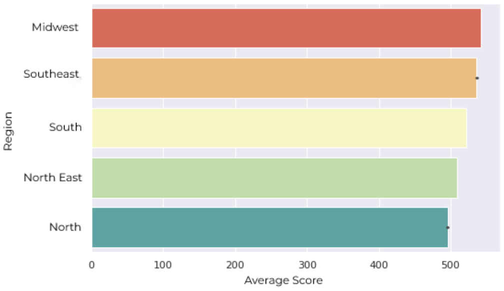

Fig. 12: Average grades by region

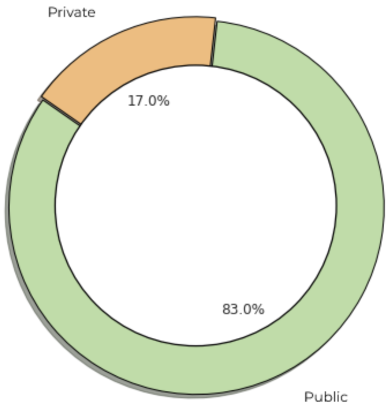

Fig. 13: ENEM 2019 participant education network

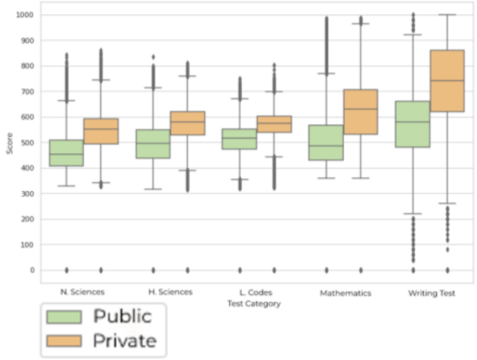

Fig. 14: Distribution of grades by education network

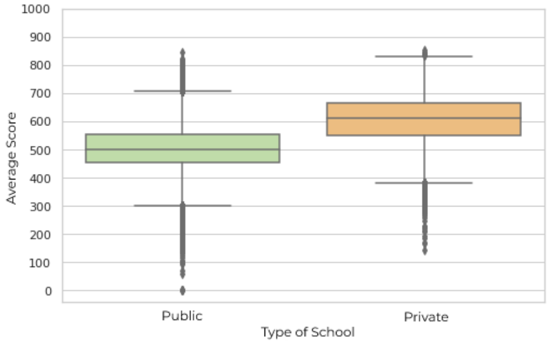

Fig. 15: Average tests per school

Figure 12 shows the average of the five tests distributed by region. As previously seen in the individual distribution of each discipline, the same classification was expected to follow. There is: The Midwest with the highest average of the notes, followed-in order, by the Southeast, South, Northeast, and North.

School

In this section, the distribution and performance of participants by the type of school in which high school was completed are analyzed. The lines of base 1 were excluded in which the candidate did not answer at which school attended high school. The classifications remained: Public, private, and external. In the entire database there was no occurrence of schools from abroad, so the analysis took place between public and private schools. After these cuts, there are 1.009,821 students from public schools and 207,299 from private schools.

Figure 13 shows the percentage of representativeness of each of these education networks in ENEM 2019.

Next, the grades obtained in each competency for each educational network were studied. The performance of each school is observable in (Fig. 14).

As you can repair, the private network gets the best results on all exam tests. The average of each school was calculated for each of the tests contemplated in the exam.

In natural sciences, private schools achieved an average of 17.31% higher than public schools. In human sciences, the average is 15.36% higher. The proof of languages and codes obtained 11.18% more than the public network. In mathematics, the result was 22.90% higher. Finally, in writing, private schools had a more expressive result, with an average of 30.70% higher than public schools.

Table 4: Distribution of participants by self-declared race

Race	Quantity	Percentage %
Brown	1.694.136	45,8
White	1.374.887	37,1
Black	453.218000	12,2
Yellow	84.752000	2,3
Undeclared	73.432000	2,0
Indigenous	21.522000	0,6

Table 5: Type of school by self-declared race

School	White %	Yellow %	Brown %	Black %	Indigenous %
Public	72,8	82,2	89,8	91,7	94,3
Private	27,2	17,8	10,2	8,3	5,7

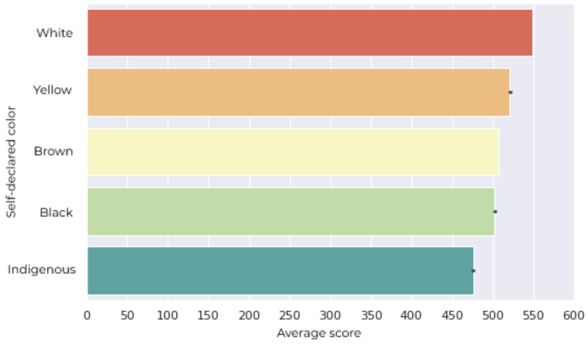

Fig. 16: Average grades per self-declared breed

Figure 15 shows the boxplot representation of the arithmetic means of the five tests for each group of schools.

As expected, it is possible to see a reflection of what was found in Fig. 12. When analyzing boxplots, it is remarkable that the lowest grades in the private network have almost the same value as the highest grades in public schools. This attests to the great difference in performance between the two education networks. On average, the private education network performs 19.75% more than the public.

Self-Declared Race

The study conducted in this section concerns the race self-declared by the participants. Table 4 shows the number and percentage of participants in each group. Racial nomenclatures were described in the same way as the form provided by INEP.

Figure 16 shows the overall mean obtained by each race group. Candidates who did not declare any race were not included in the chart. The self-declared white participants are the second most representative group and are the ones who get the best grades. Then comes, in order, the yellows, browns, blacks, and indigenous.

Table 5 shows the distribution of these races by the type of school in which high school was completed. There is a certain tendency in groups by schools. For example, whites have a higher percentage of private school students than yellows. Yellows, in turn, have a higher percentage of private institutions than browns. The average value of the breeds follows this trend, the higher the representativeness of members of private schools, the higher the average obtained. Which makes sense considering the analysis made in the previous section.

By analyzing Fig. 16 and Table 5 together, some results are surprising. For example, whites scored an average of 9.16% higher than blacks, which is not such a different difference, considering that 91.7% of black people attended the third year of high school in public schools. As seen in the previous section, the performance of participants coming from public schools is usually lower.

Historically, indigenous and black people are more marginalized racial groups, in various sectors, especially in opportunities for access to education. When analyzing Fig. 16, it is perceived that the difference in performance between races is existing, but is not abyssal. For example, the largest performance difference is between whites and indigenous peoples, where whites have an average of 15.36% higher grades. Considering that 94.3% of the indigenous peoples come from public schools, this difference is not so huge. This opens an opportunity to study how better the results of these less-favored breeds could be if everyone were on an equal footing in the opportunities for access to education.

Parents' Education

The data provided contain data on the education of the parents of the participants. This information is arranged in columns Q001 and Q002, in these columns there are the following possibilities of answers:

- A-Never studied
- B-Did not complete the 4th grade/5th grade of elementary school
- C-Completed the 4th grade/5th grade but did not complete the 8th grade/9th grade of elementary school
- D-Completed the 8th grade/9th grade of elementary school but did not complete high school
- E-Completed high school but did not complete college
- F-Completed college, but did not complete graduate school
- G-Completed graduate school
- H-I don't know

The DataFrame lines containing the H response (403.808 lines) were not considered. Answers A and B were considered just one: Never studied. The answers F and G were condensed into only one, in the form of:

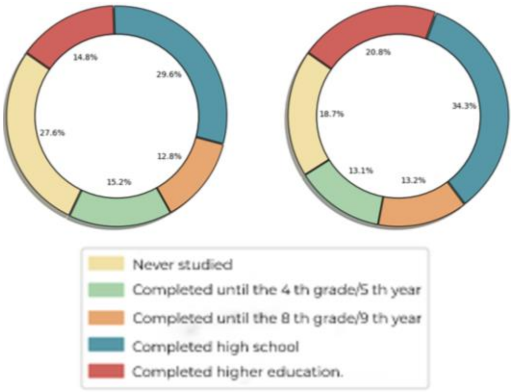

Fig. 17: Distribution of schooling of responsible

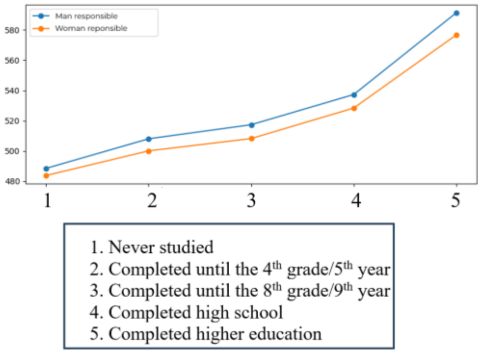

Fig. 18: Average on examination x education

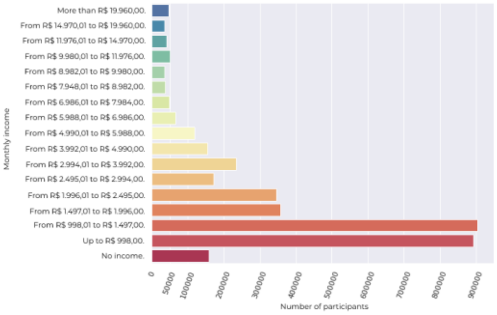

Fig. 19: Number of participants x monthly income

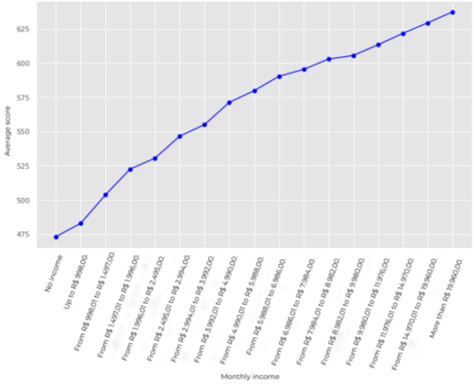

Fig. 20: Average in ENEM 2019 x monthly income

Completed higher education. After the adjustments, there are the following possible answers:

- He never studied
- Completed until the 4th grade/5th year
- Completed until the 8th grade/9th grade
- Completed high school
- He completed higher education

Figure 17 shows the distribution of the level of education of the responsible candidates of ENEM 2019.

It is possible to note that there is a considerable difference between those responsible. In the higher levels of schooling, it is noticeable that the female group is more representative. While the percentages of people with lower educational levels are more represented by male participants.

The subsequent question is whether the education of the responsible affects the performance of the participants. Figure 18 shows the general average in the exam according to the parent's level of education.

It is noticeable that the group of students with parents with higher education has higher averages generally. It is remarkable that for both sexes there is a tendency for the higher the degree of teaching of the responsible, the higher the grade obtained by the candidate.

Income

This section is intended for the analysis of the income reported by the participants and their influence on their performance. Figure 19 shows the number of participants per economic profile.

62% of the participants have an income of a maximum of R$ 1996.00. The participants with the highest purchasing power represent the smallest parts of the candidates, which reflects the Brazilian social reality.

Figure 20 shows that participants with higher purchasing power had better averages than the others overall. However, this is an isolated representation of the mean in the examination according to each income group. For an accurate investigation of the correlation between income and the performance of participants, it would take numerical data different from the participant's income, but the data provided by INEP are categorical data. When answering the socioeconomic questionnaire, participants do not report the exact value of family income, a range of values is selected in which their income falls. This data makes it categorical and makes it impossible to study the correlation between the performance and purchasing power of the participant.

Access to Technology

With the internet increasingly accessible, many activities that were previously done by physical means are performed online. One of these activities is the study.

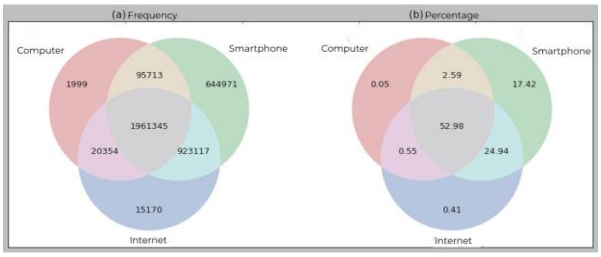

Fig. 21: Venn diagram

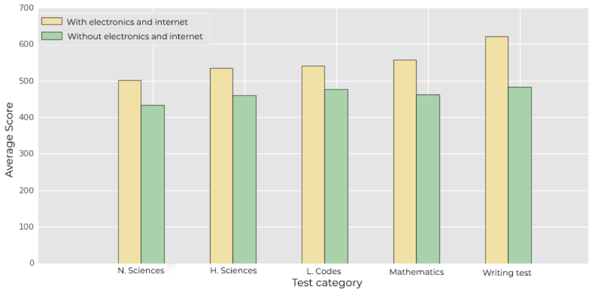

Fig. 22: Average in tests according to access to technology

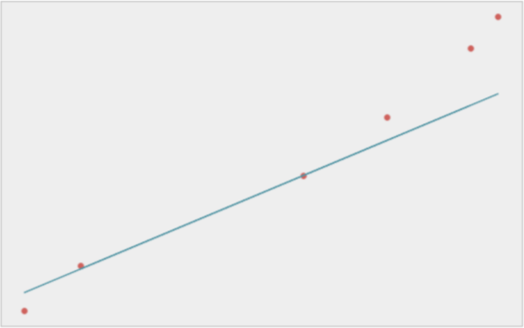

Fig. 23: Trend line obtained with training data

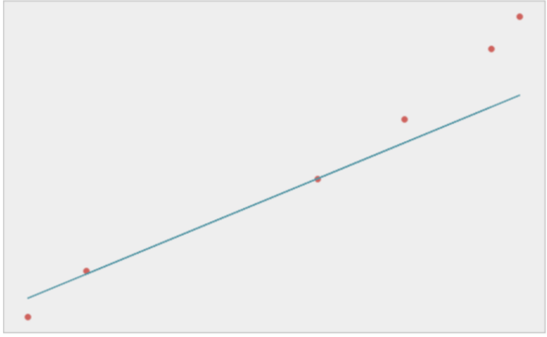

Fig. 24: Trend line with tested data

Today there is a wide range of digital content and courses that can assist in the learning journey. This section makes a study of the technologies that ENEM 2019 participants have at their disposal.

In the data provided by INEP, columns Q022 and Q024 refer, respectively, to the possession of mobile phones and computers by the participants. These columns have four possible answers, one in case you don't have any of these devices and the others to detail the amount. For the study of this section, it was considered only whether the participant has or not the equipment, disregarding the quantities he has. Finally, column Q025 refers to the applicant's access to the Internet.

For visualization purposes, three groups were created: Computer, Mobile, and Internet. Figure 21 contains a Venn diagram that shows the number and percentage of participants allocated by groups.

Only 39.278 participants (1.06%) reported not having access to any of these three technologies. A comparison was made between candidates who have access to the three technologies and those who do not have access to any of them. Figure 22 shows the overall mean obtained in each of the examination tests by these groups.

In general, participants who have access to electronics and the internet have an average higher than the others. This difference reaches the highest value in the average writing score, where it is 28.58% higher. 95% of participants who do not have a computer, mobile phone, and internet access have an income of a maximum of R$ 1,497.00. Thus, it is noted the importance of a public and free space so that students who do not have access to these learning tools can enter the virtual study environment.

Machine Learning

The ENEM has had a volatile number of participants over the years. In this section, the use of machine learning techniques is made with data regarding the number of participants per year to obtain a trend line referring to those enrolled in the exam in the coming years.

Table 6 contains the number of participants in the 1998 exams to 2019. These data were made available by INEP (2021b).

To obtain a trendline, it was necessary to train and test the data. To develop this study, Pandas, Numpy, Matplotlib and Sklearn libraries, all Python libraries, were used. It was necessary to import the necessary libraries, create the DataFrame, divide the test and training data, create the regression model, train the model and finally generate the figure with the trend line after training.

By default, 75% data were reserved for training and 25% for testing. The amount of data available for this study is not so large, since there were only ENEM data from 1998. The predictive model created makes use of linear regression to obtain a trend line for the number of subscribers per year. Where the number of participants per year is the dependent variable and the year of the exam is the independent variable. Figure 23 shows the line obtained with the training data used.

The training has generated a growing line. To continue the study, the test data were applied to investigate the result. Fig. 24 presents the result with the test data.

Table 6: Year and number of participants

Year of ENEM	Number of people	Year of ENEM	Number of people
1998	157.221	2009	4.148.720
1999	346.953	2010	4.626.094
2000	390.180	2011	5.380.856
2001	1.624,131	2012	5.791.065
2002	1.829.170	2013	7.173.563
2003	1.882.393	2014	8.722.248
2004	1.552.316	2015	7.746.472
2005	3.004.491	2016	8.627.367
2006	3.742.827	2017	6.731.341
2007	3.584.569	2018	5.513.747
2008	4.018.050	2019	5.095.270

It is remarkable that both training and test data have generated a growing trend line, which indicates that the number of participants in ENEM in the coming years tends to grow. To verify whether this trend would remain in other tests, the model was put running with the years 2021 and 2025. For the year 2021, the model resulted in a total of 8,089,014 participants, and for the year 2025 a total of 9,402,842 subscribers, which shows that the model follows the growing trend that it predicted. This is a purely numerical analysis, the model does not have variables such as the emergence of the pandemic, which resulted in a lower adherence to the last editions of the exam.

Conclusion

Social inequalities in any country have a great influence on the quality of education that the population has access to. Thus, the analysis of school performance from a socioeconomic background is of great importance for creating debates about the social differences that influence the maintenance of these inequalities.

This study is framed in the application of the concepts of data science in Brazilian educational data. The objective was the application of statistical techniques and data science for the discovery of knowledge about the participants of the ENEM 2019, in the database provided by INEP.

The results of the first studies allowed us to conclude that both the age and gender of the participants are not deterministic factors for their performance. The subsequent study showed that the Midwest has the best test results and is the second most representative in the number of participants, even though the region is less dense in the country. It was also noticeable that the Northeast and North regions occupy penultimate and last place, respectively, in all the competencies of the examination.

The following analysis was related to the type of school in which the participant completed high school. It was possible to verify that the candidates from private schools obtained higher results in all ENEM tests, with emphasis on writing, where the group of students from private schools was an average of 30.70% higher than the participants of the public. In the general, average of the five tests, private schools had a superiority of 19.75% over public institutions.

Regarding the self-declared race of the participants, a difference in the means obtained was noticeable. The classification of the groups by average obtained was whites, yellows, browns, blacks, and indigenous peoples. Although the difference in means is there, it was not such a significant difference. Another interesting observation is that the groups with higher averages have a higher representation of participants coming from private schools compared to the other groups.

It was also observed that the schooling of the parents of the participants tends to influence the result obtained in the grades. The higher the level of education of the participants responsible, the higher the score obtained in the exam. A similar trend was observed in the income analysis of the participants. The candidates were grouped by income group and it was seen that the higher the purchasing power, the higher the average obtained by the group.

Another analysis was the access to technology by the candidates. It was seen that participants who have access to technology as support for the study, have an average of higher grades than participants who do not have access to this facility. Candidates with access to technology obtained an average of 28.58% higher writing scores. It was also verified that most candidates without internet access and electronic equipment, are low-income.

After the study, it is inferred that data science is an effective way to study massive databases. It was noticeable, in a shallow way, that the social differences outline an influence on the result of the means obtained. This study makes an initial and superficial analysis of ENEM data, to find insights for future studies. It is noted that it is necessary to deepen the analysis of the socioeconomic aspects of the students, to find more direct relationships.

Acknowledgment

The authors would like to thank the Federal University of the Semi-Arid Region (UFERSA) for all the infrastructure made available.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

Thiago Oliveira de Souza: Realized all the research, obtained the results, and wrote the text in Portuguese.

Angélica Félix de Castro: Guided the conduct of the work and translated from Portuguese into English.

Amanda Gondim de Oliveira: Guided the conduct of the work and revised the text in Portuguese.

Ethics

Authors should address any ethical issues that may arise after the publication of this manuscript.

References

Horowitz, G. L., Altaie, S., & Boyd, J. C. (2010). Defining, establishing, and verifying reference intervals in the clinical laboratory; *Approved Guideline*. CLSI. https://clsi.org/media/1421/ep28a3c_sample.pdf

IBGE. (2008). Instituto Brasileiro De Geografia E Estatística-Contagem da População 2007 (2ª. Ed.). Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística-BGE. ISBN: 978-85-240-4004-7. https://biblioteca.ibge.gov.br/visualizacao/livros/liv93420.pdf

INEP. (2021a). Exame Nacional do Ensino Médio (ENEM). Site oficial do Governo Federal Brasileiro Ministério da Educação. https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem

INEP. (2021b). Instituto Nacional De Estudos E Pesquisas Educacionais Anísio Teixeira. Microdados. Site oficial do Governo Federal Brasileiro-Ministério da Educação. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados

Ministério da Educação. (2018). IDEB-Apresentação. Site oficial do Governo Federal Brasileiro-Ministério da Educação. http://portal.mec.gov.br/conheca-o-ideb

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*, *1*(1), 51-59. https://www.liebertpub.com/doi/full/10.1089/big.2013.1508

Vickery, R., (2021). Fundamental Statistical Concepts for Data Science. https://towardsdatascience.com/8-fundamental-statistical-concepts-for-data-science-9b4e8a0c6f1c