**Name: Gabriel Augusto Nascimento da Silva Costa**

**Descriptive Statistics Treatment and Analysis - Microdata - ENEM201**

**GITHUB Repository: Treatment-of-data-ENEM-2019**

**Preliminary code for data loading and treatment: Load_Microdados.py**

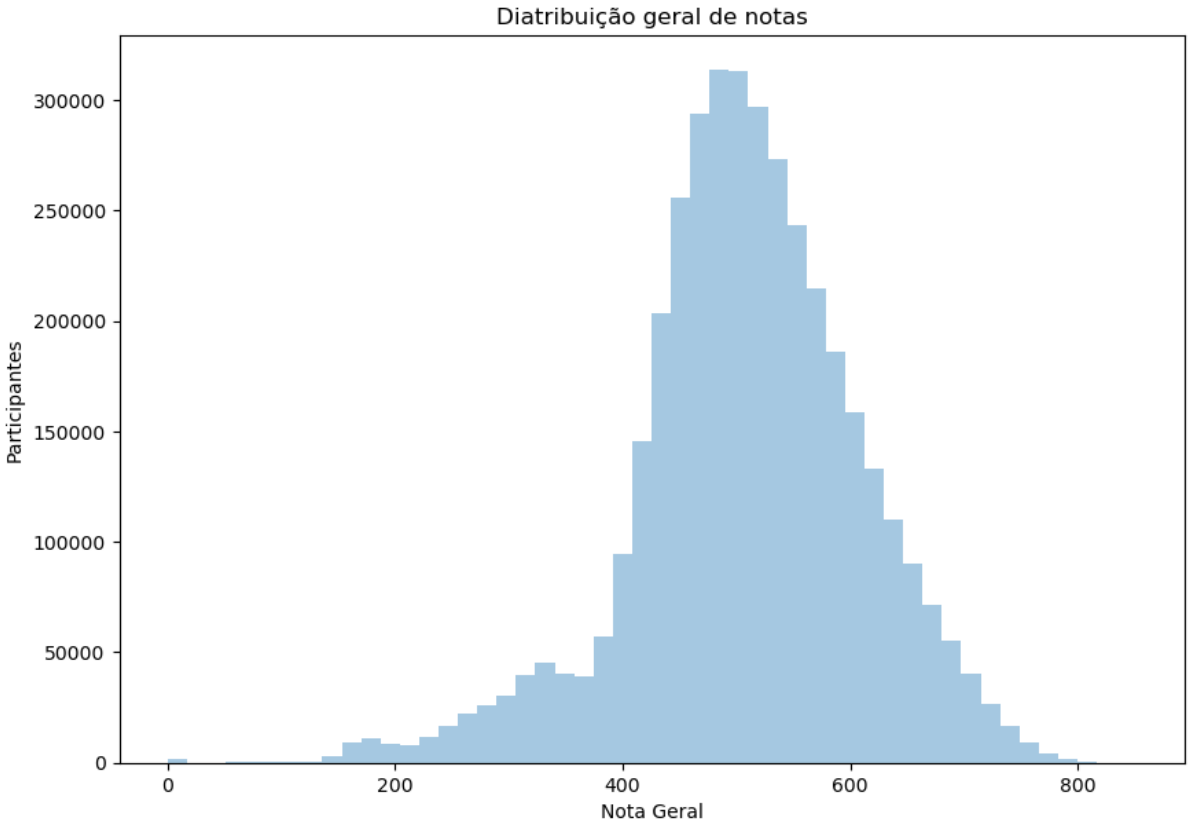
**Plotting and analysis: Análise_Estatística.py, Análise_gabarito.py**

**Note: Partitioning of the data and use of remote compilation platforms (Google Collab) were necessary for some analyses due to dataset size and the available hardware limitations.**

Starting from some direct and fundamental inferences, here are some basic distributions.

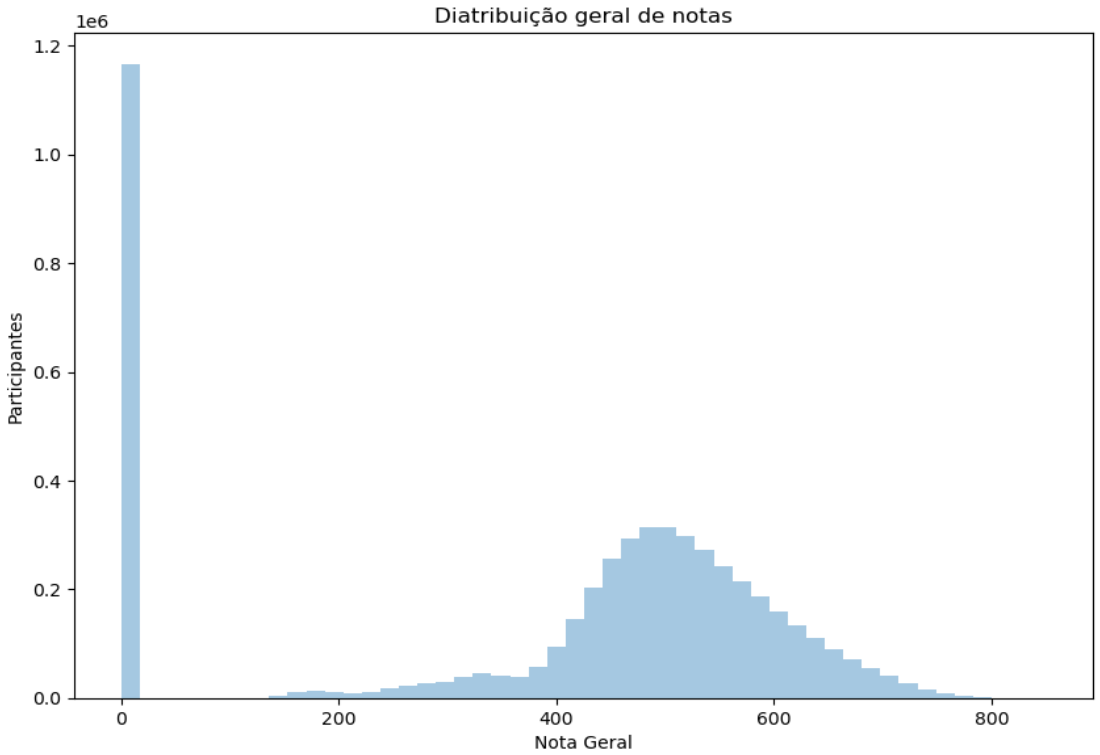
**Overall score distribution:**

For this distribution, only participants who had full attendance throughout the test were considered, missing values due to non-attendance were disregarded.




**Histogram of continuous variable 'nota_geral' (Overall Score) with full attendance.**
**Mean -> 508.72**
**Median -> 508.88**

On the contrary, if we consider non-attendance as 0, we have quite a different distribution:




**Histogram of continuous variable 'nota_geral' considering absences as 0**
**Mean -> 392.01**
**Median -> 477.12**

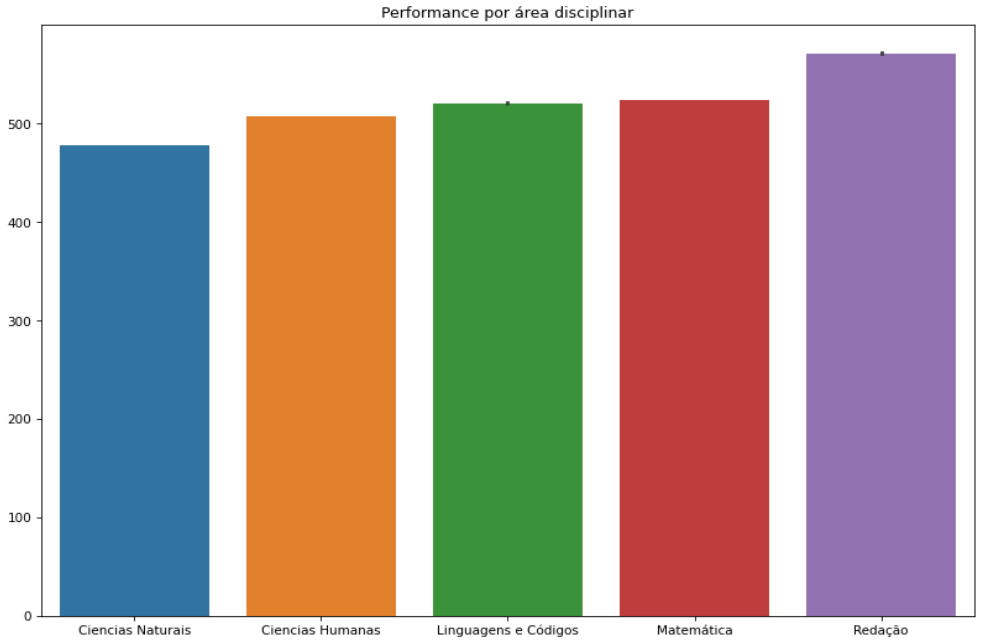
Based on the last distribution, it is possible to infer that there is a large number of non-attendances in the tests, as demonstrated by the following occurrence table, where each row represents a common combination among various candidates (Ex: First row = Present in all tests; Last row = Eliminated in all tests).

| Natural Science | Human Science | Math | Languages and Grammar | Count | Percentage |
|---|---|---|---|---|---|
| Present | Present | Present | Present | 3702008 | 72,65577683 |
| Missing | Missing | Missing | Missing | 1160010 | 22,76640885 |
| Missing | Present | Missing | Present | 219245 | 4,302912309 |
| Present | Missing | Present | Missing | 8027 | 0,157538266 |
| Missing | Disqualified | Missing | Disqualified | 3670 | 0,072027586 |
| Disqualified | Present | Disqualified | Present | 1892 | 0,037132478 |
| Present | Disqualified | Present | Disqualified | 398 | 0,007811166 |

| | | | | | |
|---|---|---|---|---|---|
| Disqualified | Missing | Disqualified | Missing | 16 | 0,000314017 |
| Disqualified | Disqualified | Disqualified | Disqualified | 4 | 7,85042E-05 |

It is interesting to note that the total number of absences (not attending to any test) represents nearly 23% of the registrations, while cases of elimination represent just over 0.1%. For this reason, most subsequent analyses only consider candidates who took the test at some point.

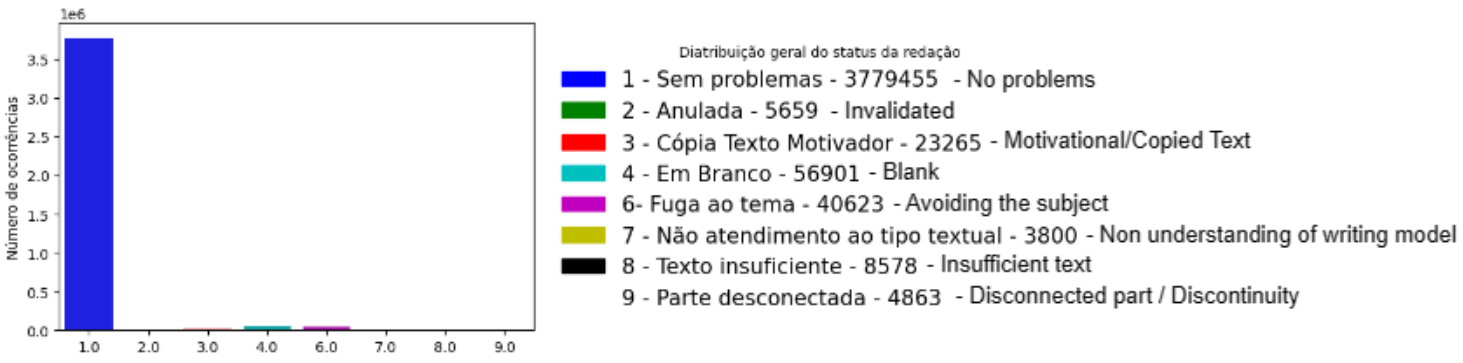
**Score comparison by subject:**



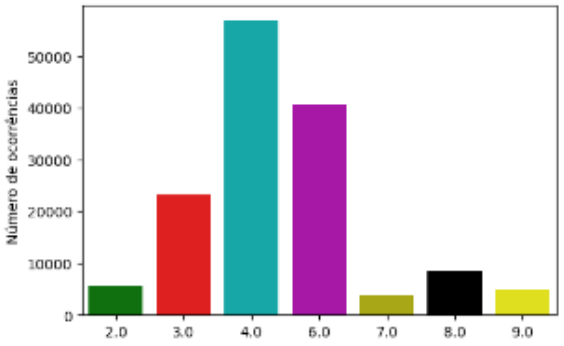

**Bar graph comparing score central tendency in different disciplines.**

Although the relationship does not necessarily allow us to make a more generalist inference, we can inform ourselves on the lowest scoring subject pertaining to the 2019 test.

**Essay Status Distribution:**



As can be seen, abnormal occurrences during the essay represent a small minority of all cases. For better visualization purposes, we can isolate the abnormal statuses and analyze them separately as provided below.

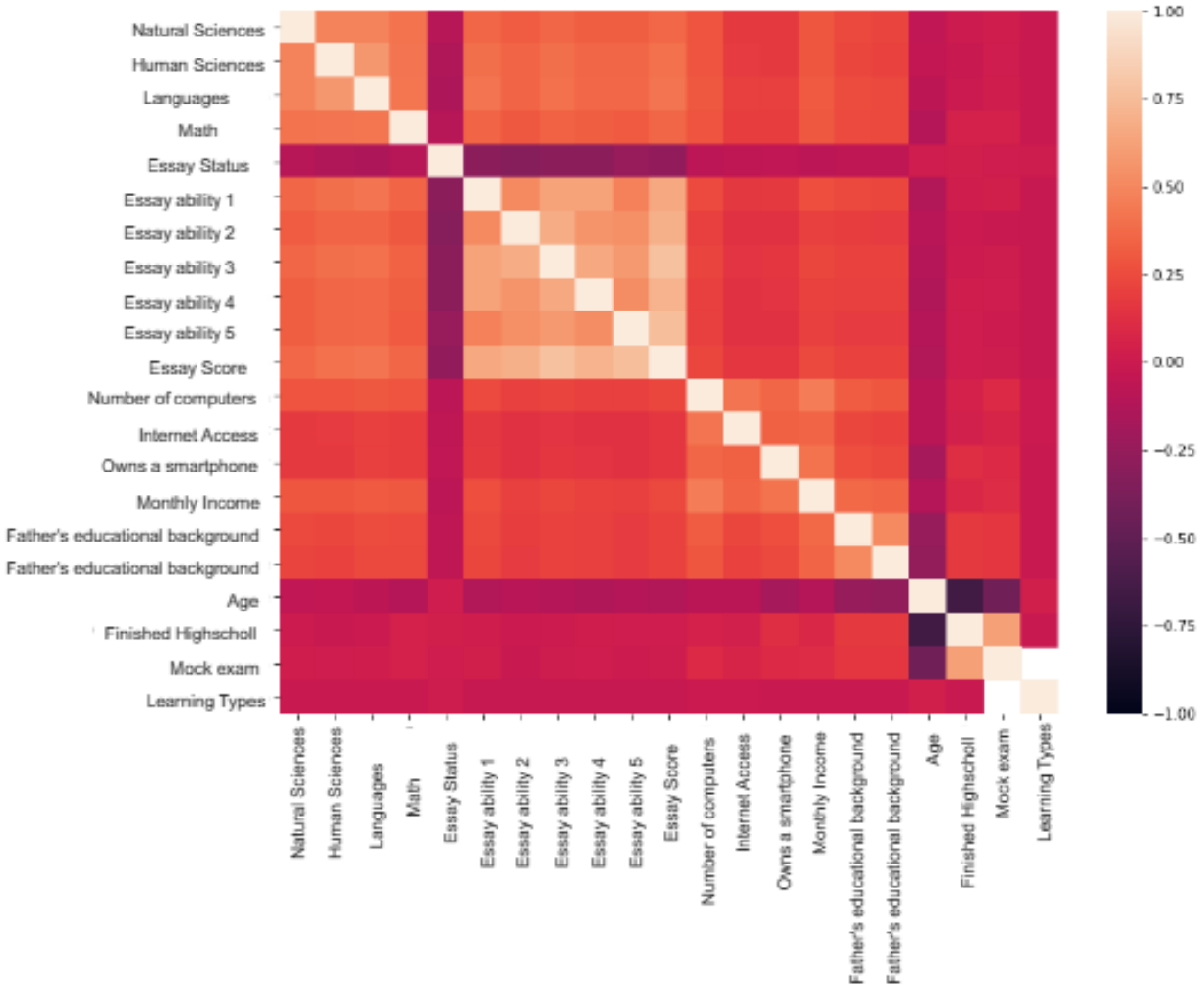



Diatribuição geral do status da redação

2 - Anulada - 5659 - Invalidated
3 - Cópia Texto Motivador - 23265 - Motivational/Copied Text
4 - Em Branco - 56901 - Blank
6- Fuga ao tema - 40623 - Avoiding the subject
7 - Não atendimento ao tipo textual - 3800 - Non understanding of writing model
8 - Texto insuficiente - 8578 - Insufficient text
9 - Parte desconectada - 4863 - Disconnected part / Discontinuity

According to the graph, after cases in which the essay is left blank, the most common errors committed by the candidates are, respectively, deviation from the theme and copying the motivating text. It should be noted that in both graphs, only the candidates who attended were considered..

**Variable correlation analysis:**



The main goal of this analysis is to try to discover relationships, even if subtle, between the candidate's performance and different variables provided by the dataset.
Three statistical correlation methods were used to construct a heatmap, in an attempt to locate latent correlations between parameters.
;The assessed methods were Pearson correlation, Kendall Tau, and Spearman correlation.
After consulting literature and documentation regarding the methods, it was found that the Kendall Tau correlation method is the most appropriate for studying associations between continuous and ordinal variables.

The variables analyzed were selected as those judged most relevant in terms of their influence on the candidate's performance. The interpretation of the graph is given by the Kendall Tau correlation values table. In other words, the closer the correlation value is to the extremes (1 for positive correlation, and -1 for negative correlation), the stronger the association between the variables. With the table in mind, we can more easily guide ourselves through the dataset in search of meaningful variables.

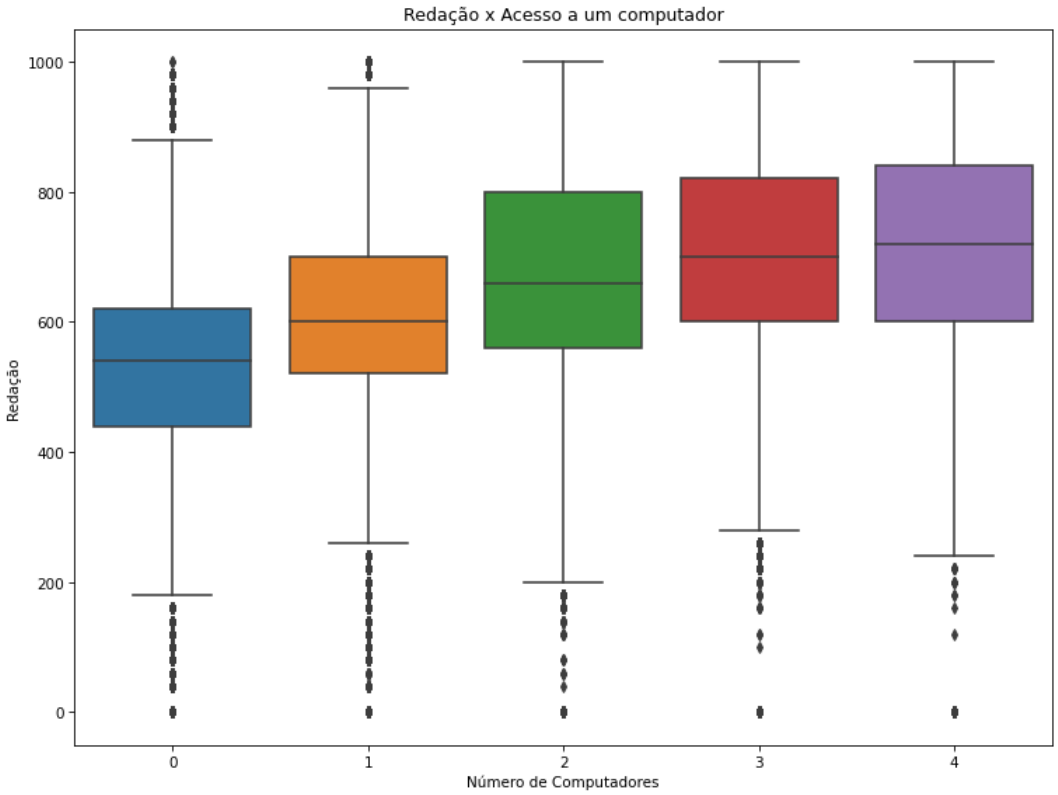
***Kendall's tau-B values:***

- Less than + or - 0.10: very weak
- + or -0.10 to 0.19: weak
- + or - 0.20 to 0.29: moderate
- + or - 0.30 or above: strong

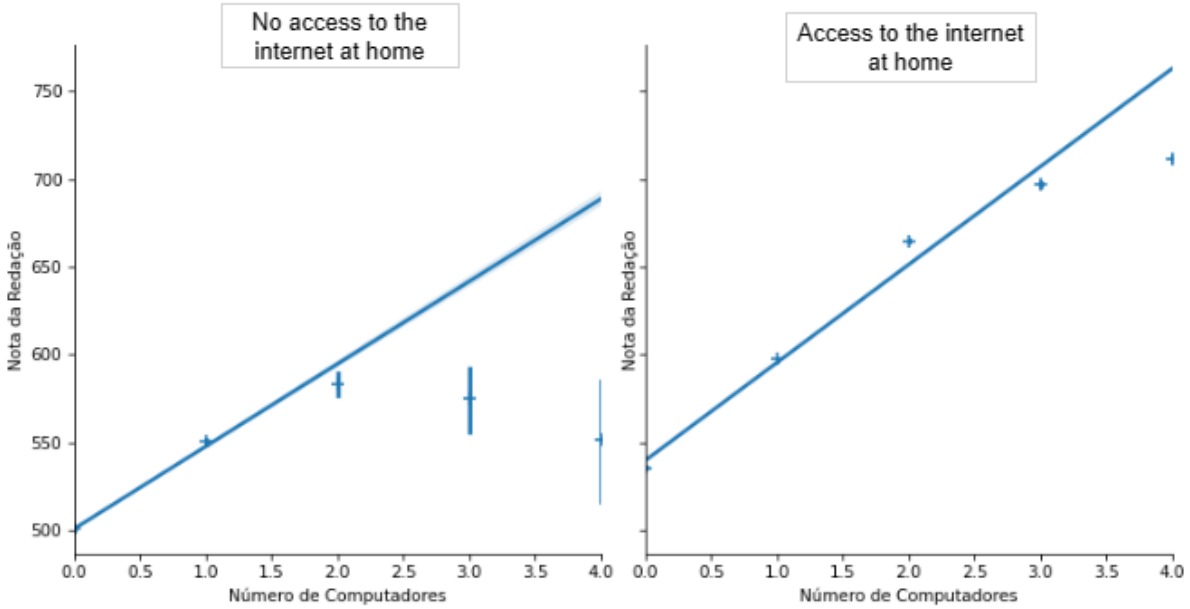
**Insights and inferences regarding test performance and its correlated variables from the heatmap.**

**Access to information and Essay Performance:**

Staying informed about both national and international news and events is essential for a good foundation in your writing, as it is common for the topic to involve global knowledge of current events which requires a certain level of access to information. This relationship can be seen through the correlation between the variables that indicate this (number of computers at home, cell phones, internet, etc.) and the writing assignment score.




**Boxplot Number of computers at home x Essay Score -> Kendall Tau = 0.229654**

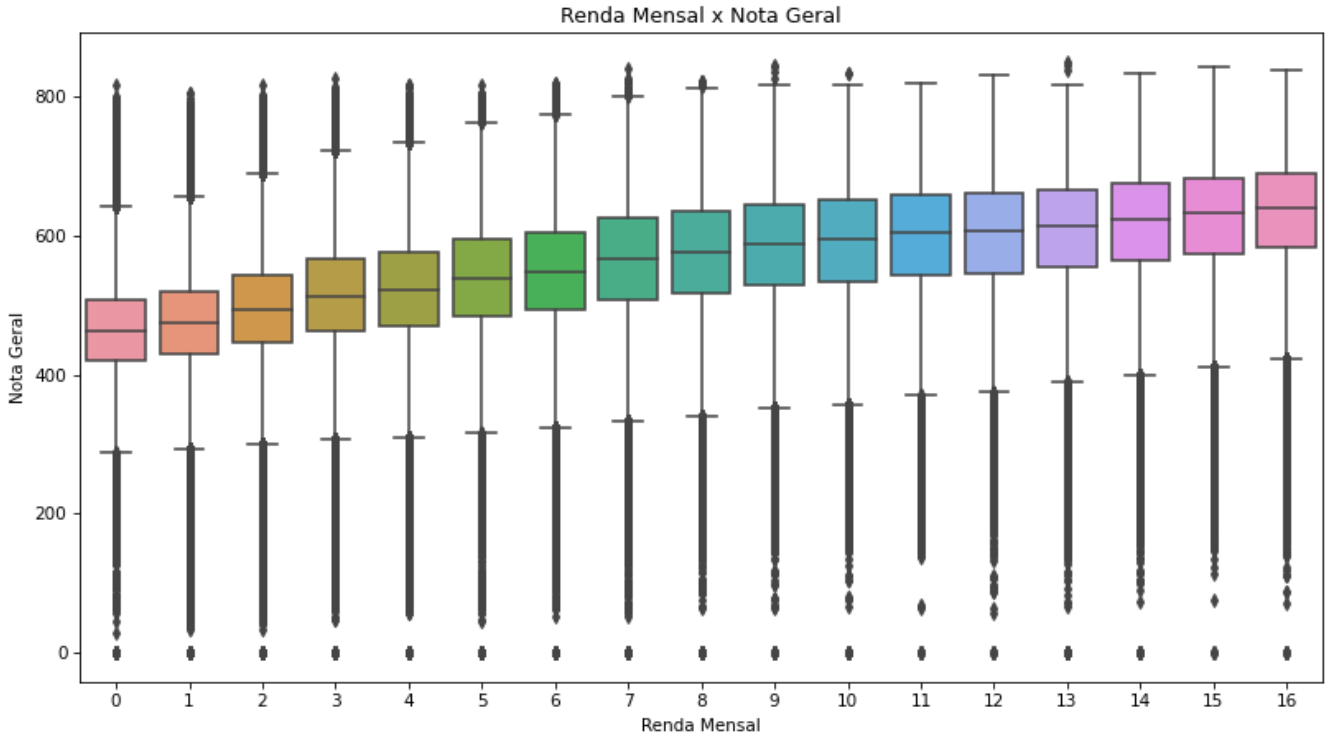



**Linear Regression -> Number of computers at home x Essay Score - > grouped by - access to internet**

The positively correlated relationship remains in both cases of internet access, but it is observed that candidates with access to the internet have a higher central tendency for their grades.

**Monthly Income e Overall Score:**

Starting from the same principle, monthly income was found to be strongly correlated (Tau values of 0.30 or higher) with performance in all subjects. Because parameters such as type



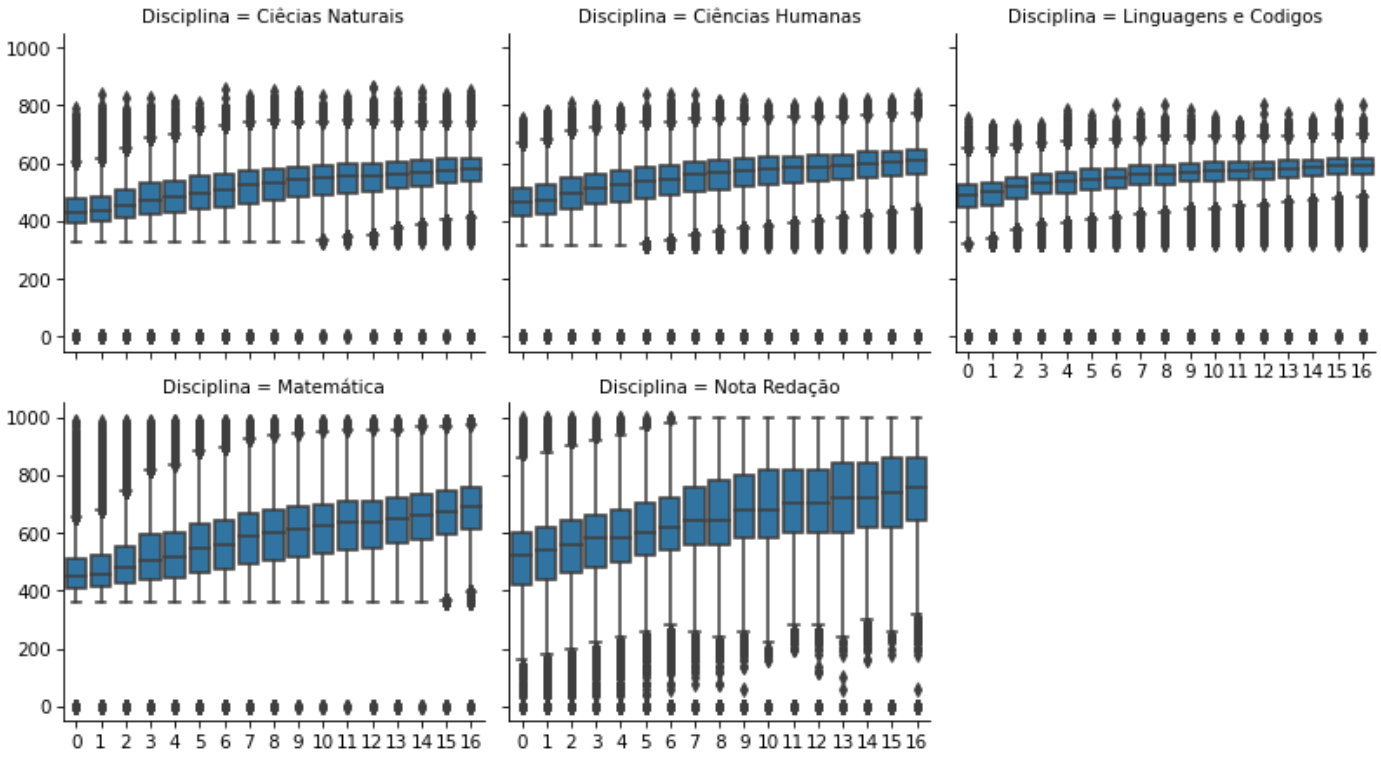

of education and access to information are intrinsically dependent on the family's monthly income.

**Boxplot Monthly Income x Overall Score -> Kendall Tau = 0.312418**

The relationship between the grades for each subject and the monthly income of the candidates' families is presented below. According to the provided dictionary, the monthly income is distributed ordinally and categorically in 16 levels.

| |
|---|
| Nenhuma renda. |
| Até R$ 998,00. |
| De R$ 998,01 até R$ 1.497,00. |
| De R$ 1.497,01 até R$ 1.996,00. |
| De R$ 1.996,01 até R$ 2.495,00. |
| De R$ 2.495,01 até R$ 2.994,00. |
| De R$ 2.994,01 até R$ 3.992,00. |
| De R$ 3.992,01 até R$ 4.990,00. |
| De R$ 4.990,01 até R$ 5.988,00. |
| De R$ 5.988,01 até R$ 6.986,00. |
| De R$ 6.986,01 até R$ 7.984,00. |
| De R$ 7.984,01 até R$ 8.982,00. |
| De R$ 8.982,01 até R$ 9.980,00. |
| De R$ 9.980,01 até R$ 11.976,00. |
| De R$ 11.976,01 até R$ 14.970,00. |
| De R$ 14.970,01 até R$ 19.960,00. |
| Mais de R$ 19.960,00. |

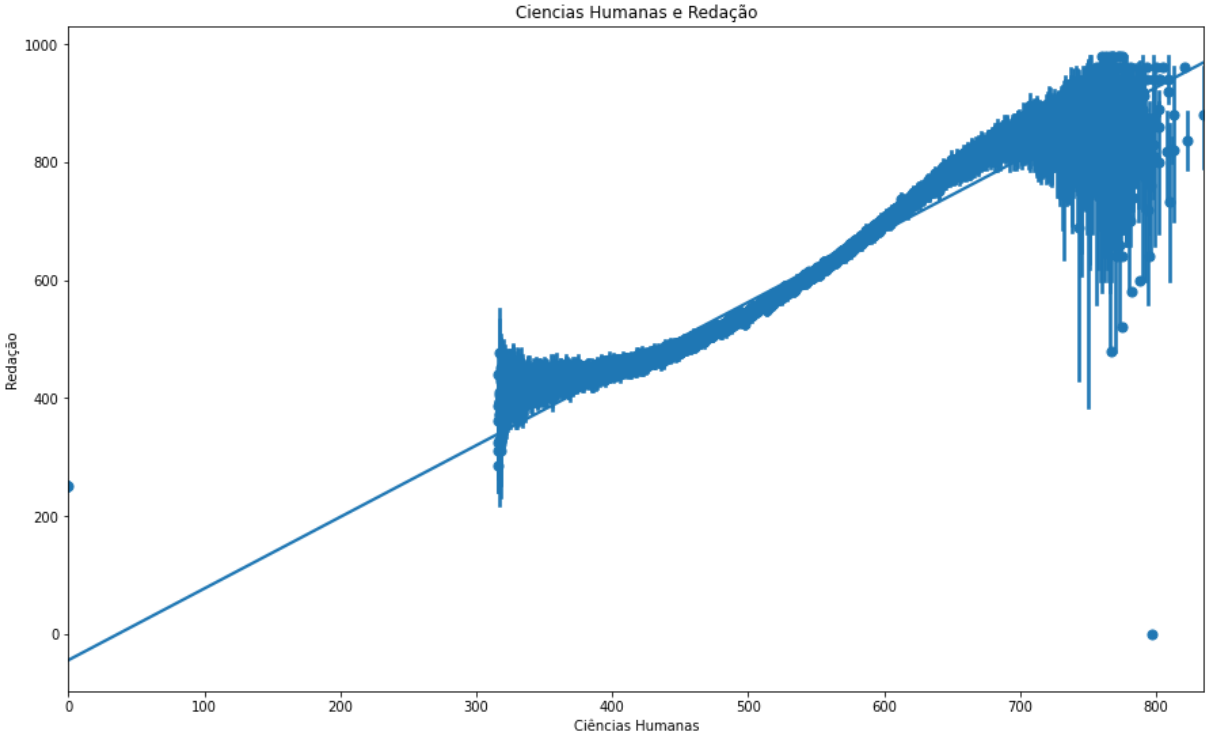
**Boxplot Monthly Income x Subject's score -> Kendall Tau = [0.293 0.291 0.309 0.303 0.238**



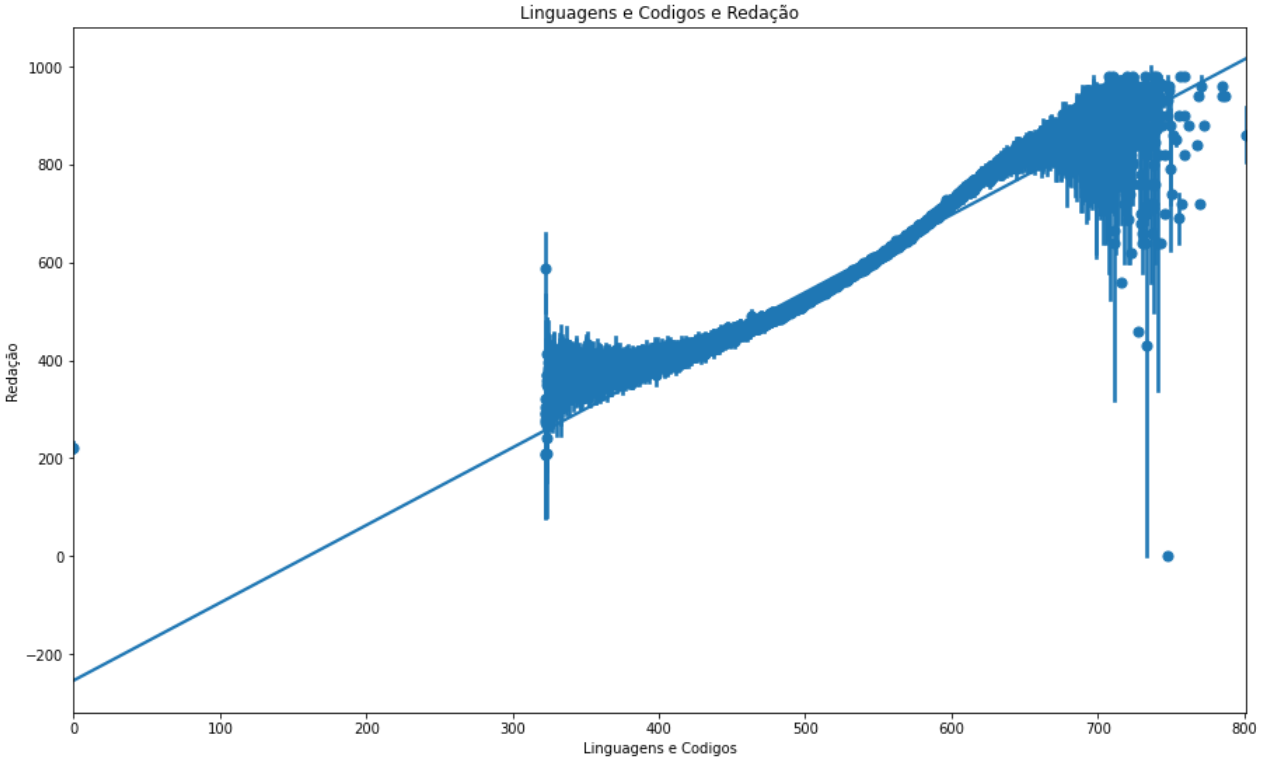

Analyzing both the graphs and correlation values, we can infer that mathematics is the subject most influenced by monthly income.

**Humanities, Languages and Grammar x Writing:**

Due to the nature of these subjects and the writing topics, it is expected that there is a correlation between these variables. We can observe that both Humanities and Languages exert a great influence on the candidates' writing performance, that is, statistically, candidates who excel in these subjects tend to do better in competencies such as mastery of the Portuguese language and application of historical concepts in their arguments.



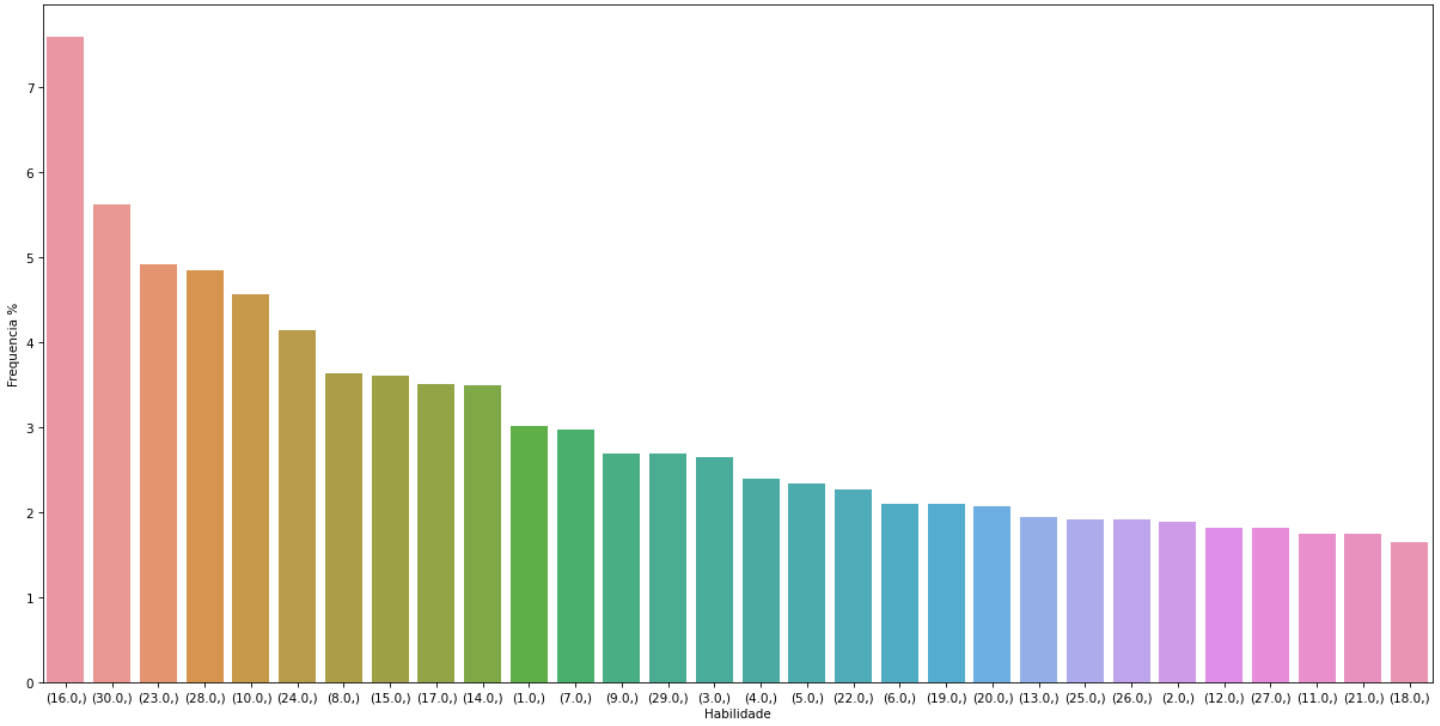

**Human Sciences and Essay Score regression plot, Kendall Tau = [0.402]**



**Languages and Essay Score regression plot, Kendall Tau = [0.415]**

**Answer key, competencies, and skills analysis:**

Through the same line of reasoning as the analysis above, it is interesting to try to find out which skills these subjects' (Humanities and Languages) candidates find the most difficulty. According to the ENEM reference matrix provided by INEP, each question from different subjects has a series of distinct competencies and skills that the candidate must master in order to answer them correctly.
Below is the Languages subject Analysis:




**Frequency of Error Distribution Based on Different Skills**

To develop the above distribution, it was necessary to iterate through all the questions, answer sheets, answer keys and candidate's codes in order to properly map the skills in which candidates were unsuccessful. As we can observe, skills numbered 16 and 30 were the most wrongly answered by the candidates, this may mean it contains a greater inherent score weight, or something related to the exam's model (harder questions corresponding to these skills in that specific year).

**H16 -** Relacionar informações sobre concepções artísticas e procedimentos de construção do texto literário.

**H16 - Ability to find the relationship between artistic concepts and literary work development.**

**H30 -** Relacionar as tecnologias de comunicação e informação ao desenvolvimento das sociedades e ao conhecimento que elas produzem.

**H16 - Ability to find the relationship between communication technologies, society development information and the knowledge it produces.**