UDACITY

DATA ANALYST NANODEGREE

GABRIEL SALVADOR BARÇANTE BARROS

PISA 2012 Exploration Analysis

Brazil

2021

1. Dataset

The data consists of information regarding 490,000 students who took part in the PISA 2012, including student performance in math, reading, and science. The dataset can be found in here, with feature documentation available here. The clean dataset used in the explanatory analysis is the **pisa2012_clean.csv**.

2. Summary of Findings

In the exploration, I found that all the plausible values, for each subject (math, reading, and science), have a normal distribution, with the peak between 400 and 600, approximately in 500. Moreover, I found that the plausible values have a strong positive relationships with each other, i.e., each subject has a strong positive correlation with each other.

The highest parental ISEI and the highest parental education in years have a weak to moderate positive correlation with the plausible values. On the other hand, the month of birth has a weak negative correlation with each plausible value. Meanwhile, the age variable has a weak positive relationship.

There are also some interesting relationships between the plausible values and the ordinal variables (cars, phones, televisions, computers, and books), where the median of the plausible values has a growing value in the ordinal categories. The only exception is with the variable cars, where the median is greater with students that have two cars at home, rather than with three or more.

Another relationship that prove to be interesting is between the plausible values and the countries. The same countries remain with high averages for all the three subjects (math, reading, and science). Moreover, in the world maps plotted in the exploratory analysis, a preponderance of dark colors, which indicates high values, were seen in Europe, Oceania (Australia and New Zealand), North America (EUA and Canada), China, Japan and South Korea.

Finally, I investigated how the relationship between the average of the plausible values and the highest parental ISEI / highest parental education in years, is influenced by the number of computers or books at home. Regardless of the increase of the two numeric variables, the average of the plausible values is lower for students who don't have any computer or have fewer than 10 books at home.

3. Key Insights for Presentation

For the presentation, I focus on features that could help me see if there is inequality in academic achievement, i.e., if environment variables could influence the students' performance.

So, I explore on how the parents' situation (highest parental ISEI and highest parental education in years) influence their children performance in the PISA 2012. In addition, I analyze which countries have the highest averages in all three subjects. It isn't a surprise that these are in rich countries, like in Europe, Oceania (Australia and New Zealand), North America (EUA and Canada), China, Japan and South Korea.

Finally, I introduce a categorical variable: number of computers at home. I analyze the influence that this variable introduce at the relationship between the average of the plausible values by subject and each of following numeric variables: the highest parental ISEI and the highest parental education in years.

4. References

- PISA Data Analysis Manual: SPSS® SECOND EDITION
- PISA 2012 Technical Report
- Glossary Of Statistical Terms - OECD
- Geopandas: Mapping and Plotting Tools
- Data Visualization: How To Plot A Map with Geopandas in Python?
- How to iterate over rows in a DataFrame in Pandas?