Trang Tran, Anmol Kaur

Analysis of High School Student Performance & Demographics

Description of Data

The data is chosen from Kaggle, and the set contains the student achievement for two Portuguese high schools. It collected the data through school reports, questionnaires, which included student grades, demographics, social, parents, and school-related features. The two dataset that were given are divided into the subjects Mathematics and Portuguese. The math dataset has 29 columns and 349 entries while the Portuguese set has 423 rows but some number of columns. The columns are a mix between integers and strings.

Cleaning the Data

1. Removed columns in both datasets that were not necessary for our analysis.
2. Rename columns to understand the data better for us.
3. Sort the datasets based on age ascending.
4. Removed all the rows where students are going to MS (school), to focus only on one school.
5. Modified to new csv files.
6. Merged dataset into one based on the same student id and added Portuguese grades to the end of the columns.
7. Renamed the grade columns to divide math and portuguese grades.

Analysis and Visualization

- 250 students missed class at least once and 69 students while 47 students missed class once and failed a class
- Highest Math score was 20 while 19 was in Portuguese, the student with the highest math score does not have the highest portuguese score
- Correlation between Absences and Final Math score is 0.04

Challenges, Research and Solutions

The team faced challenges integrating the Kaggle API in Google Colab. Issues included errors with API credentials, dataset path specification, account permissions, and API updates affecting the code. Key steps for success involve uploading the Kaggle API key, ensuring its JSON format, setting permissions, and maintaining dataset path accuracy. Users should also have the necessary Kaggle account permissions, a stable internet connection, and stay informed about API updates.

Research Expansion in the Future

For future research, we will consider advanced data cleaning and enhancing user-friendly visualizations. Investigate collaborative analysis methods and integrate external APIs. Prioritize reproducibility through improved documentation and enrich the dataset with additional external sources.