Analysis of High School Student Performance & Demographics

CIS 9650 Group Project

By: Anmol Kaur and Trang Tran Phuong

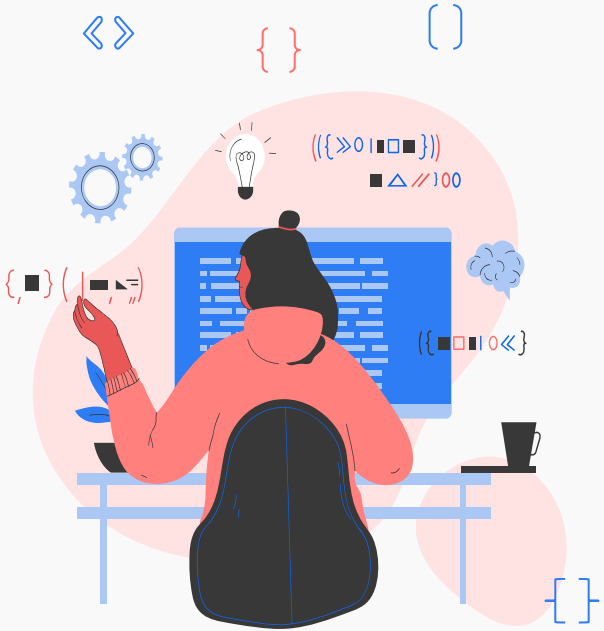
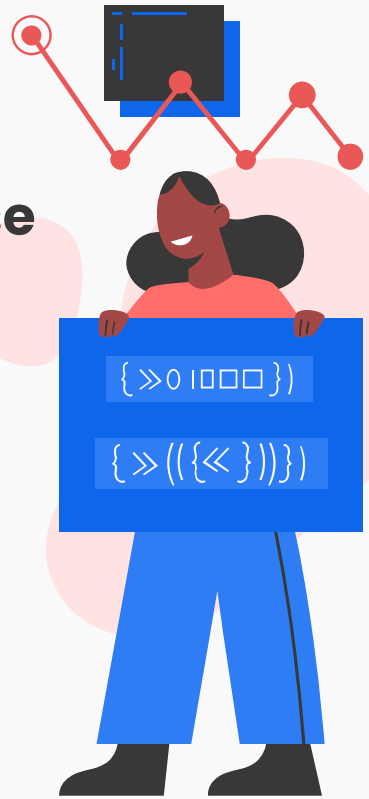


Table of contents



Introduction | Data Description

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 same number of columns. The columns are a mix between integers and strings.

Data Import

```
[1] !pip install -q kaggle
```

```
[2] from google.colab import files
    files.upload()
```

Choose Files kaggle.json

- **kaggle.json**(application/json) - 68 bytes, last modified: 12/8/2023 - 100% done

```
Saving kaggle.json to kaggle.json
{'kaggle.json': b'{"username":"anmolkaur155","key":"b34917a05d40a982e6ce0463bac1721e"}'}
```

```
[3] ! mkdir ~/.kaggle
```

```
[5] ! cp kaggle.json ~/.kaggle/
```

```
[4] ! chmod 600 ~/.kaggle/kaggle.json
```

```
chmod: cannot access '/root/.kaggle/kaggle.json': No such file or directory
```

```
[6] ! kaggle datasets list
```

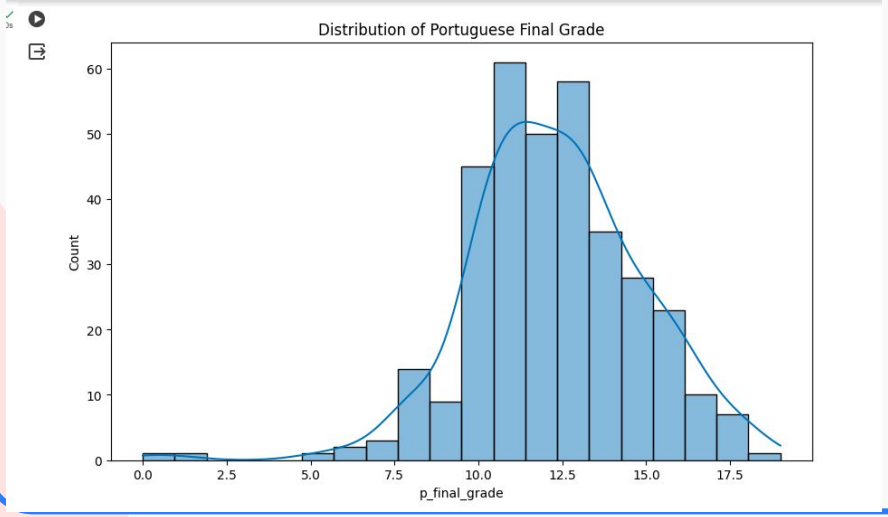
```
Warning: Your Kaggle API key is readable by other users on this system! To fix this, you can run 'chmod 600 /root/.kaggle/kaggle.js
ref                                                            title                                              size  last
-------------------------------------------------------------  -------------------------------------------------  -----  ----
thedrcat/daigt-v2-train-dataset                                DAIGT V2 Train Dataset                              29MB  2023
muhammadbinimran/housing-price-prediction-data                 Housing Price Prediction Data                      763KB  2023
thedevastator/spotify-tracks-genre-dataset                     Spotify Tracks Genre                                 8MB  2023
```

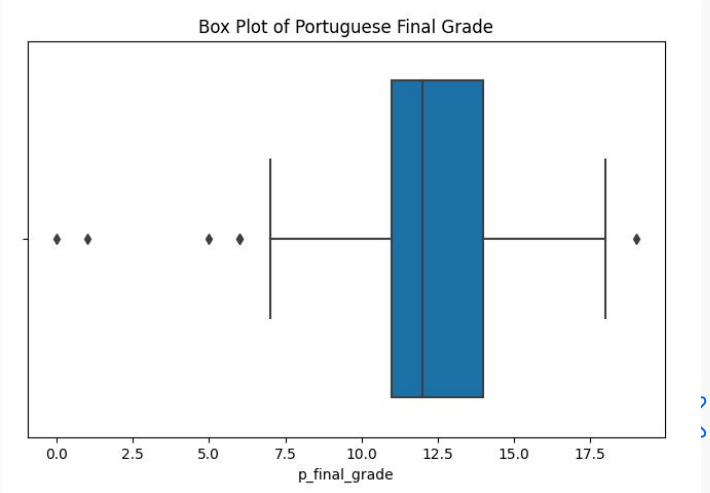
Cleansing the Data

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

Analysis and Visualizations

- 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

Challenges:

- Integration of Kaggle API in Google Colab
- Dataset path specification issues
- API updates impacting code functionality

Key Steps for Success:

- Upload Kaggle API key
- Verify JSON format
- Set correct permissions
- Ensure accurate dataset path

User Recommendations:

- Ensure a stable internet connection
- Stay informed about Kaggle API updates.

Research Expansion in the Future

Future Research Goals:

- Advanced data cleaning techniques
- Enhance user-friendly visualizations
- Investigate collaborative analysis methods
- Integrate external APIs

Prioritizing Reproducibility:

- Improve documentation for enhanced reproducibility

Enriching the Dataset:

- Consider incorporating additional data from external sources

Thanks for Listening!