CAPSTONE PROJECT 1: MILESTONE REPORT

MINH H. LE

1. Problem statement

Education is the passport to the future. My goal is to help educational institutions, and parents to improve the student academic performance. The problems that I want to solve from this project is to find out which attributes, including student grades, demographic, social and school related features, have an effect on high school student performance and from those effective attributes, predicting the student performance. Based on my analysis, schools and parents will know which factors and reasons affect the student performance. With that knowledge, they can help the students to increase their performance, preparing them for the future.

2. Description of the dataset

The data that I am using was collected by using school reports and questionnaires. The data include two datasets (CSV files), one is performance in mathematics and the other is performance in Portuguese. The data are acquired from the UCI Machine Learning Repository (URL: https://archive.ics.uci.edu/ml/datasets/Student+Performance#).

The first step is to find unique values of each attribute (feature) for both datasets (Mathematics and Portuguese) to figure out whether it has missing values or not. There are 32 attributes in both datasets. Most of the features are categorical (nominal and ordinal) data, except absences (number of school absences), G1 (first period grade), G2 (second period grade), and G3 (final grade, also the output target).

In [5]:
```
# Looking for missing data (Mathematics)
for i in mat.columns:
    print(i,mat[i].unique())
```

```
school ['GP' 'MS']
sex ['F' 'M']
age [18 17 15 16 19 22 20 21]
address ['U' 'R']
famsize ['GT3' 'LE3']
Pstatus ['A' 'T']
Medu [4 1 3 2 0]
Fedu [4 1 2 3 0]
Mjob ['at_home' 'health' 'other' 'services' 'teacher']
Fjob ['teacher' 'other' 'services' 'health' 'at_home']
reason ['course' 'other' 'home' 'reputation']
guardian ['mother' 'father' 'other']
traveltime [2 1 3 4]
studytime [2 3 1 4]
failures [0 3 2 1]
schoolsup ['yes' 'no']
famsup ['no' 'yes']
paid ['no' 'yes']
activities ['no' 'yes']
nursery ['yes' 'no']
higher ['yes' 'no']
internet ['no' 'yes']
romantic ['no' 'yes']
famrel [4 5 3 1 2]
freetime [3 2 4 1 5]
goout [4 3 2 1 5]
Dalc [1 2 5 3 4]
Walc [1 3 2 4 5]
health [3 5 1 2 4]
absences [ 6  4 10  2  0 16 14  7  8 25 12 54 18 26 20 56 24 28  5 13 15 22  3 21
  1 75 30 19  9 11 38 40 23 17]
G1 [ 5  7 15  6 12 16 14 10 13  8 11  9 17 19 18  4  3]
G2 [ 6  5  8 14 10 15 12 18 16 13  9 11  7 19 17  4  0]
G3 [ 6 10 15 11 19  9 12 14 16  5  8 17 18 13 20  7  0  4]
```

Figure 1: Mathematics dataset

In [6]:
```
# Looking for missing data (Portuguese)
for i in por.columns:
    print(i,mat[i].unique())
```

```
school ['GP' 'MS']
sex ['F' 'M']
age [18 17 15 16 19 22 20 21]
address ['U' 'R']
famsize ['GT3' 'LE3']
Pstatus ['A' 'T']
Medu [4 1 3 2 0]
Fedu [4 1 2 3 0]
Mjob ['at_home' 'health' 'other' 'services' 'teacher']
Fjob ['teacher' 'other' 'services' 'health' 'at_home']
reason ['course' 'other' 'home' 'reputation']
guardian ['mother' 'father' 'other']
traveltime [2 1 3 4]
studytime [2 3 1 4]
failures [0 3 2 1]
schoolsup ['yes' 'no']
famsup ['no' 'yes']
paid ['no' 'yes']
activities ['no' 'yes']
nursery ['yes' 'no']
higher ['yes' 'no']
internet ['no' 'yes']
romantic ['no' 'yes']
famrel [4 5 3 1 2]
freetime [3 2 4 1 5]
goout [4 3 2 1 5]
Dalc [1 2 5 3 4]
Walc [1 3 2 4 5]
health [3 5 1 2 4]
absences [ 6  4 10  2  0 16 14  7  8 25 12 54 18 26 20 56 24 28  5 13 15 22  3 21
  1 75 30 19  9 11 38 40 23 17]
G1 [ 5  7 15  6 12 16 14 10 13  8 11  9 17 19 18  4  3]
G2 [ 6  5  8 14 10 15 12 18 16 13  9 11  7 19 17  4  0]
G3 [ 6 10 15 11 19  9 12 14 16  5  8 17 18 13 20  7  0  4]
```

Figure 2: Portuguese dataset

The next step is to find outlier values of the non-categorical features for both datasets using the z-score method. From the empirical rule, around 99.7% of the z-score will be within -3 and 3. The outlier will have the z-score value greater than 3 or less than -3. I calculate the z-score of each value in each feature and take the absolute value of it. If the value is greater than 3, that value is an outlier. Then I compare the outliers with the range of values of the non-categorical features provided in the dataset description (from the website) to check whether the outliers are due to incorrectly entered or measuring data or not. After close examination, there is no missing value in both datasets. However, there were outliers in both datasets. Based on information from the dataset description, since none of the outliers are due to incorrectly entered or measuring data, the outliers are kept in the datasets.

3. Initial findings from exploratory analysis

The goal of my project is to figure out which attributes, including student grades, demographic, social and school-related features, have an effect on high school student performance and from those attributes, predict the student performance (final grade). To see correlations between pairs of independent variables and between an independent and a dependent variable, I use correlation matrices using Pearson parametric correlation test. For both datasets, the correlation matrices show that there is a moderate positive relationship between mother's education and father's education, Pearson correlation coefficients are around 0.6, and between workday alcohol consumption and weekend alcohol consumption of students, Pearson correlation coefficients are around 0.6. However, the correlation matrices also show that there is a strong positive linear relationship between first-term grade and second-term grade, and between the term grade (both terms) and final grade. For both datasets, the relationship between the second term grade and final grade is slightly stronger than between the first term grade and final grade.

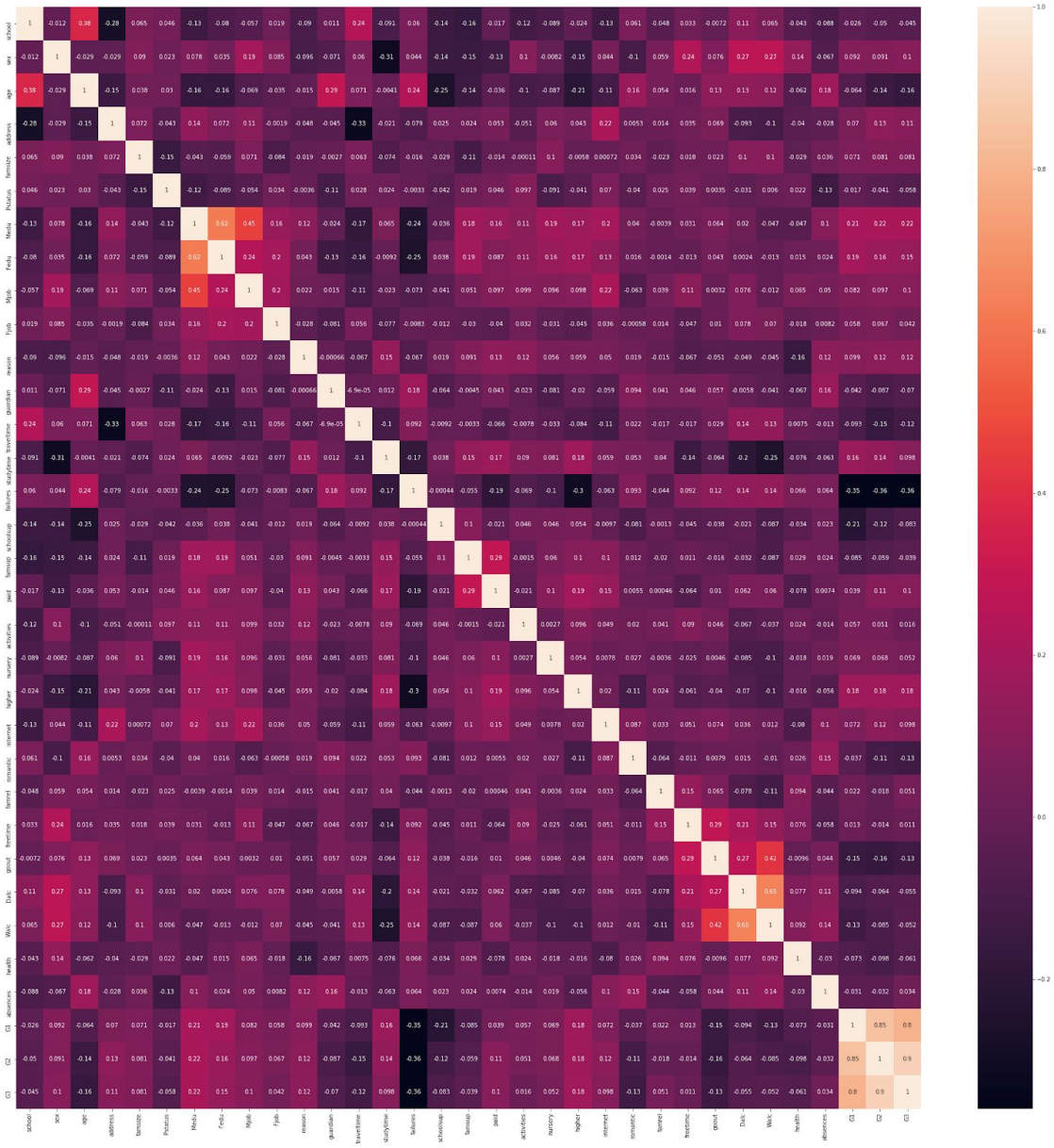

Figure 3: Correlation matrix heatmap - Mathematics

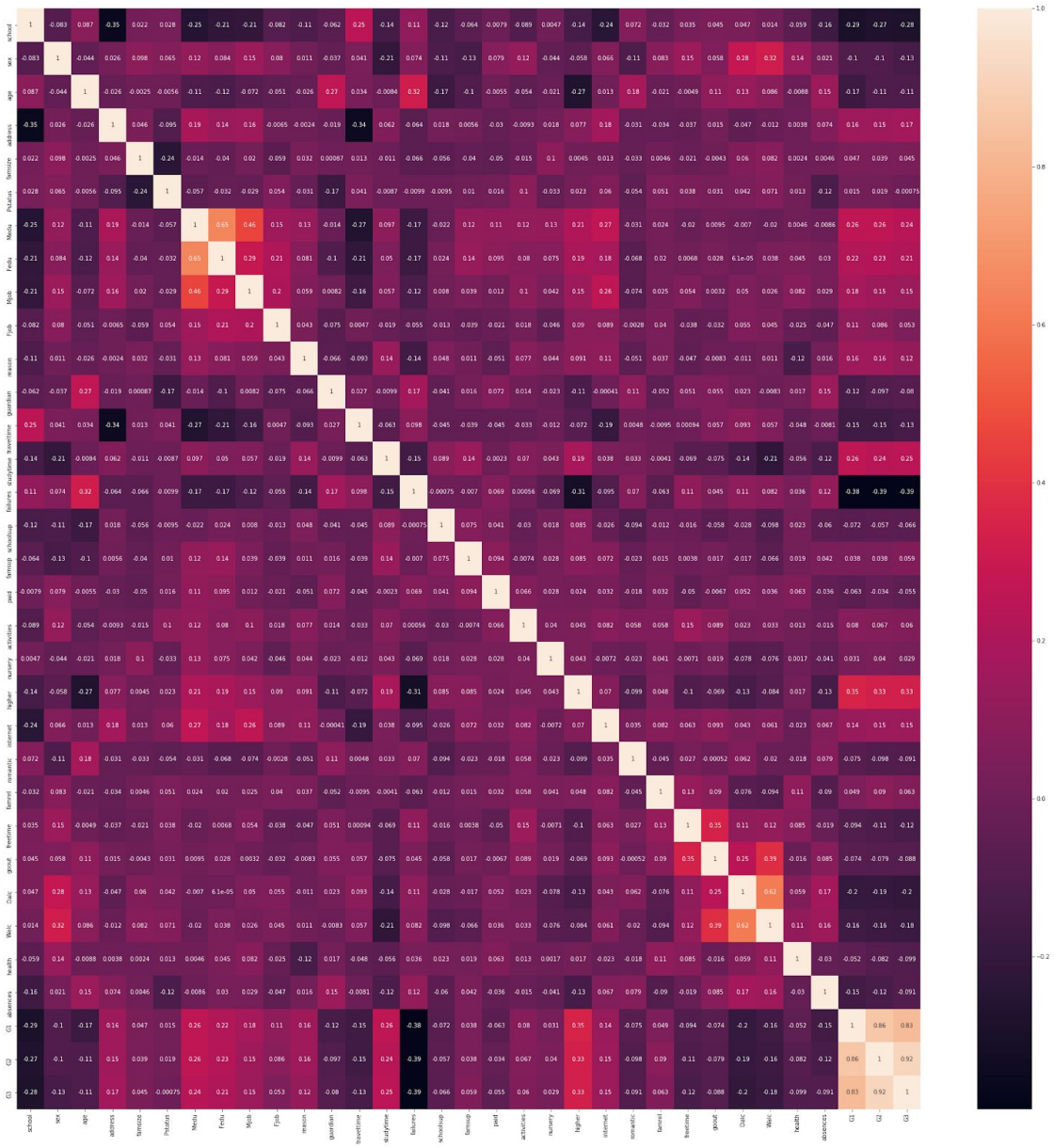

Figure 4: Correlation matrix heatmap - Mathematics

To see the relationship of the final grade between subgroups of each attribute, I compare the mean of the final grade of each subgroup in the same attribute to each other. To test the significant differences of the mean, I use one-tailed t-test for the means of two independent samples with the significance level (alpha) equals 0.5 and with the assumption that the population means from the two groups are not equal. For gender, the average mathematics final grades of male students are higher than

those of female students. In contrast, the average Portuguese final grades of female students are higher than those of male students. For wanting higher education, the average final grades in both mathematics and Portuguese of students who want higher education are higher than those of students who do not want a higher education. For extra educational support, the average final grades in both mathematics and Portuguese of students who have extra educational support are lower than those who do not have. For time spending on studying mathematics, surprisingly, there is no difference in the average final grade of students who spend a lot of time studying (more than 10 hours) and students who don’t spend much (less than 2 hours). However, for Portuguese, students who spend more time studying (more than 10 hours) have higher average final grades than those who don't (less than 2 hours). For health, surprisingly, students who are the least healthy (1) have higher average final grade final grades in both subjects than students who are the healthiest (5). This can be because healthy students are more likely to do more extracurricular activities than those who are not healthy. Therefore, they will spend less time studying, which can affect their final grades.