Table of content

Introduction

Exploratory Data Analysis (EDA)

Data Visualisation

Comparing features

Data Preparing

Data Modelling

# Introduction

In the field of education, predicting student performance is very important for educators.

In this way, based on a certain amount of data, students can increase their ability to learn and increase their learning opportunities. Educators can also improve teaching methods, identify problems, and plan ahead based on student data.

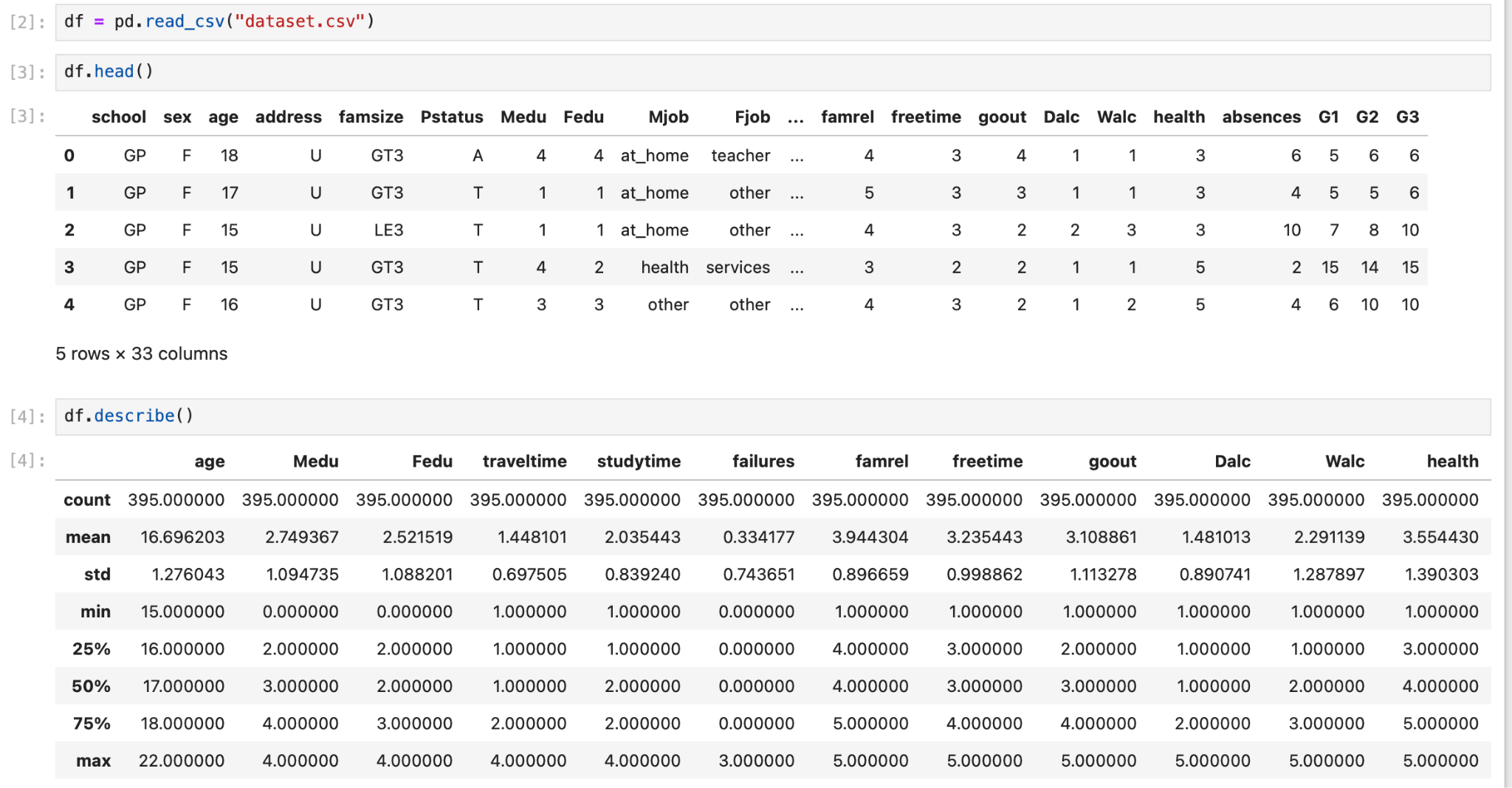
This data contains specific information, starting with the age and gender of the students, and each row represents one student. This data contains specific information, starting with the age and gender of the students, and each row represents one student. This data will be used to examine the factors that influence student enrollment in graduate school.

# Exploratory Data Analysis (EDA)



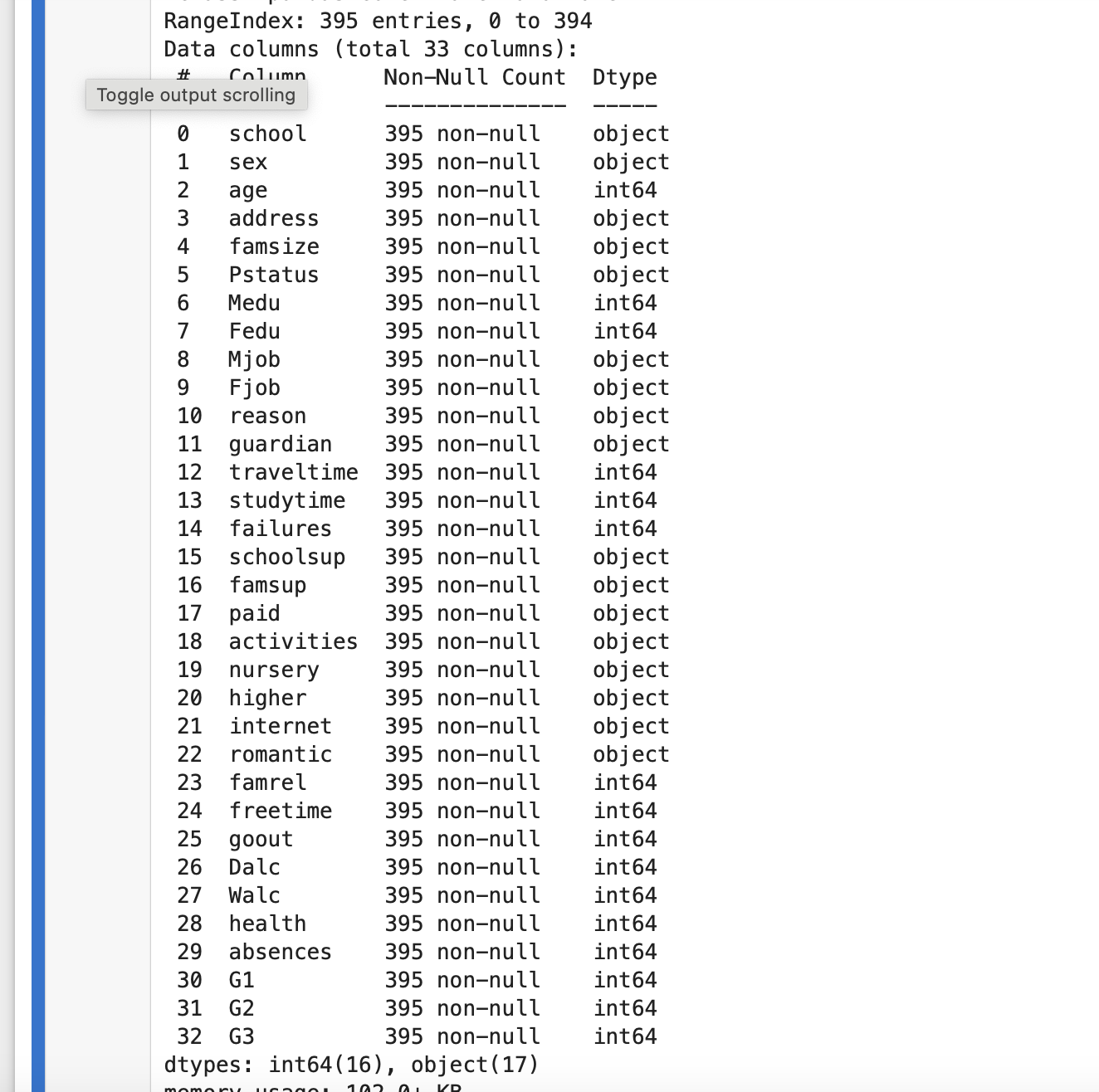
*Figure 1. Important libraries*

First of all, to do a proper job we need to import all the necessary libraries. In this case I import a few machine learning modeling libraries and simple libraries. So that I can visualize my dataframe.



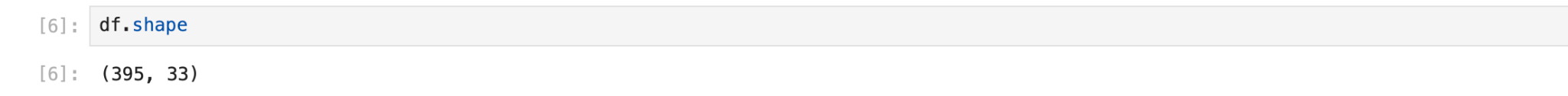
*Figure 2. Head of the Dataset and describe*

By using the head function I am able to quickly overview my dataset. In this case my dataset seems as it has a lot of categorical columns and it has 33 columns.

And the description function showed me information about numerical column statistics.

*Figure 3 . Dataset info*

df.info method gives me information about columns, data type,non-null value which is helpful in finding missing values and inspection. In my dataset there are 32 columns and 394 rows and as i seeing there is a lot of object data type so it need to be changed to visualize the data.



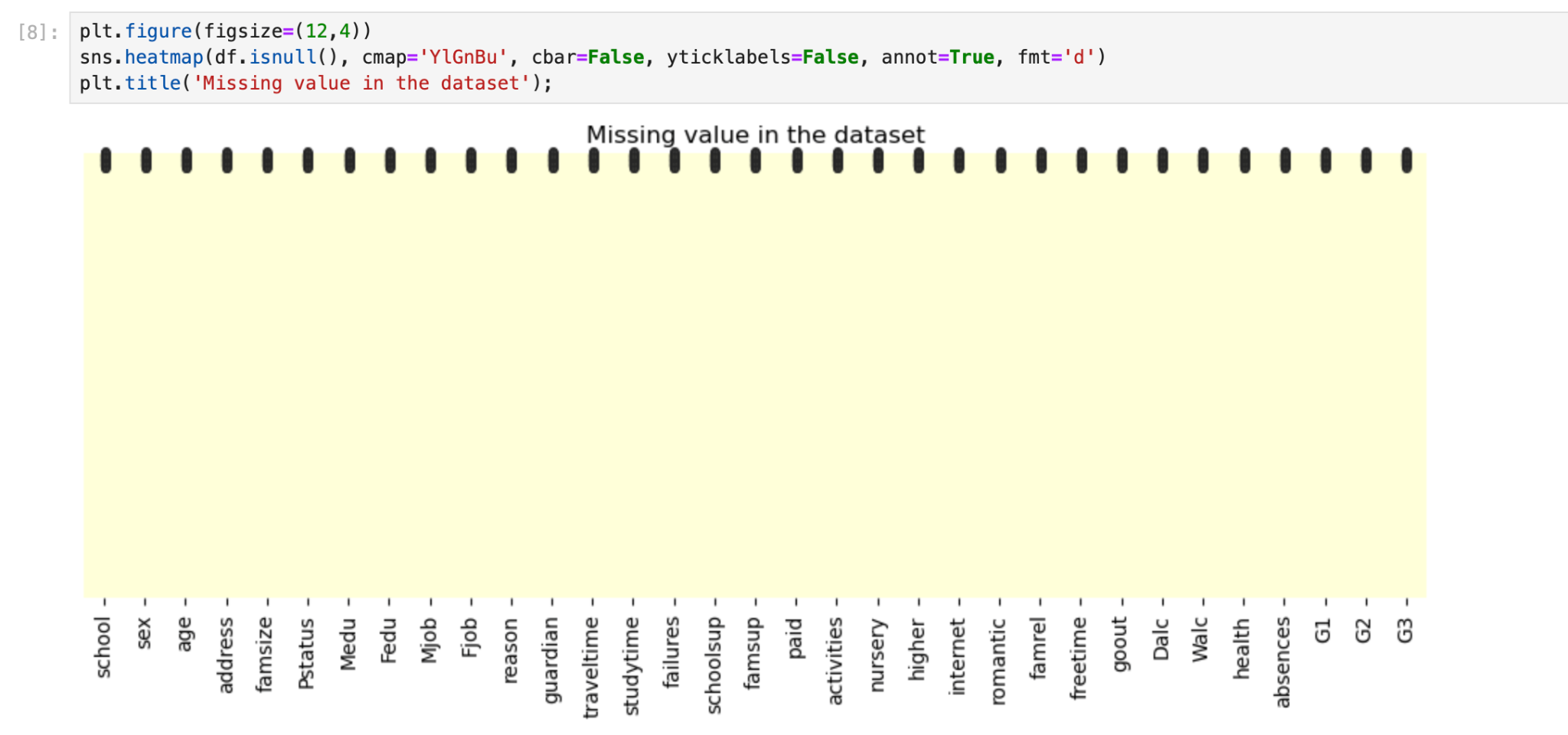
*Figure 4 . Dataset shape*

This method is very helpful to see my dataset structure. It showed me my dataset has 33 columns and 395 rows. But this is my raw datasets structure so the next step will be to check missing values.



*Figure 5 . Checking missing value*

By using df.isnull().sum() I checked my dataframe to see if it has a missing value or not and this is one of the simple steps in EDA. Usually there is always some missing value in the datasets I worked on before but this time there is no missing value. If there was a missing value I would run dropna or fillna method. But before running, machine learning model need to check unique values as well such as NaN.



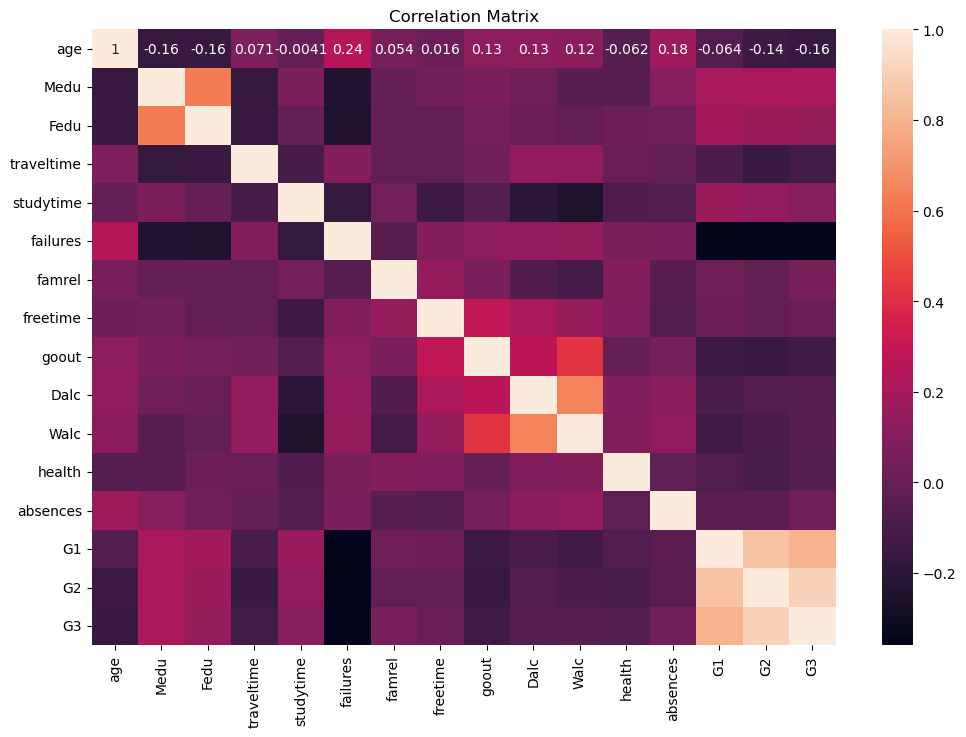
*Figure 6. Visualize missing value*

This method is very helpful in seeing the missing values in the dataset. In this dataset there is no missing value so visualization looks clean.

# Data visualization

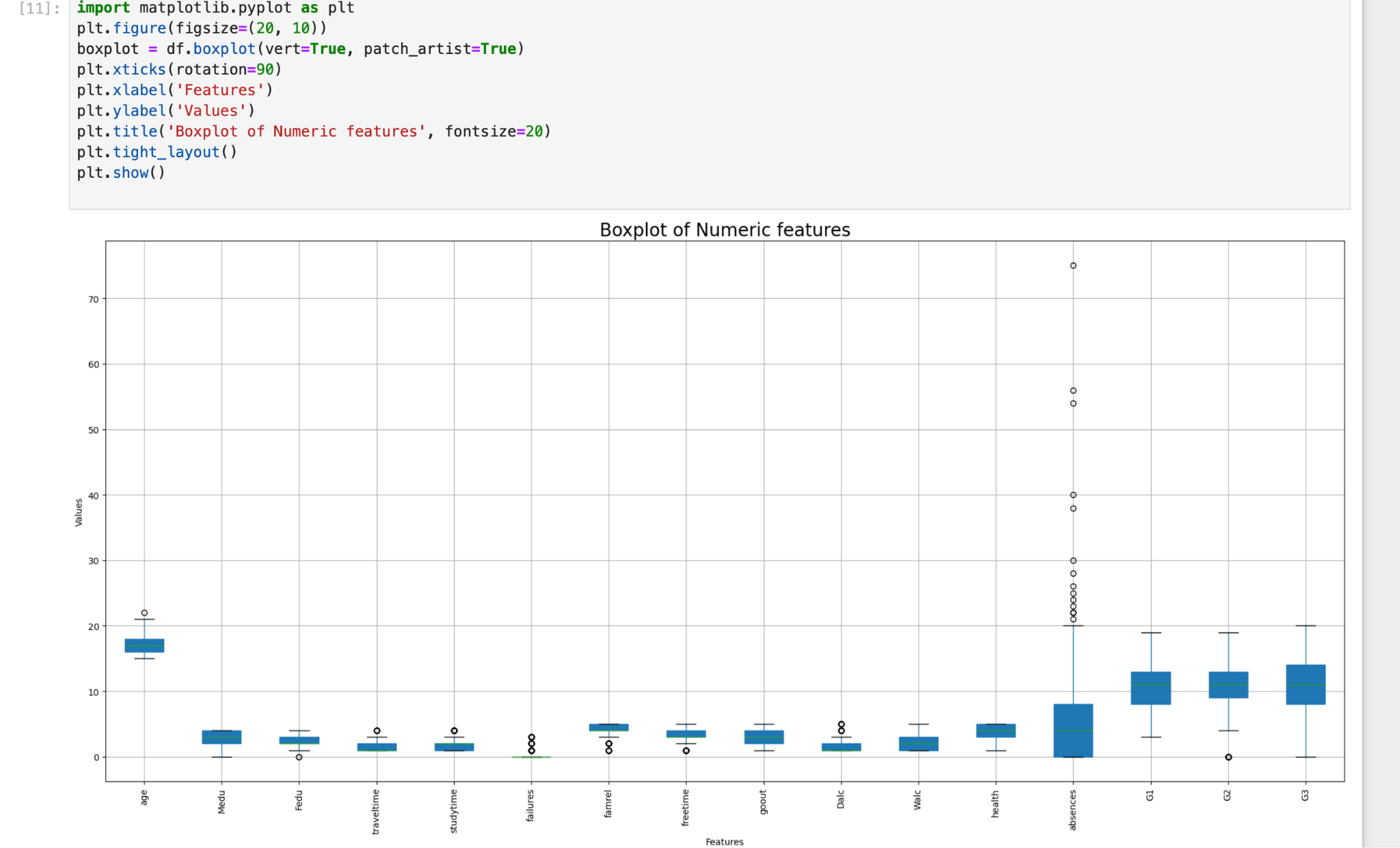
Data visualization allow me to identify patterns, trends and relationships between the variables and its helpful to understanding the structure of dataset.

In this project I have done correlation matrix and boxplot of numeric features. I found some correlation between the features and some outliers from the columns.



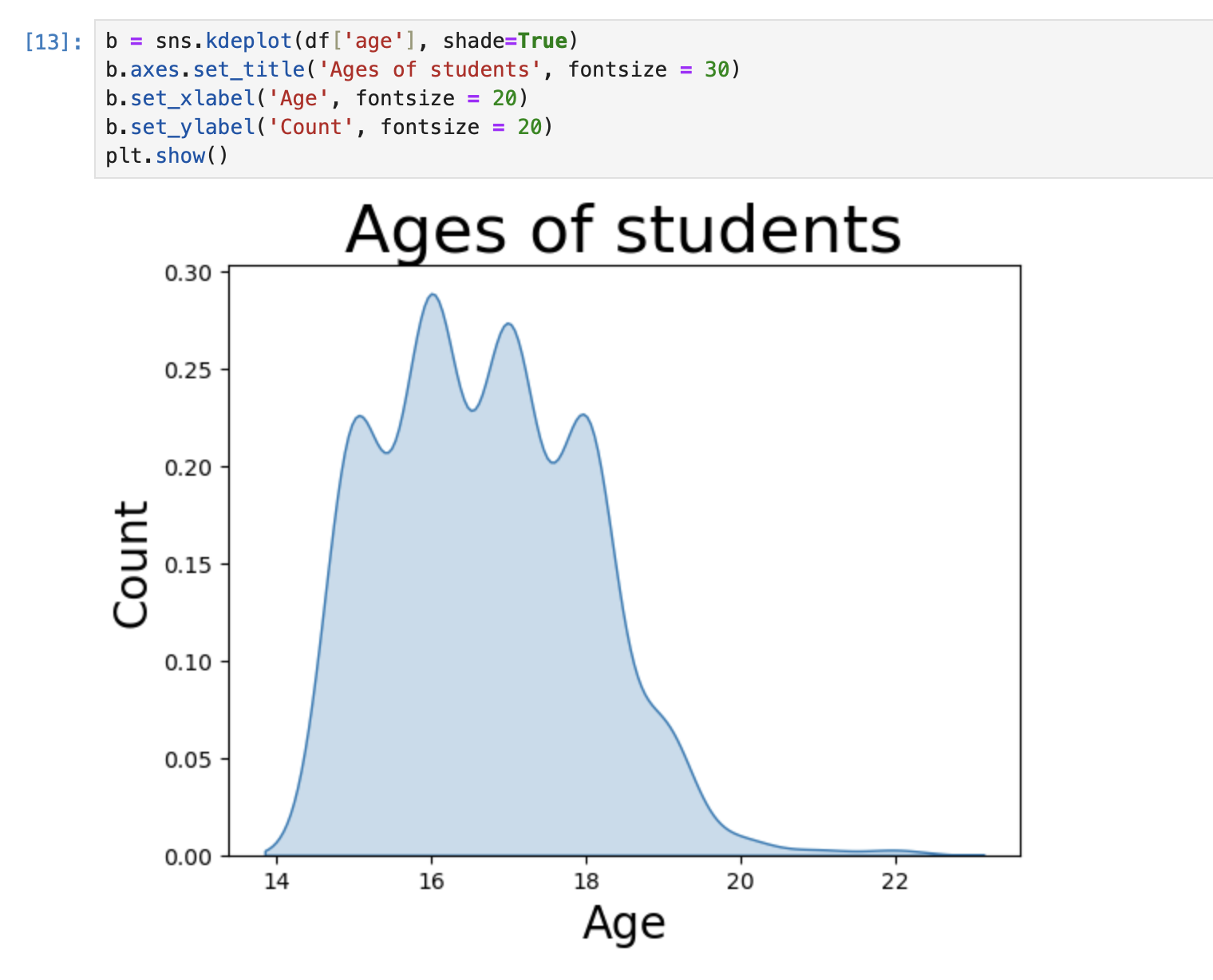
*Figure 7. Correlation matrix*

It tells me how closely related different things are in a dataset. With this visualization I can understand which features have strong connections. In this dataset parents' education and study time has strong connection with higher grades. Interesting part is that traveling time has a connection with failures.



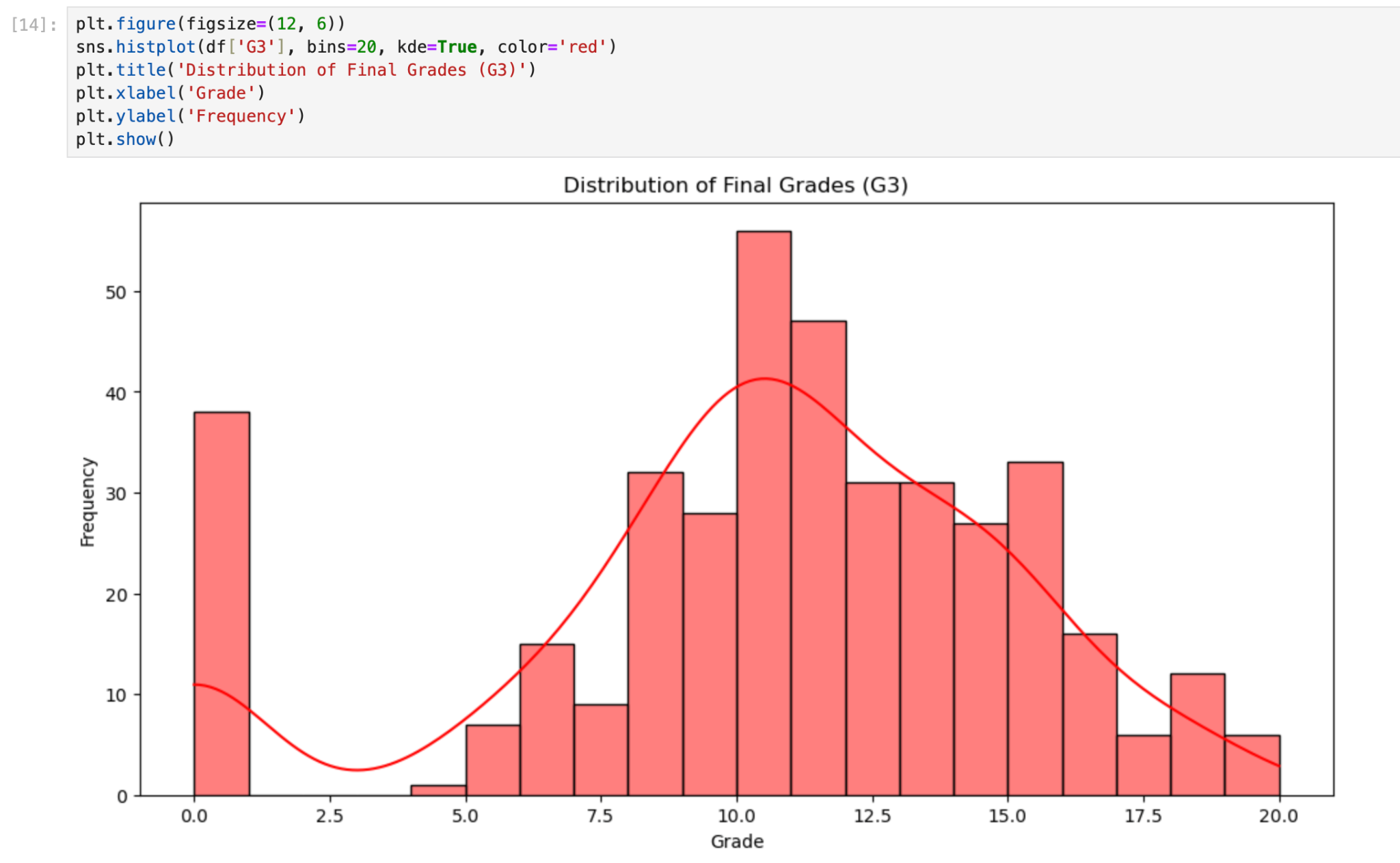
*Figure 8. Boxplot of Numeric features*

I checked my dataset with a box plot to identify its distribution, skewness and outliers. Here I found outliers from the column of absences.



*Figure 9. Age of students*

Most students in dataset is 16-18 years old.

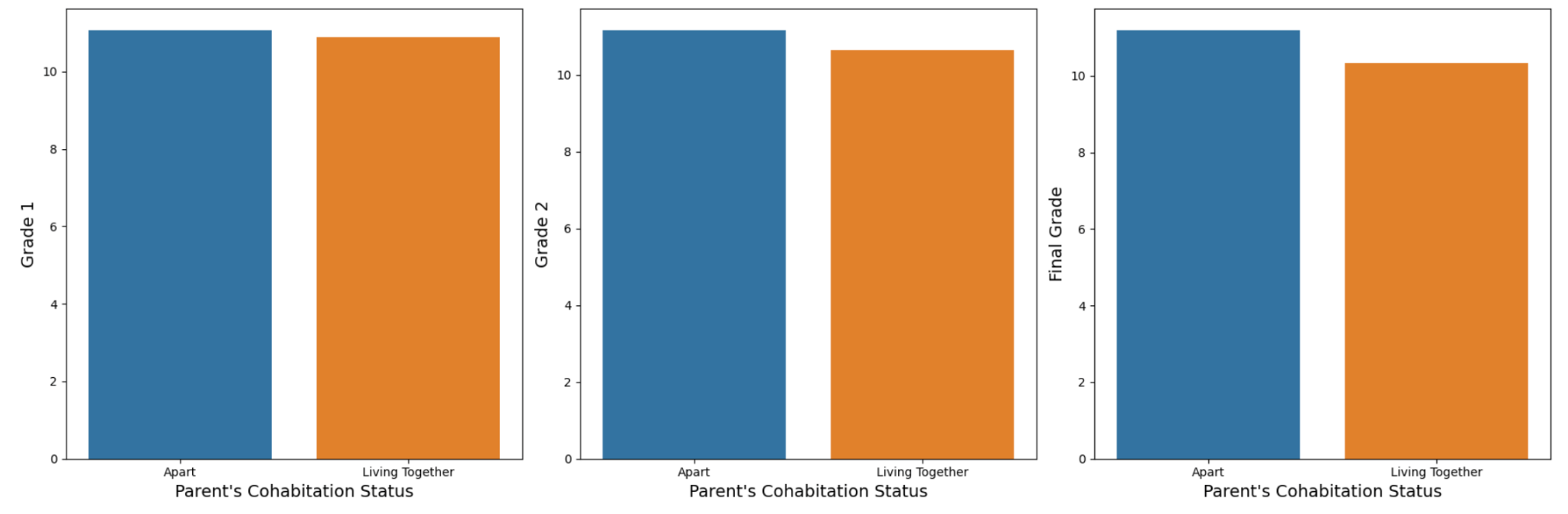


*Figure 10. Distribution of final grades (G3)*

# Comparing features

Comparing features is an important part of exploratory data analysis and model development. It helps in gaining insights into the data and improving model performance, and making informed decisions in the data analysis.

Also comparing features is a good way to identify highly correlation between features.If data set high dimensional, comparing features is helpful to decide to remove the features to reduce its high dimension. Moreover, it's easy to explain the dataset to somebody else.

*Figure 11. Parents status and grades compare*

As I see here students whose parents are apart have good scores at Grade 1, Grade 2, Grade 3. Result was unexpected. Also students whose parents are living together losing the scores from grade 2.