Table of content

Introduction

Exploratory Data Analysis (EDA)

Data Visualisation

Comparing features

Data Preparing

Data Modelling

Conclusion

Reference

# 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.

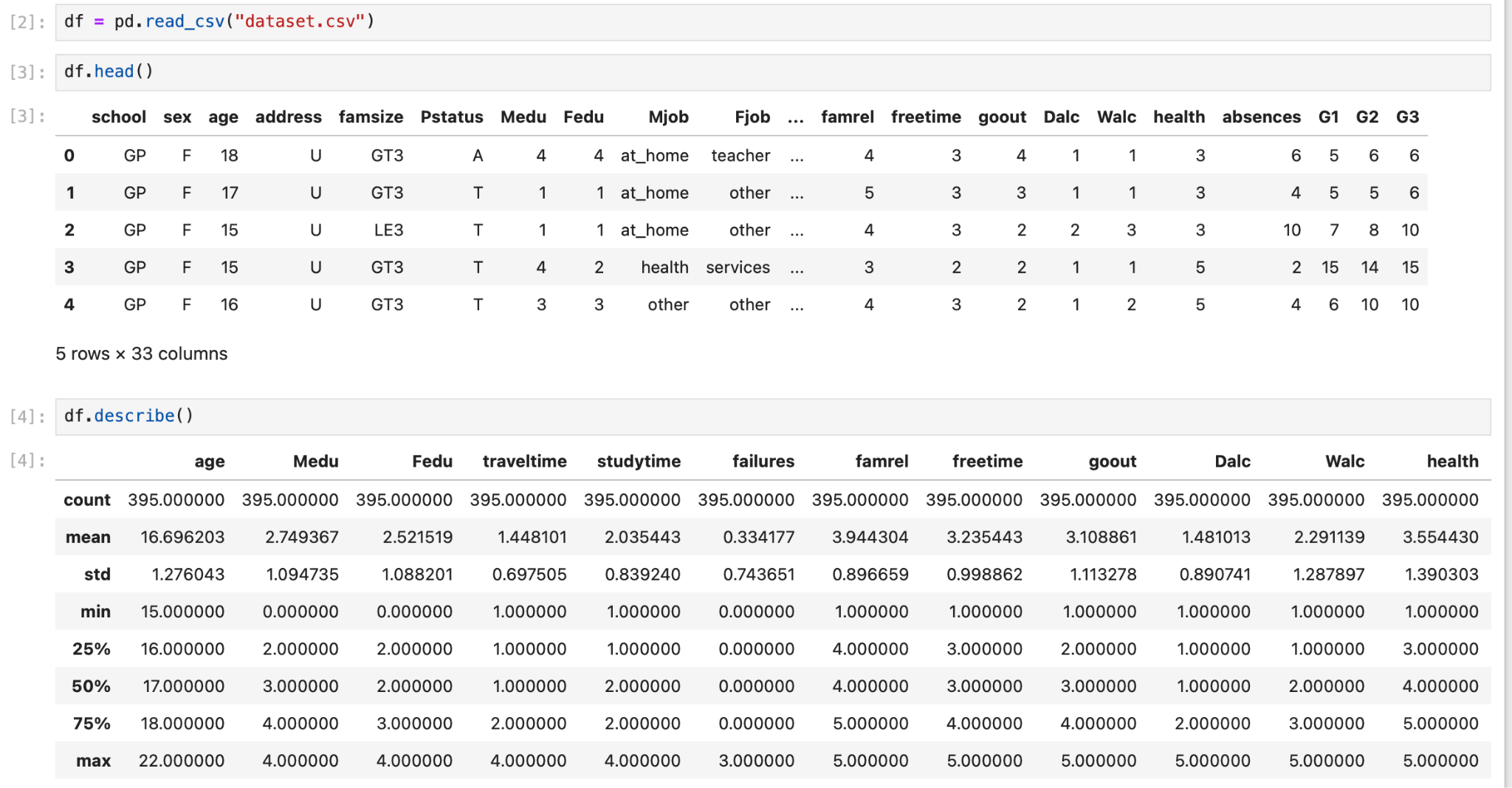
# Exploratory Data Analysis (EDA)



*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.

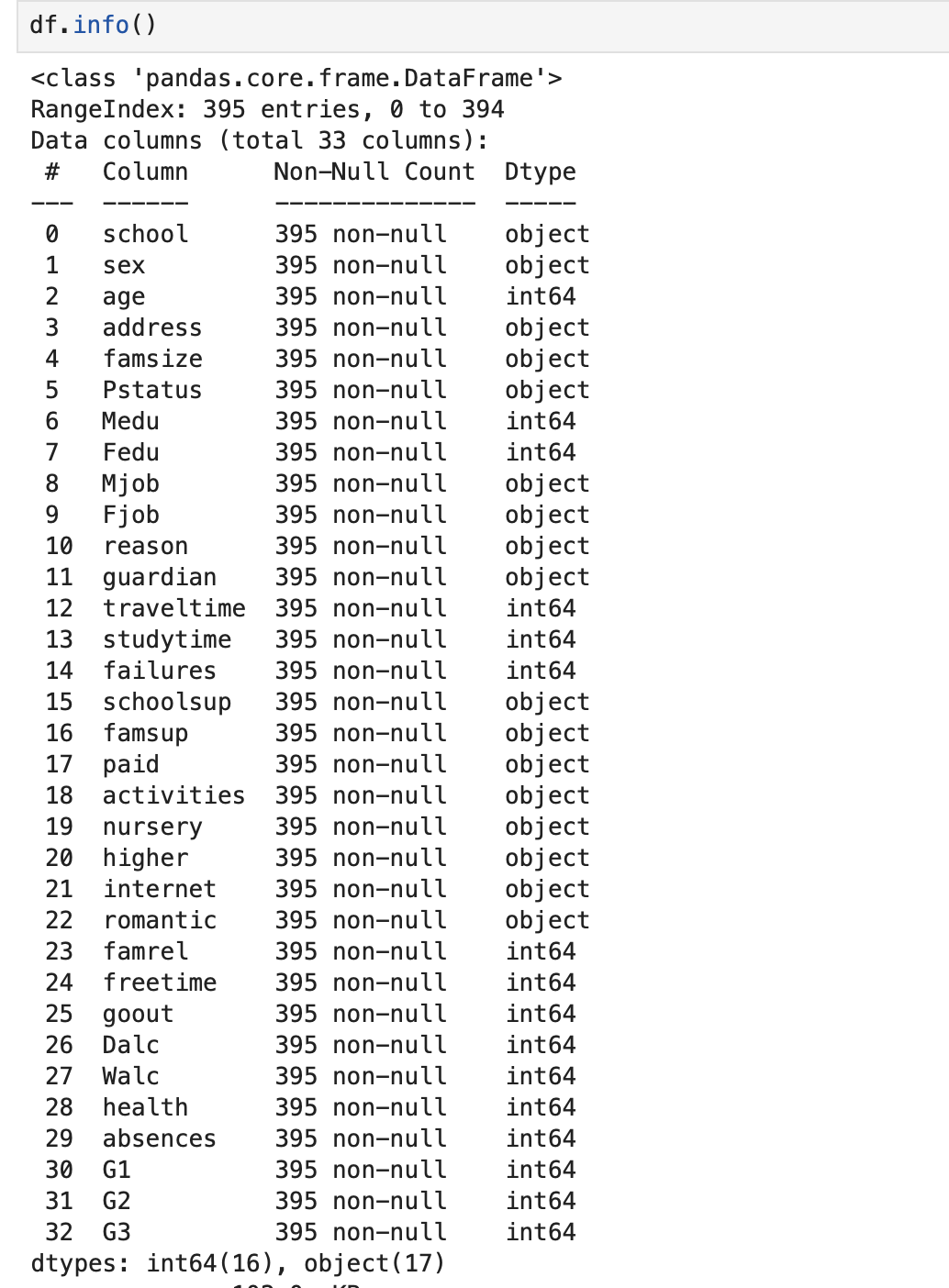
*(Word count 152)*



*Figure 2. Head of the Dataset and describe*

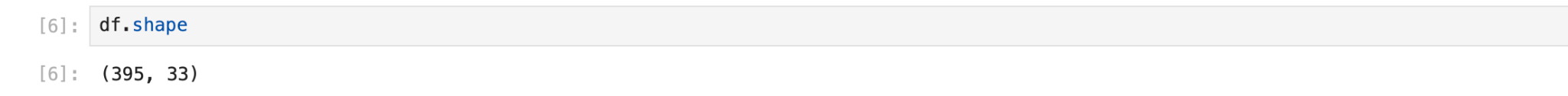
By using the head function I am able to quickly overview my dataset. In this case my dataset seems as it has a lot of categorical columns and it has 33 columns.

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

*(Word count 51)*

*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.

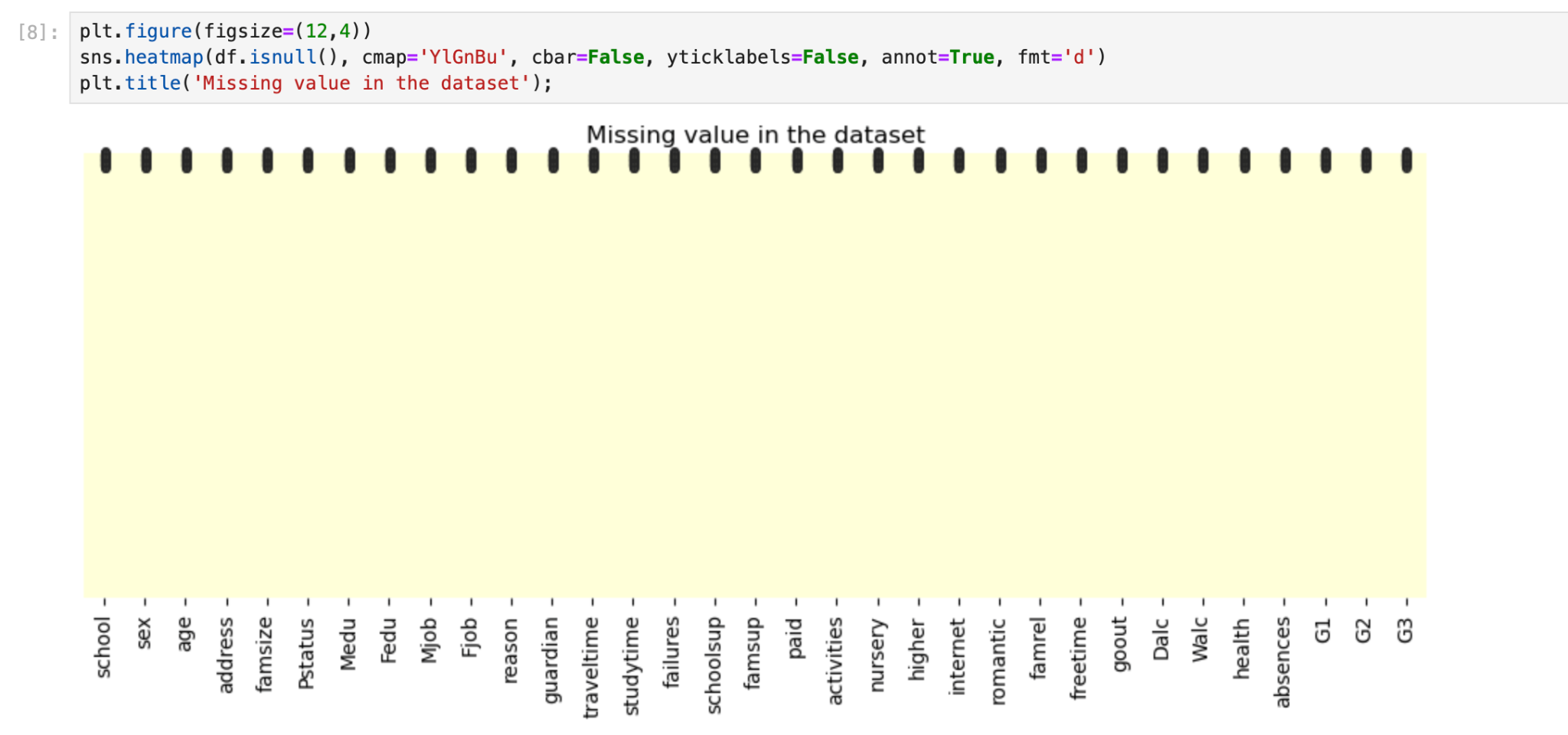


*Figure 5 . Checking missing value*

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

*(Word count 43)*

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*

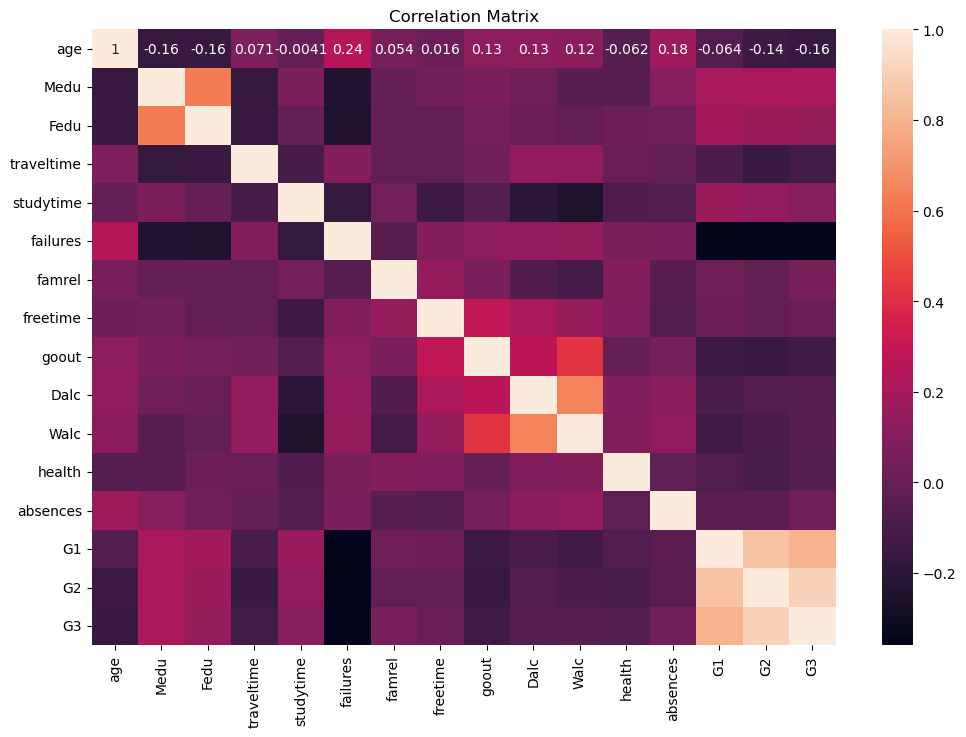
This method is very helpful in seeing the missing values in the dataset. In this dataset there is no missing value so visualization looks clean.

# Data visualization

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

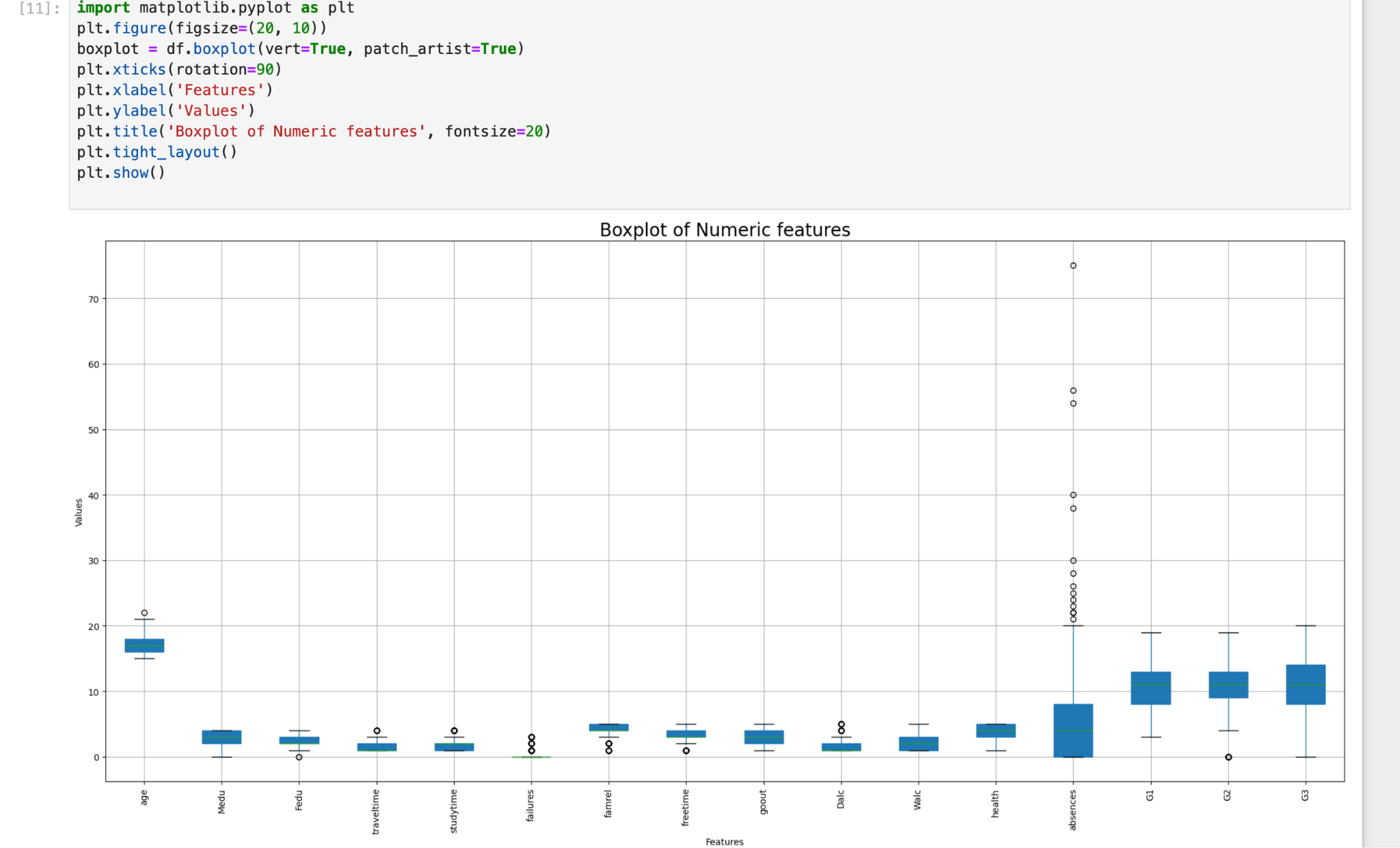
In this project I have done correlation matrix and boxplot of numeric features. I found some correlation between the features and some outliers from the columns.

*(Word count 101)*



*Figure 7. Correlation matrix*

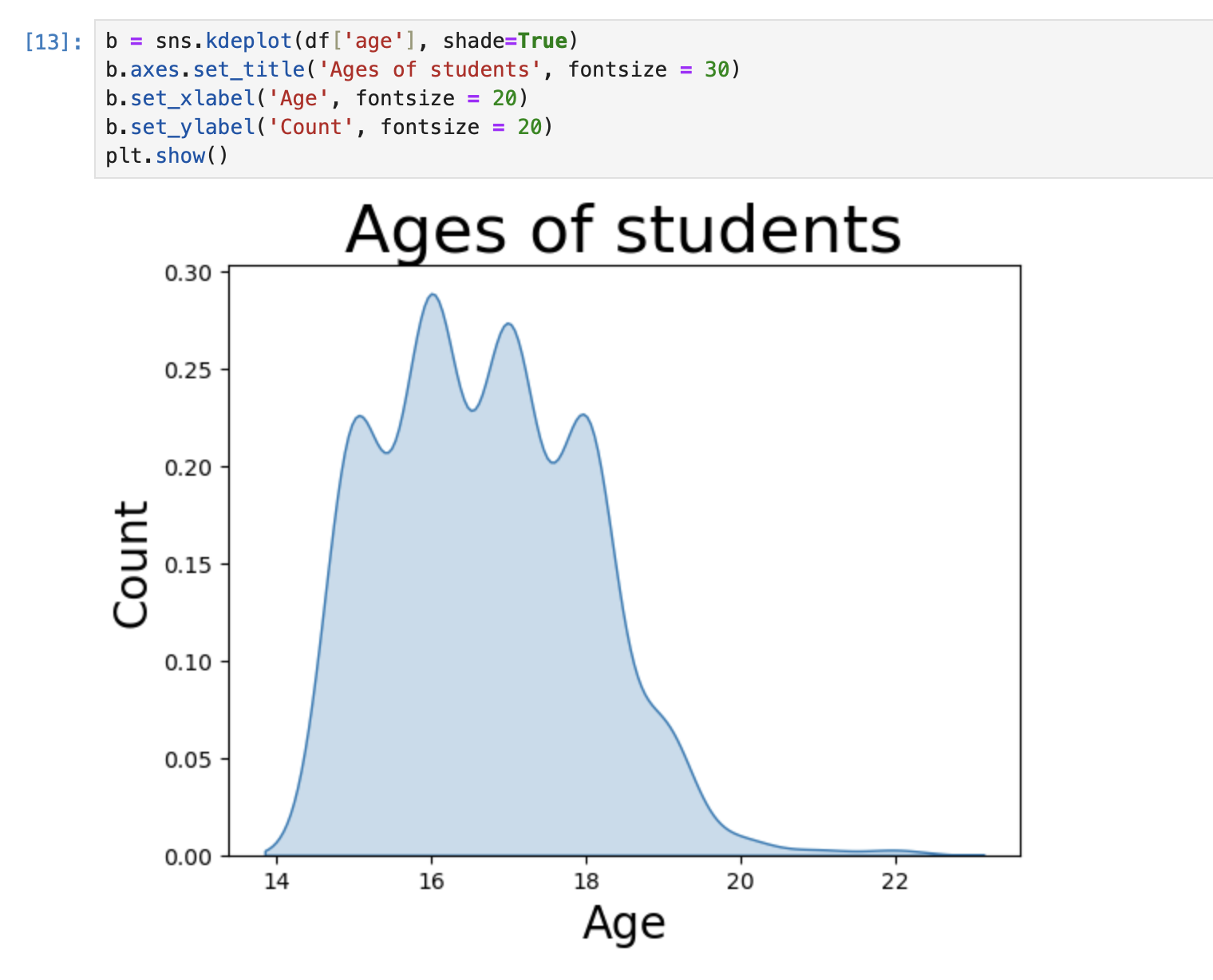
It tells me how closely related different things are in a dataset. With this visualization I can understand which features have strong connections. In this dataset parents' education and study time has strong connection with higher grades. Interesting part is that traveling time has a connection with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



*Figure 8. Boxplot of Numeric features*

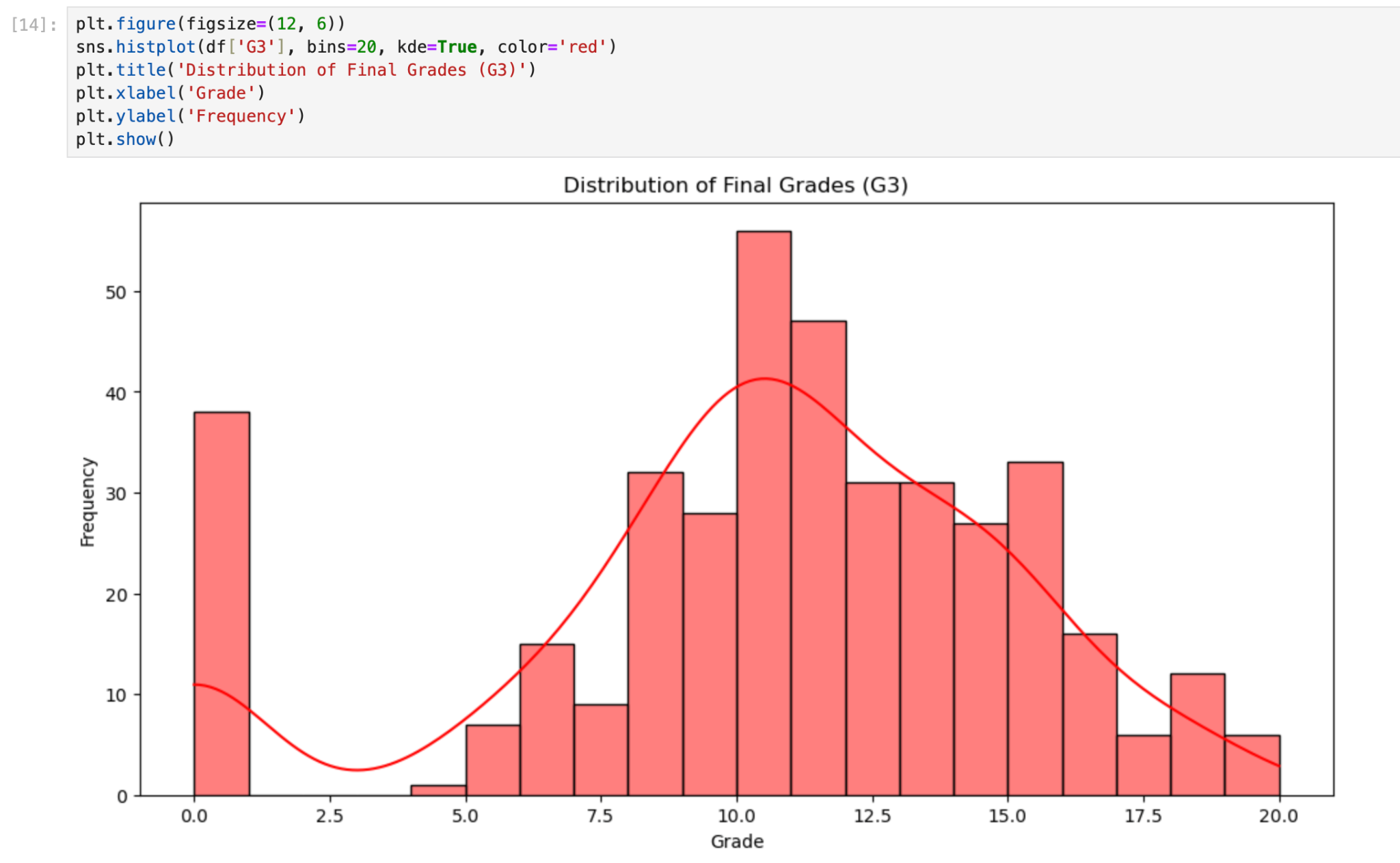
I checked my dataset with a box plot to identify its distribution, skewness and outliers. Here I found outliers from the column of absences.

*(Word count 78)*



*Figure 9. Age of students*

Most students in dataset is 16-18 years old.



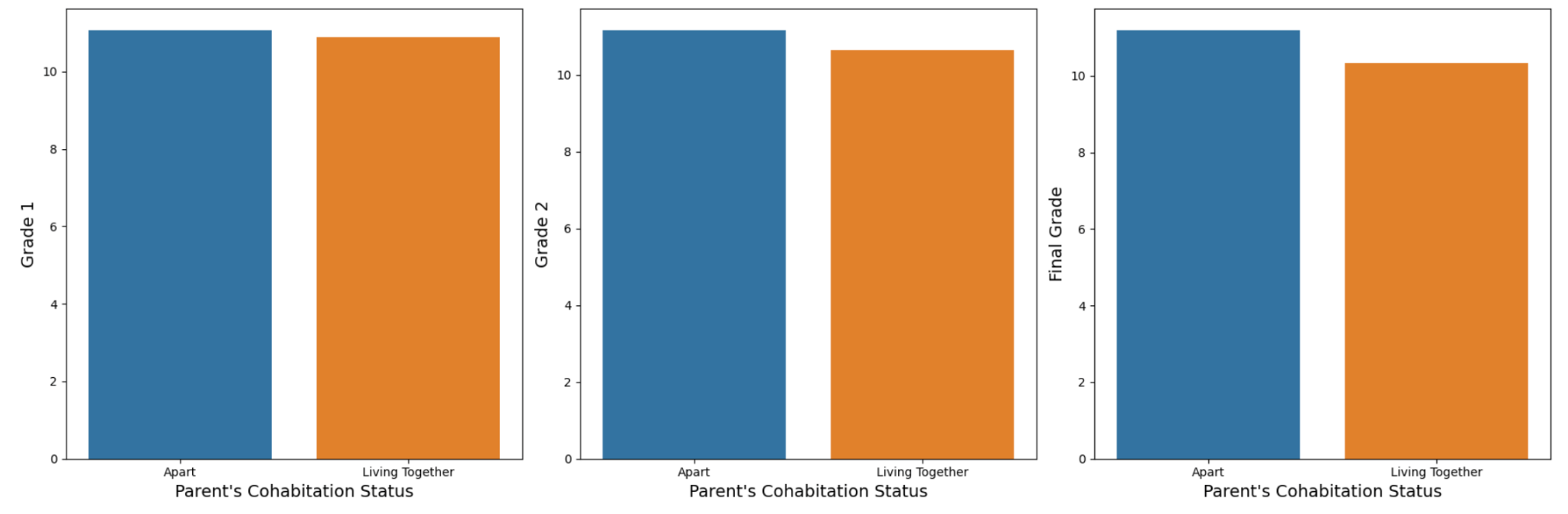
*Figure 10. Distribution of final grades (G3)*

*(Word count 7)*

# 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.

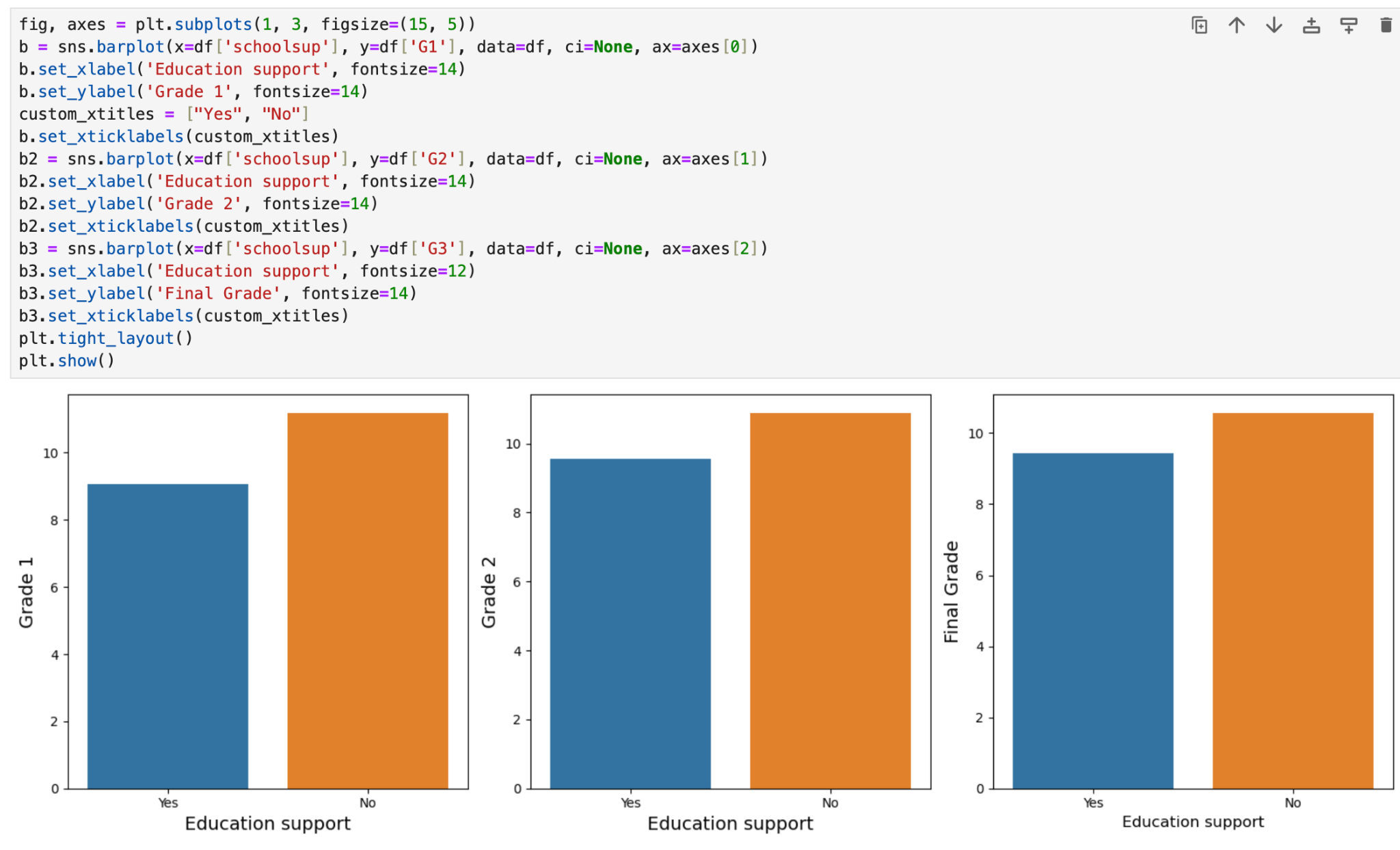
Also comparing features is a good way to identify highly correlation between features.If data set high dimensional, comparing features is helpful to decide to remove the features to reduce its high dimension. Moreover, it's easy to explain the dataset to somebody else.



*Figure 11. Parents status and grades compare*

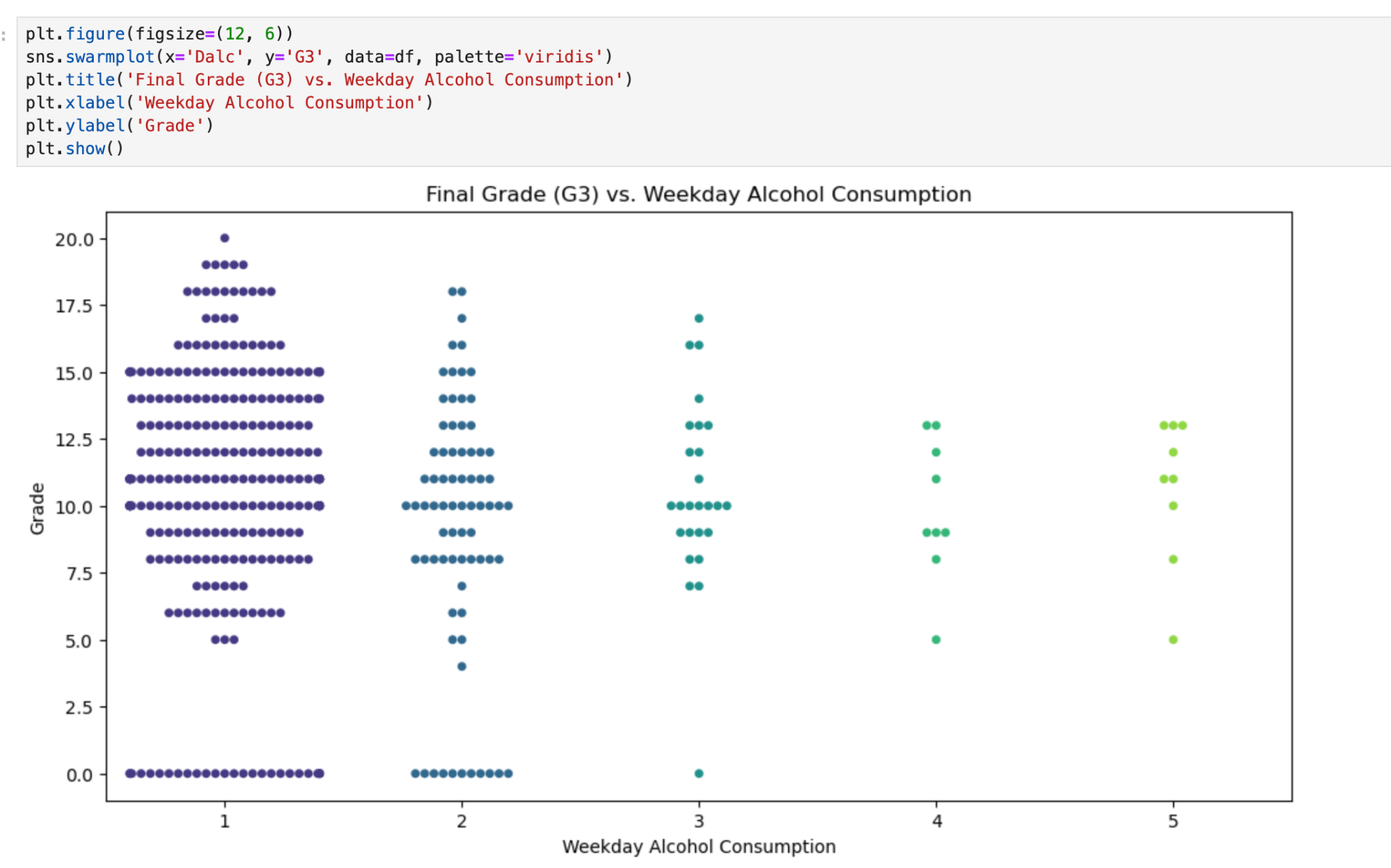
*(Word count 78)*

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 are losing the scores from grade 2.



*Figure 11. Education support*

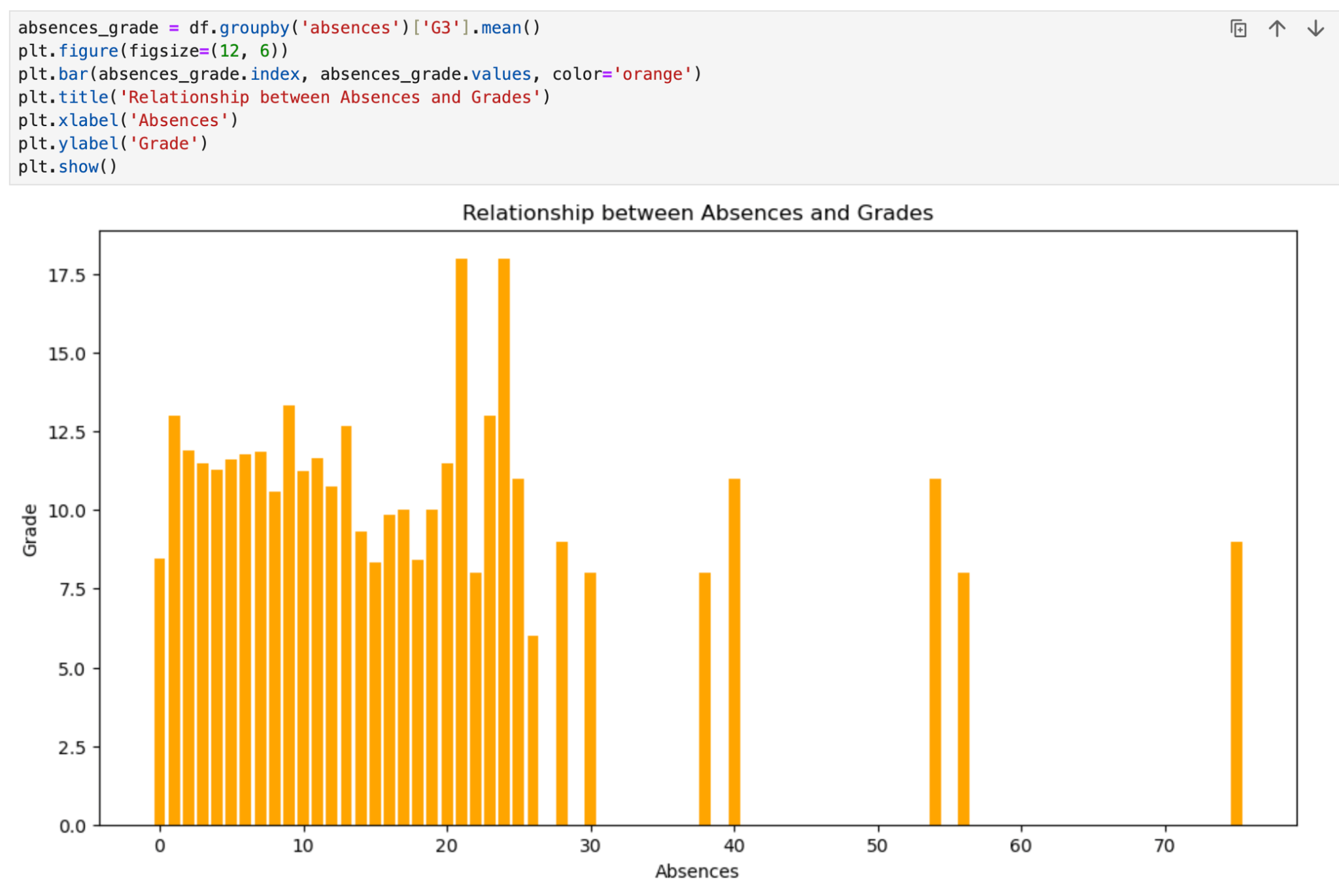
These bars showed students who have education support have lower scores than those who don't.



*Figure 12. Final grade vs Week day alcohol consumption*

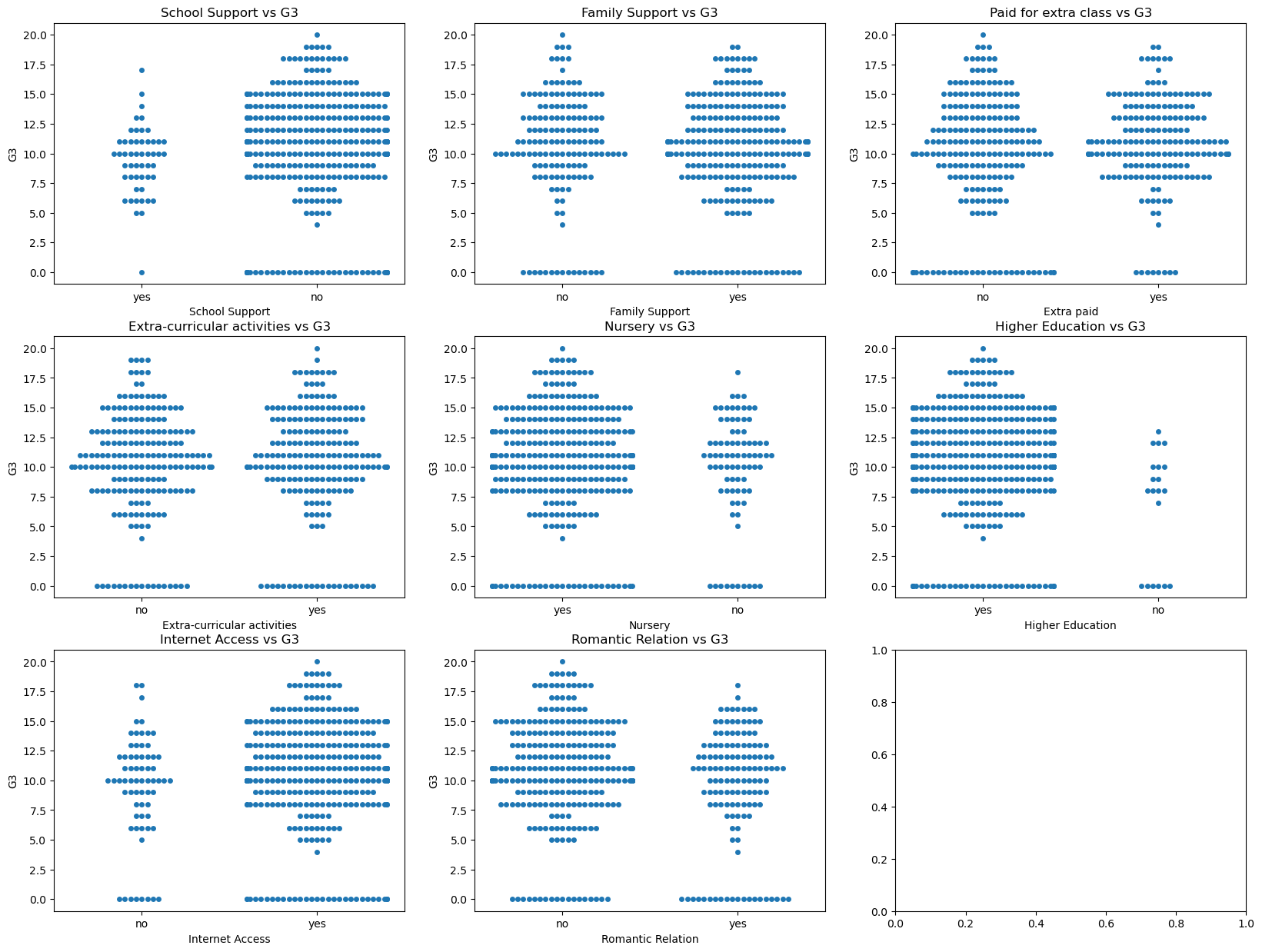
*(Word count 55)*

This graph showed me that students who is drinking 1 or 2 times in week days has normal scores but students using alcohol more than 3 times in week days has low scores.



*Figure 13. Relationship between absences and grades*

This graph says students whose absences are more than 30 are getting lower scores. So educators might focus on students absences reason to get them high scores. *(Word count 67*



*Figure 13. Students grades vs some features*

By using this method I can see a multi graph which is comparing students' grades with important features.

In this graph some information was found. For example having nursery, higher education, internet access are effects to get higher scores. Students who have romantic relationships have lower scores than those who don't.

# Data preparing

Data preparation and handling is very important to build machine learning models. Also its improved data quality, reduced overfitting. Earlier I checked the missing value (*Figure 5. Checking missing values)*  and found zero which is good. But almost half of my features are categorical so it need to be changed to numeric.



*Figure 13. Converting categorical values*

*(Word count 105)*

This code is performing mapping operations on specific columns and changed my categorical datas to numerical. After this code all my categorical features changed to 0 and 1, and on some columns 0 to 5.

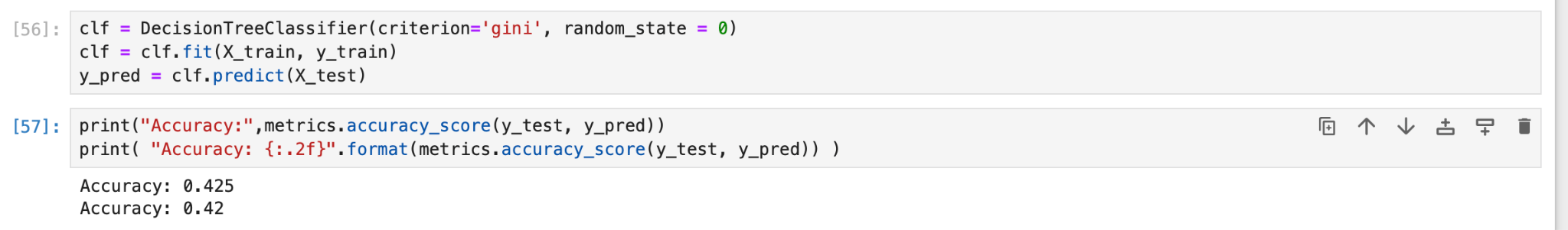


*Figure 13. Converting categorical values*

This code prepares the data for training a machine learning model by separating the features and target variable, performing one-hot encoding, and splitting the dataset into training and testing sets. It also prints out the shapes of the different data arrays to verify their dimensions.

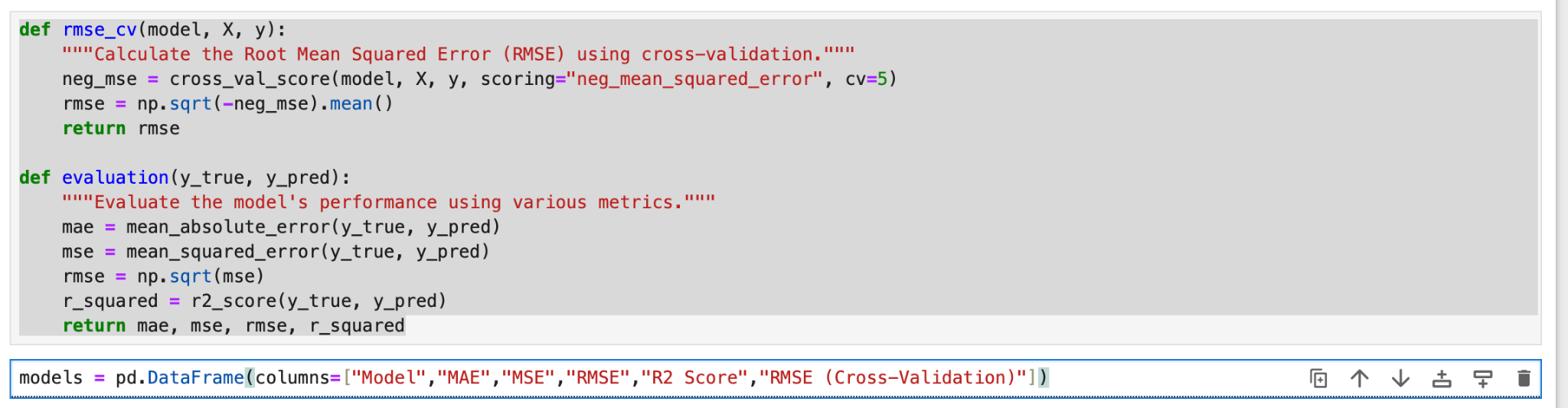
*(Word count 80)*

# Data modeling



*Figure 13. Building Decision tree classifier model*

In this section I'm trying to build a decision tree classifier model but accuracy 0.42 is very low to do proper prediction.



*Figure 14. Preparing model*

With this code I'm preparing a machine learning model for Linear regression and Random forest regressor.

*(Word count 43)*





*Figure 15. Linear regression model*

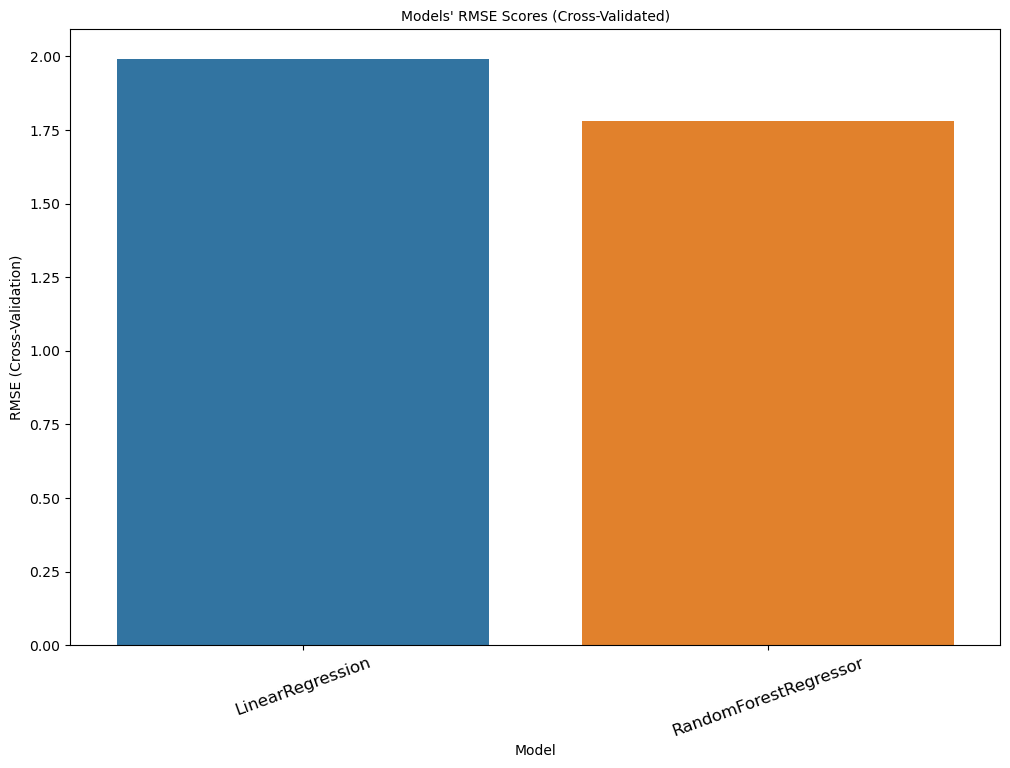
This model is 74.3 percent explains the variance in the target variable.



*Figure 16. Random forest regressor*

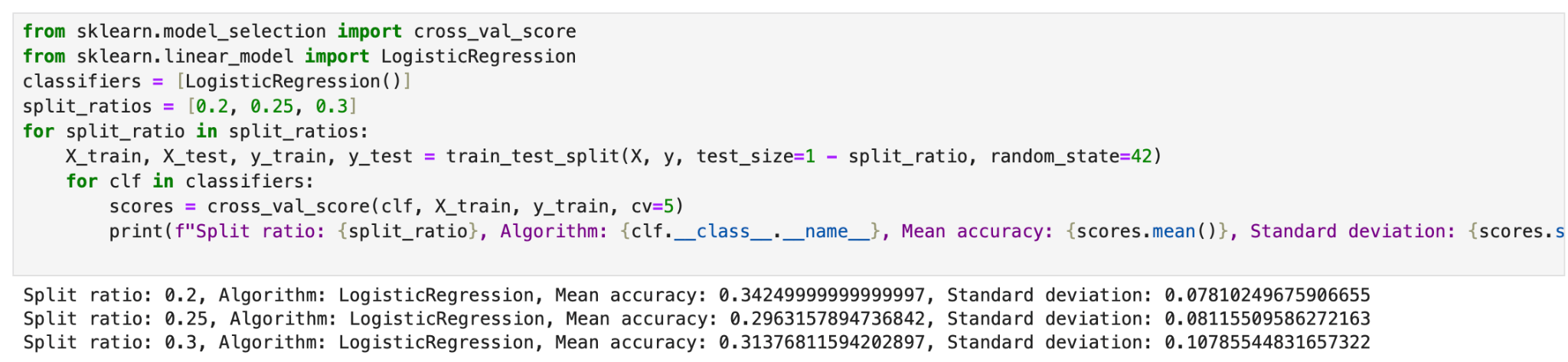


*(Word count 13)*



*Figure 16. Displaying machine learning models*

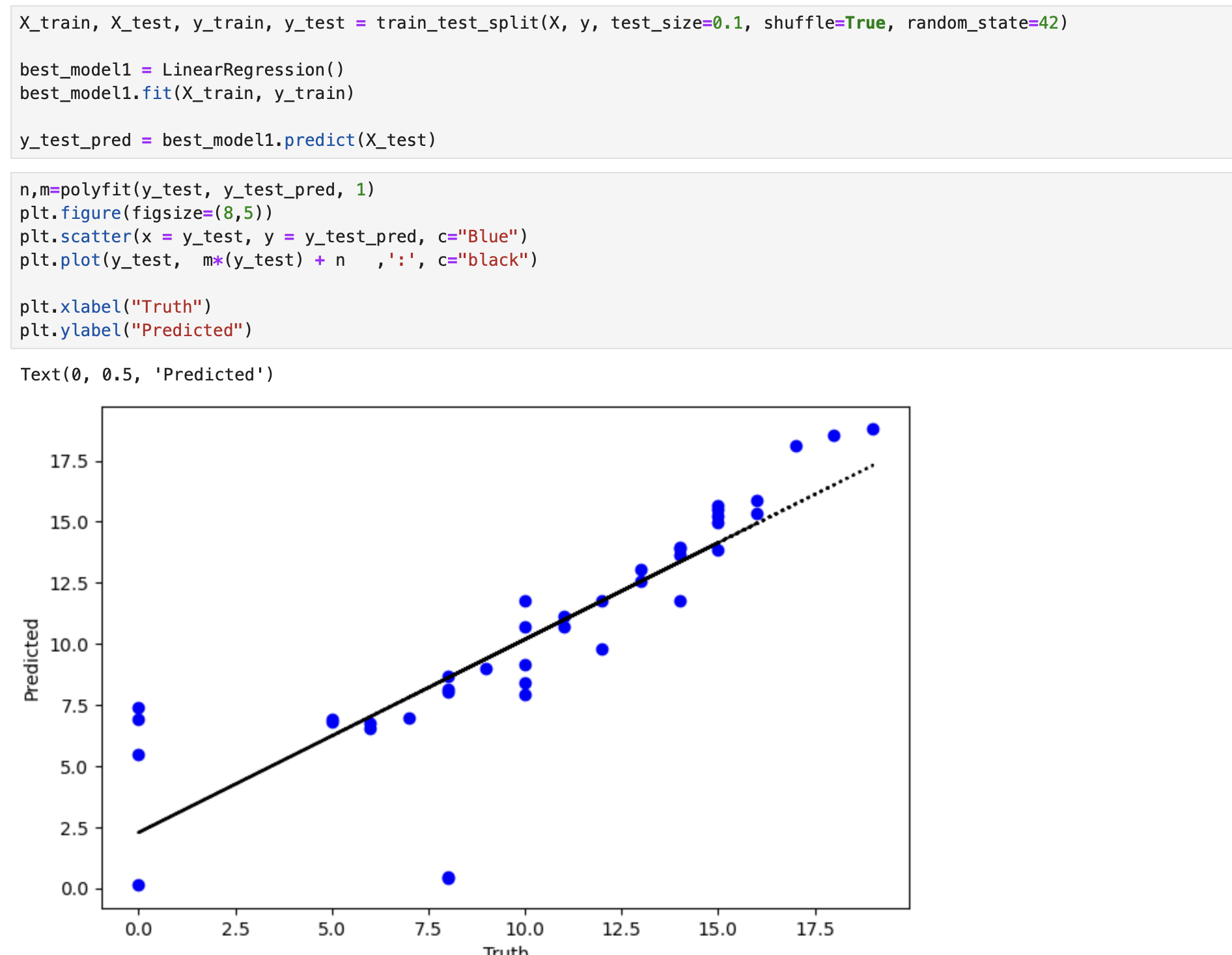
From this graph I can see Linear regression is most suitable for my dataset.



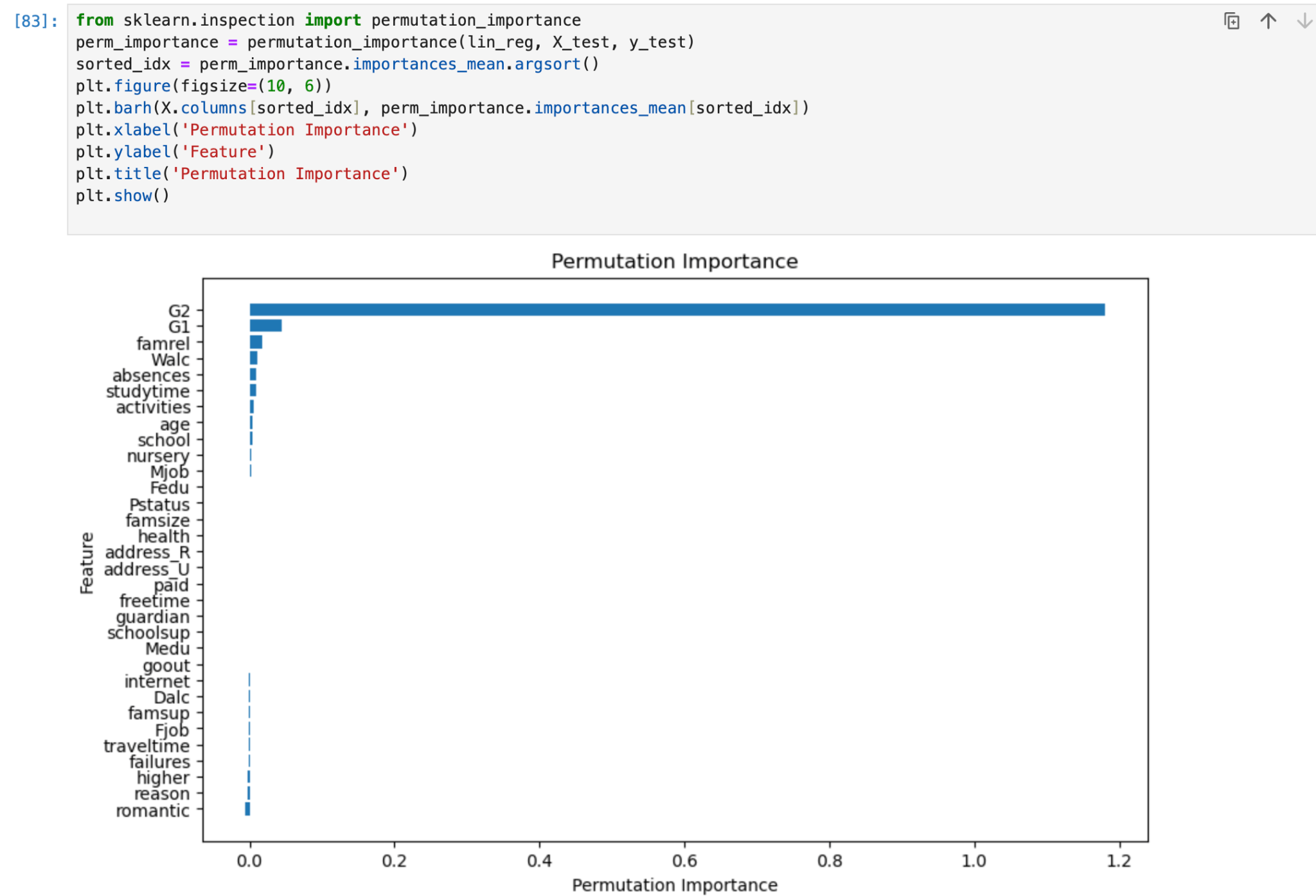
*Figure 17. Logistic Regression*

Before going further I checked Logistic Regression but the result is not better than Linear regression.

*(Word count 31)*



*Figure 17. Linear Regression prediction*



*Figure 18. Feature Importance*

As we see, students who study well in the second grade are more likely to study well in Grade 3. And in the middle columns such as parents status, freetime looks like they are not important with the target variable.

# Conclusion

By working on this dataset I found that what situation is most important for students to get a higher score in their grades. And I identify interesting things such as having school support doesn't actually affect students' performance and students whose parents are together have lower points than who doesn't. The result of my work result is to study well in the grade 3(G3) is highly depends on how study at G2.

Implementing this model is very important to educators and students. Finding the reason why students perform well or not is the opportunity to prevent failure and would support on their perform.

*(Word count 145)*

*TOTAL WORDS: 1179*

# Reference

* P. Cortez and A. Silva. (April, 2008) *Using Data Mining to Predict Secondary School Student Performance.* [online] Available at: https://www.kaggle.com/datasets/dipam7/student-grade-prediction/data
* McKinney, W. (2017). *Python for Data Analysis*. 2nd ed. Beijing: O’Reilly Inc.
* Sham, K. (2020). *Linear Regression in Python with Scikit-Learn*. [online] Analytics Vidhya. Available at: https://medium.com/analytics-vidhya/linear-regression-in-python-with-scikit-learn-e1bb8a059cd2.
* ‌OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model].
* https://chat.openai.com/chat (https://chat.openai.com/chat)