**Machine Learning 5710 (Assignment 4)**

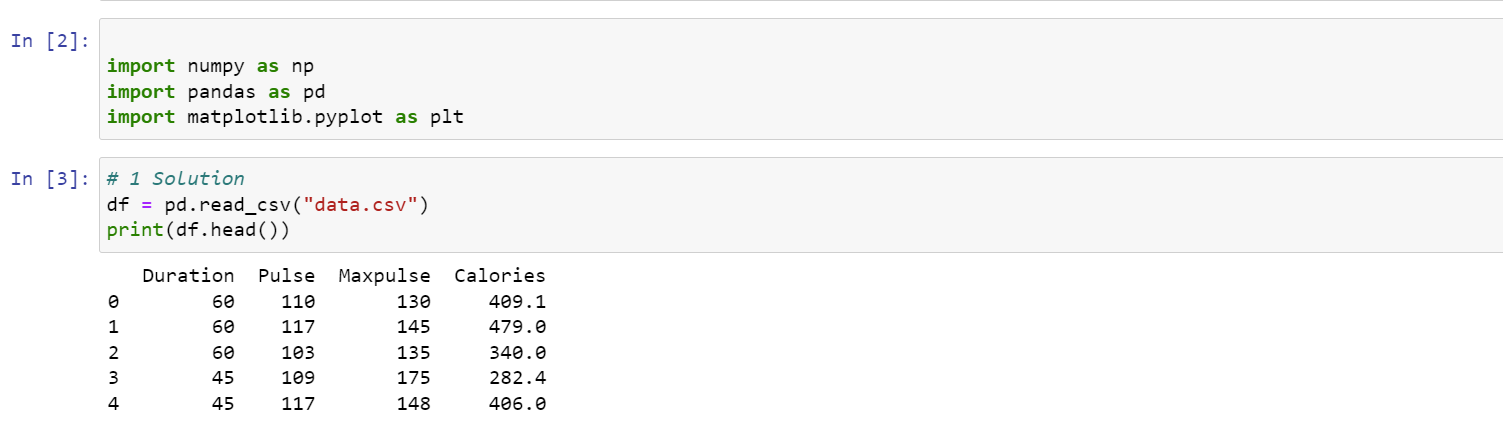
**Name - Nikhil Manikya**

**700734200**

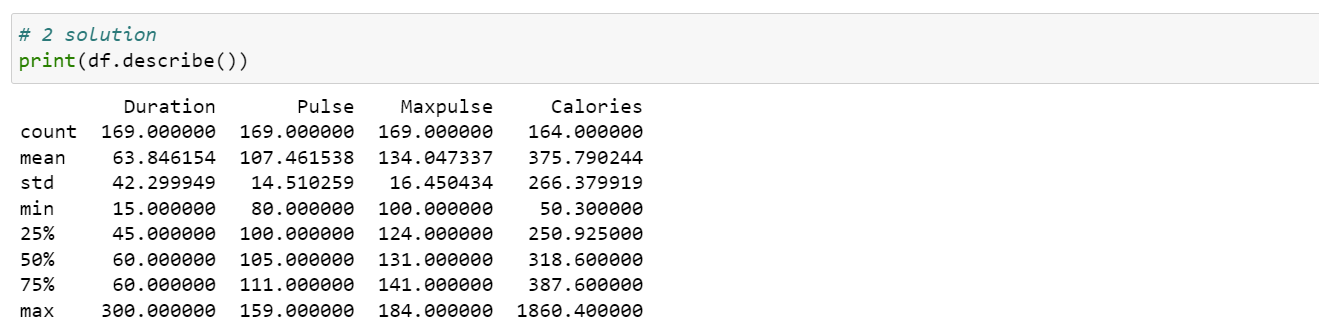
**Video link-** https://drive.google.com/file/d/1iugScaWVRshcGY7QMfEIeFg6NovVt8\_p/view?usp=sharing

**Git link -** https://github.com/NikhilManikya123/ML5710\_Assignment-\_4

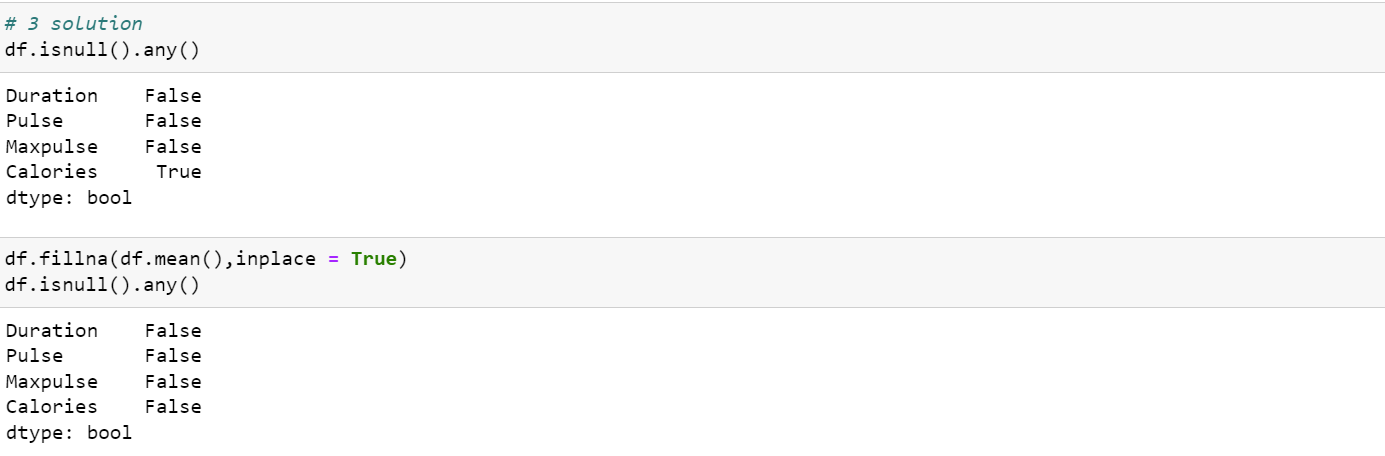
1. Read the provided CSV file ‘data.csv’.



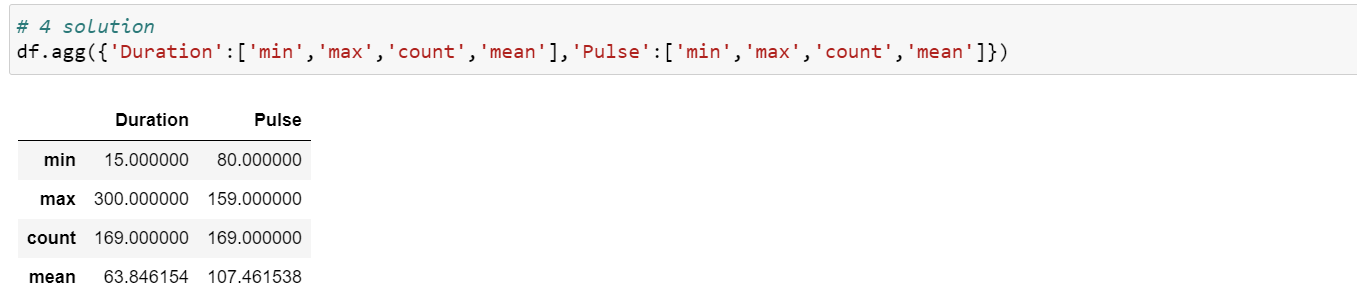
2. Show the basic statistical description about the data.



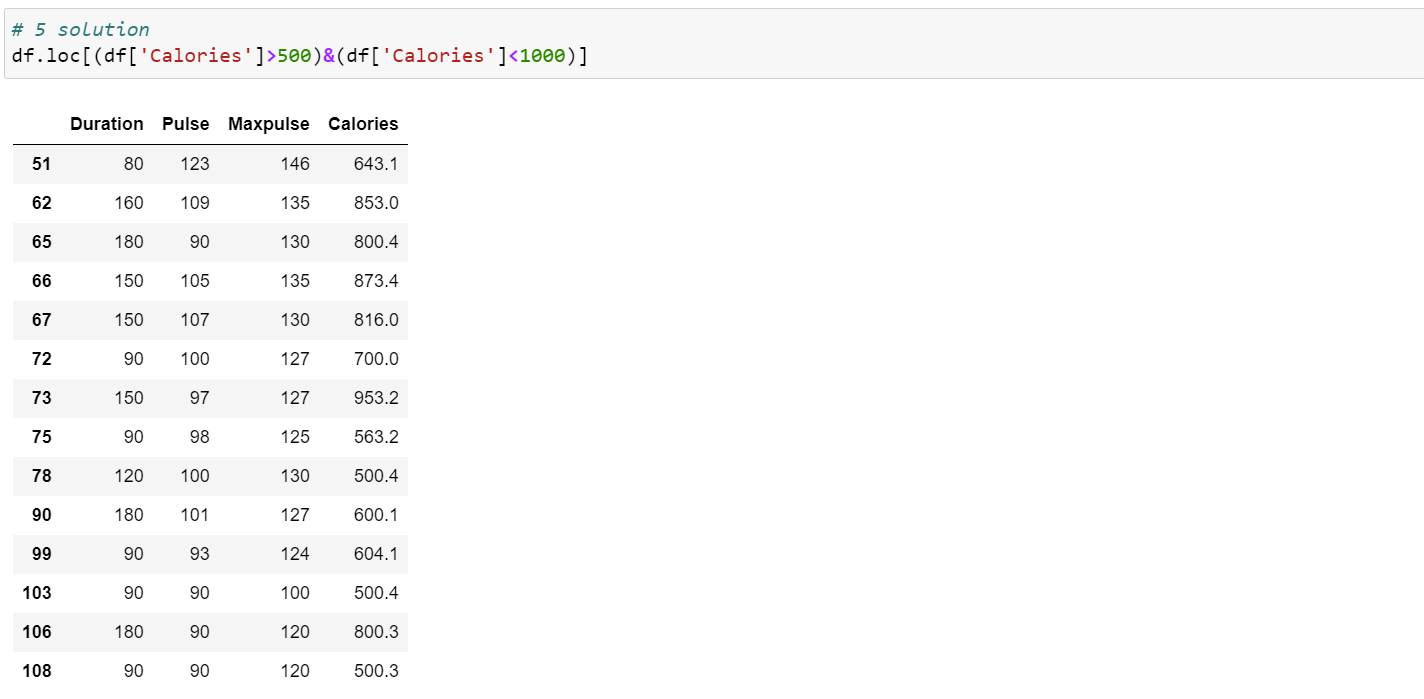
3. Check if the data has null values. a. Replace the null values with the mean



4. Select at least two columns and aggregate the data using: min, max, count, mean.



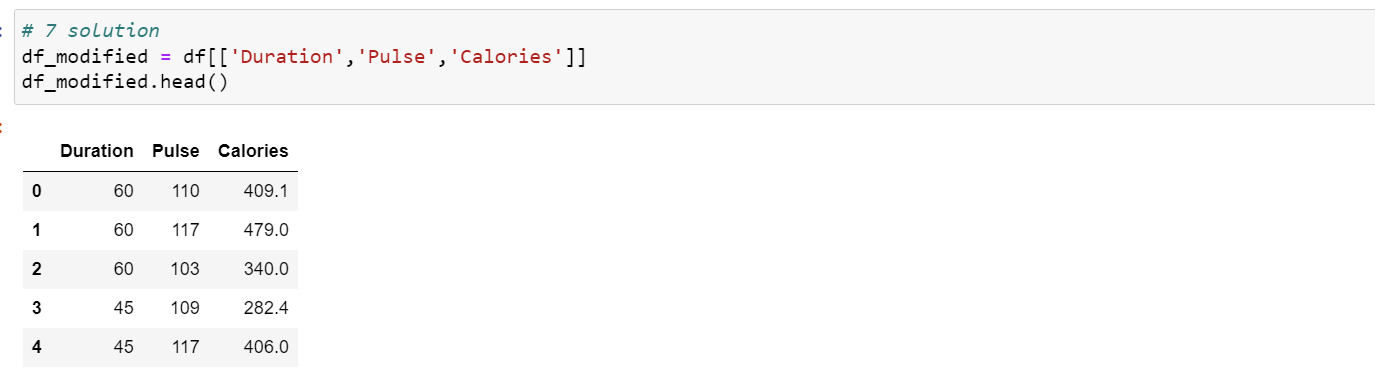
5. Filter the dataframe to select the rows with calories values between 500 and 1000.



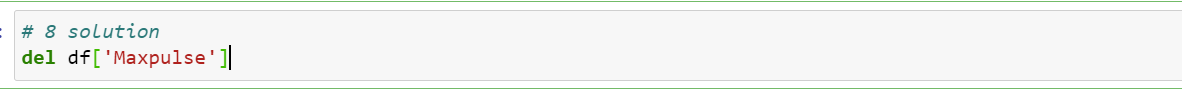
6. Filter the dataframe to select the rows with calories values > 500 and pulse < 100.



7. Create a new “df\_modified” dataframe that contains all the columns from df except for “Maxpulse”.



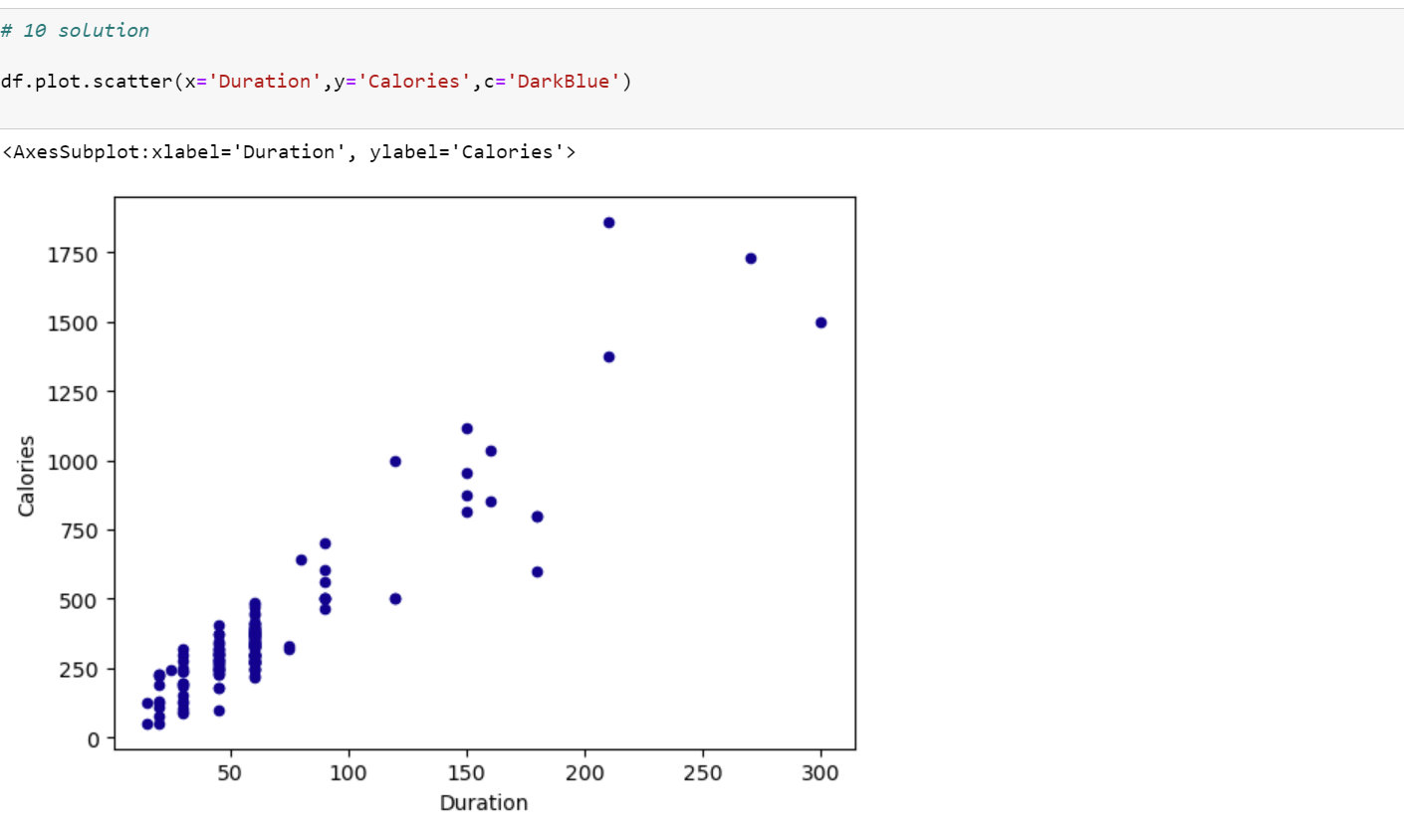
8. Delete the “Maxpulse” column from the main df dataframe



9. Convert the datatype of Calories column to int datatype.



10. Using pandas create a scatter plot for the two columns (Duration and Calories).

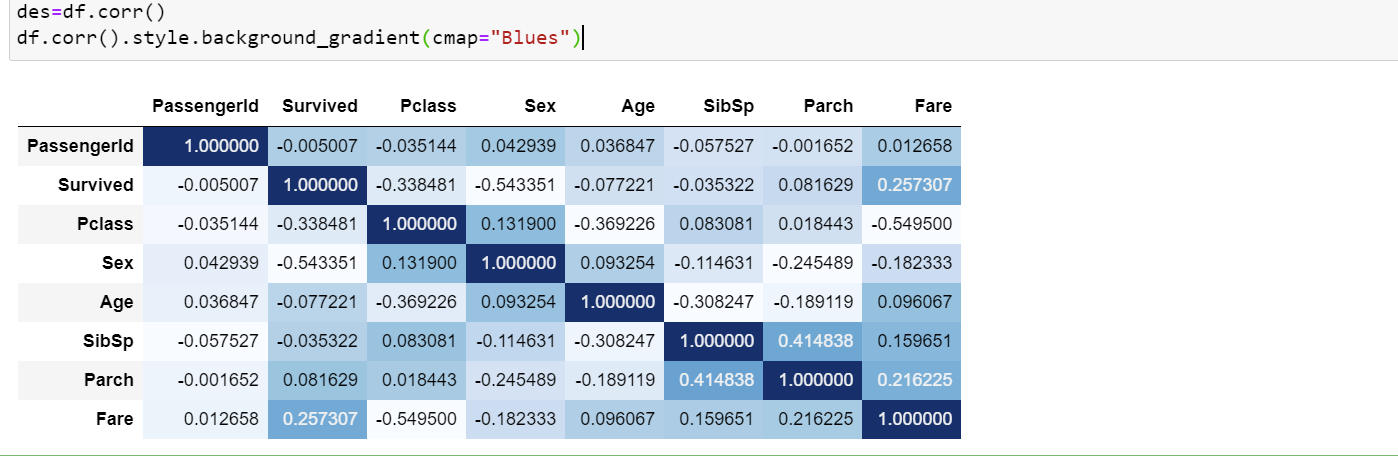


**(Titanic Dataset)**

1. Find the correlation between ‘survived’ (target column) and ‘sex’ column for the Titanic use case inclass. a. Do you think we should keep this feature?



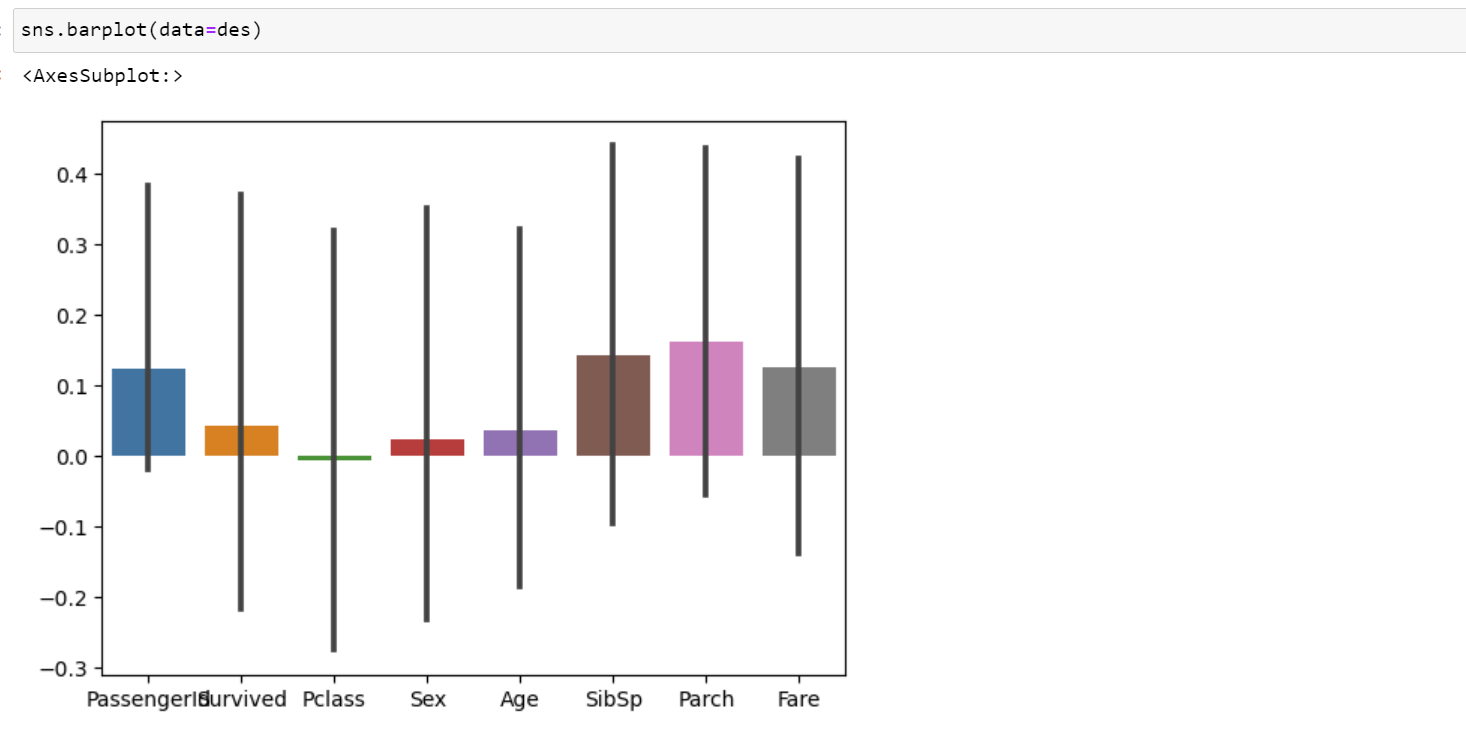
When the correlation coefficient is less than 0, there is a negative (inverse) correlation. Yet, the strength of the negative correlation between two securities may change over time (and they are almost never exactly correlated all the time). The model's results remain unchanged when a correlated characteristic is removed. The highly correlative feature should always be eliminated first, followed by the least correlated one.



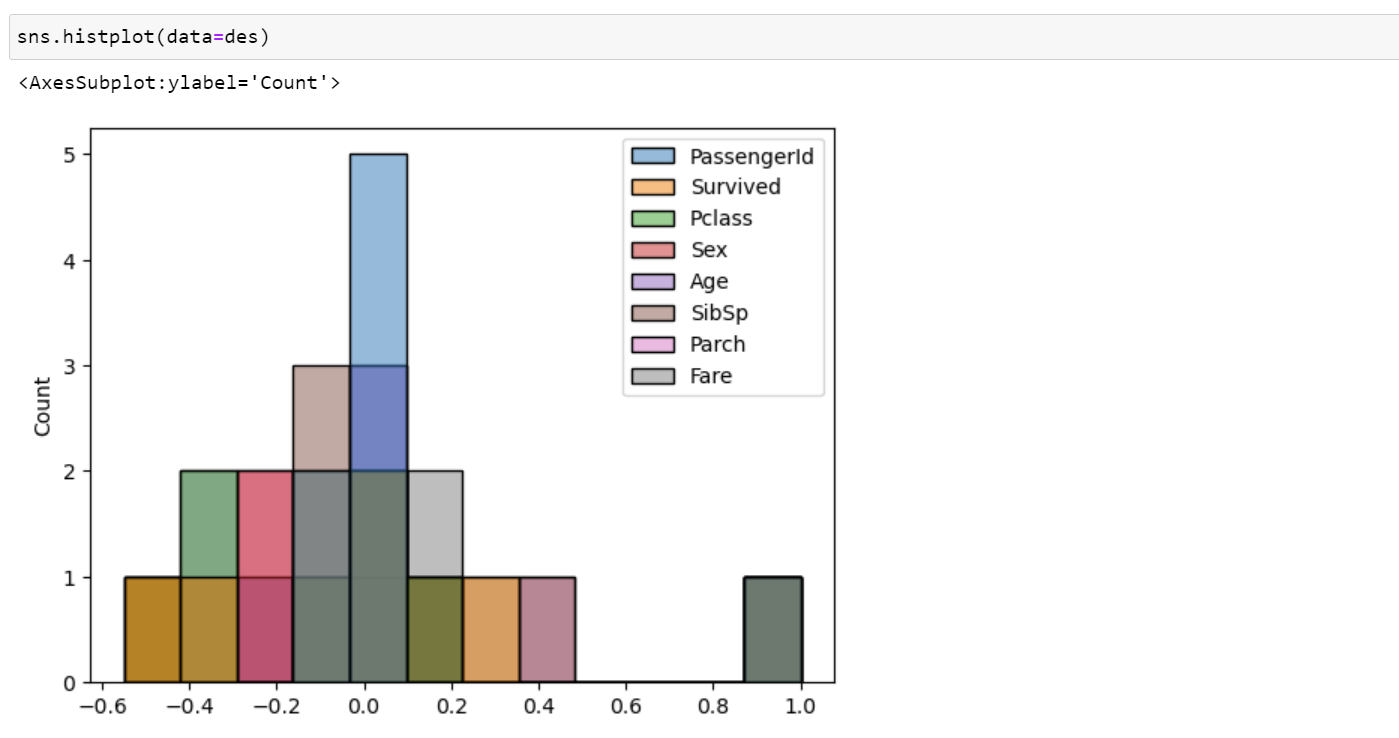
We can find the correlation directly by using background grandient visualizatiom.

2. Do at least two visualizations to describe or show correlations.

Bar plot visualization:



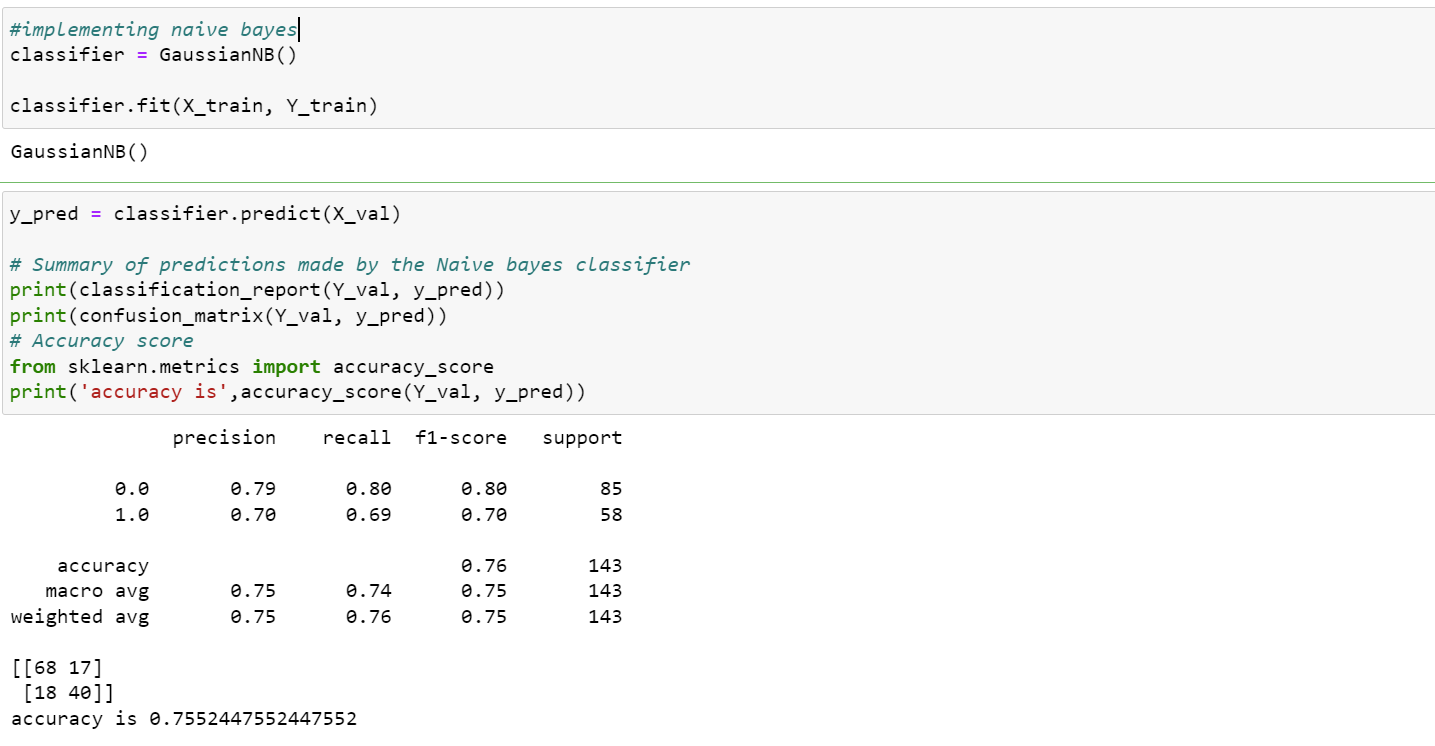
Histogram visualization:



3. Implement Naïve Bayes method using scikit-learn library and report the accuracy.



Now, Implementing now naive bayes and getting the summary including accuracy



Accuracy obtained is 0.76

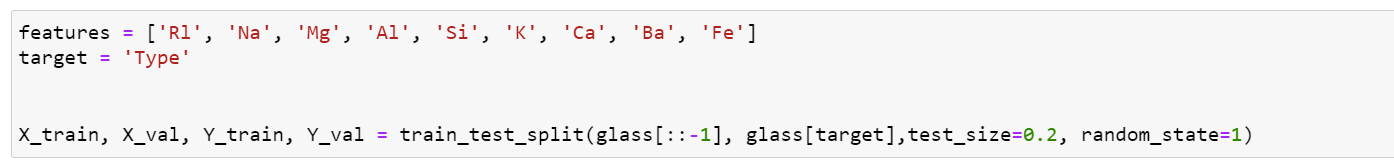
**(Glass Dataset)**

1. Implement Naïve Bayes method using scikit-learn library.

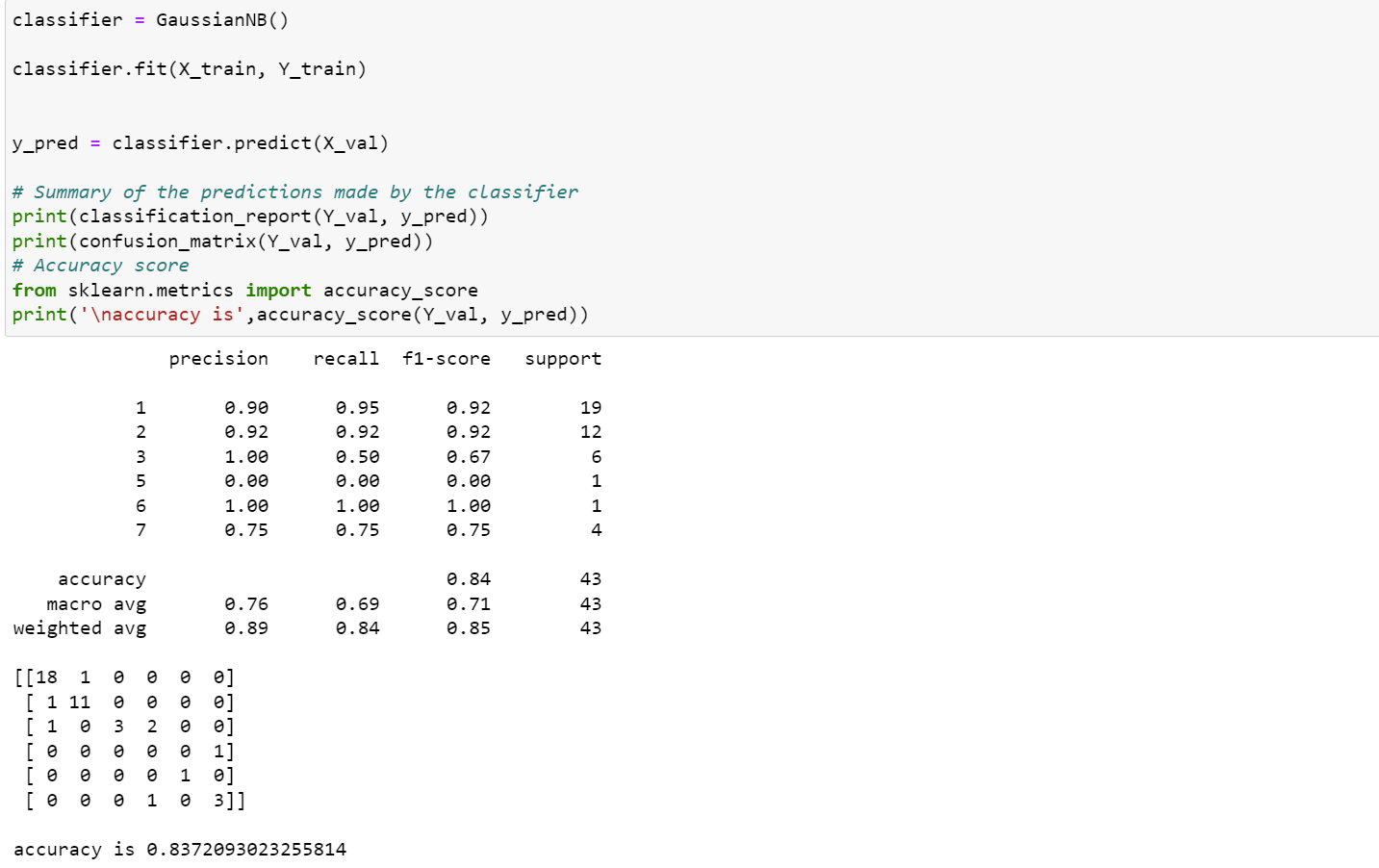
a. Use the glass dataset available in Link also provided in your assignment.



b. Use train\_test\_split to create training and testing part.



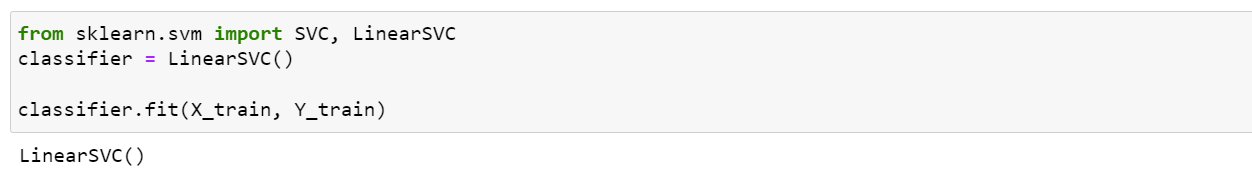
2. Evaluate the model on testing part



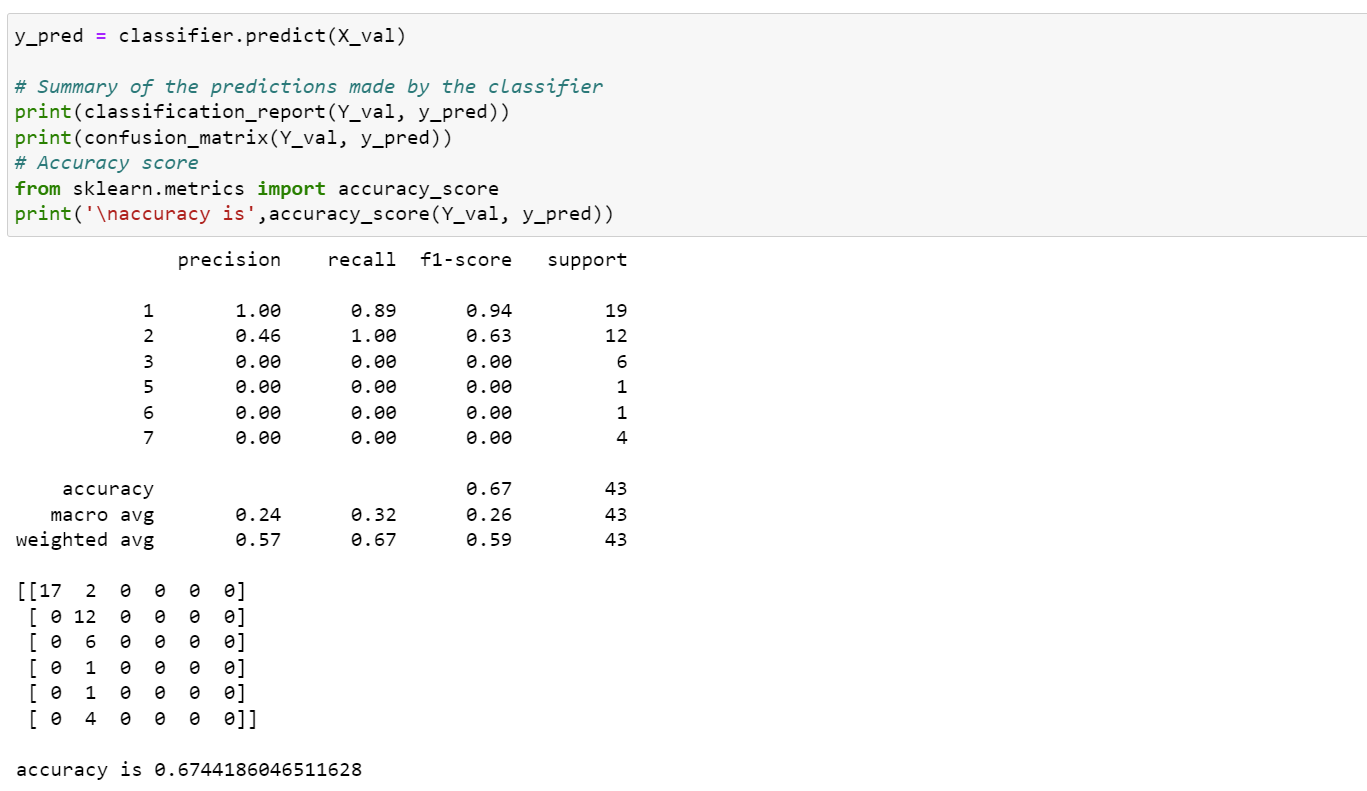
1. Implement linear SVM method using scikit library

a. Use the glass dataset available in Link also provided in your assignment.

b. Use train\_test\_split to create training and testing part.

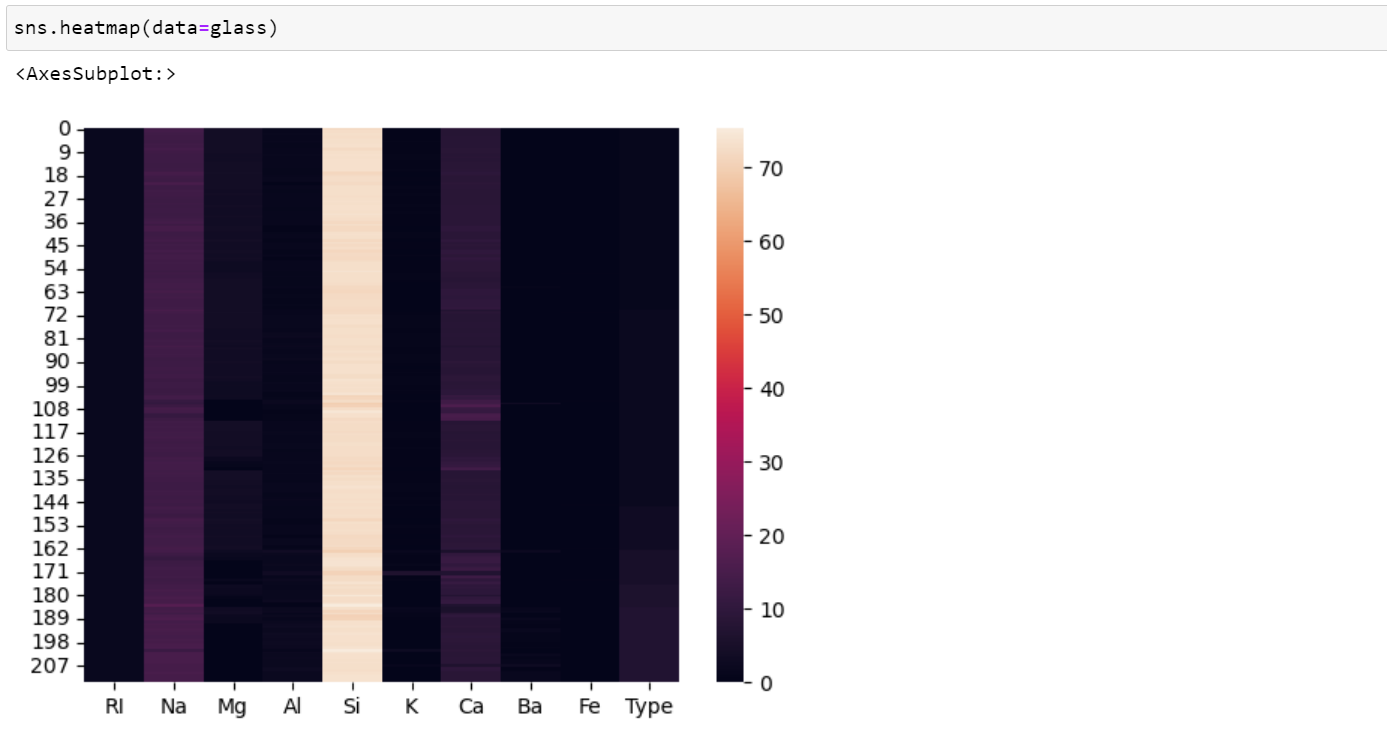


2. Evaluate the model on testing part

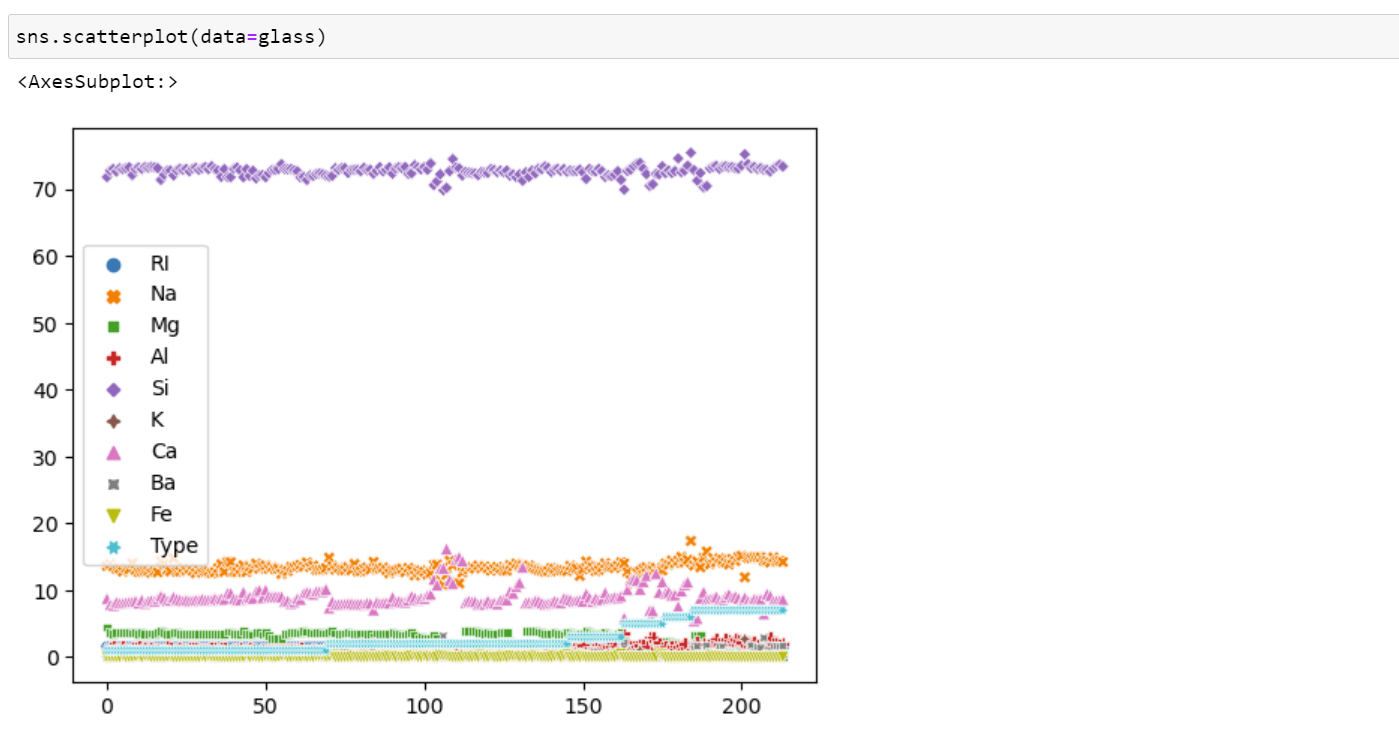


Do at least two visualizations to describe or show correlations in the Glass Dataset.

Heat map visualization:



Scatter plot visualization:



Which algorithm you got better accuracy? Can you justify why?

The Naive Bayes approach is superior to the Support Vector Machine method for data visualization, according to the aforementioned accuracy scores. Each algorithm performs differently depending on a number of variables. Thus, a small number of algorithms are effective for a small number of problems but not for others. We can evaluate different algorithms by evaluating the model, and then declare which one is best.