**Machine Learning (Assignment # 4)**

**Name: Mounika Kandula.**

**StudentId: 700747570**

**GithubLink:** https://github.com/MounikaKandula/MLAssignment4

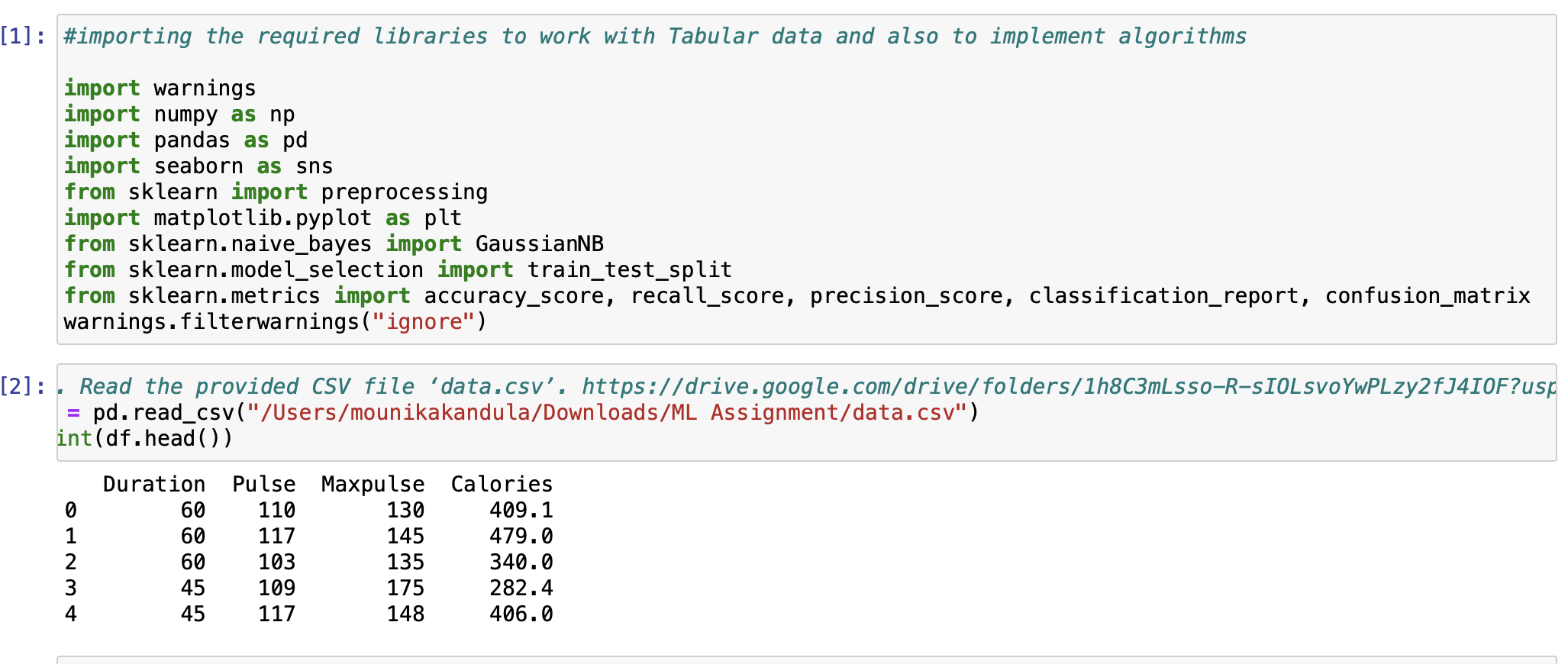
**VideoLink:** https://drive.google.com/file/d/143SSUNb0OF-jGvOHzOnOItZJiCGfByoX/view?usp=share\_link

**Question 1:**

Read the provided CSV file ‘data.csv’.

https://drive.google.com/drive/folders/1h8C3mLsso-R-sIOLsvoYwPLzy2fJ4IOF?usp=sharing

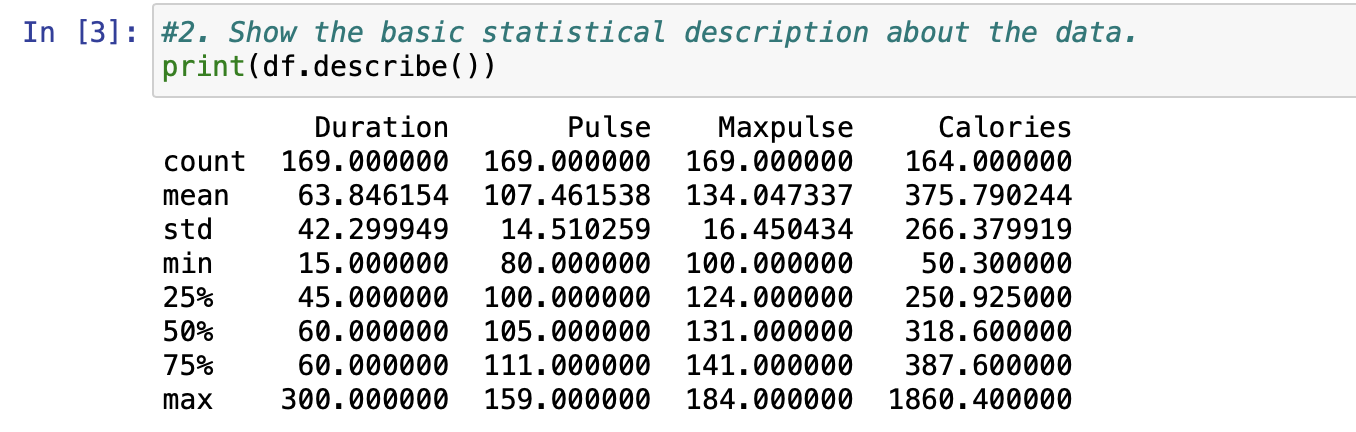
**Solution :**



**Explanation:** Reading the csv file.

* Show the basic statistical description about the data.

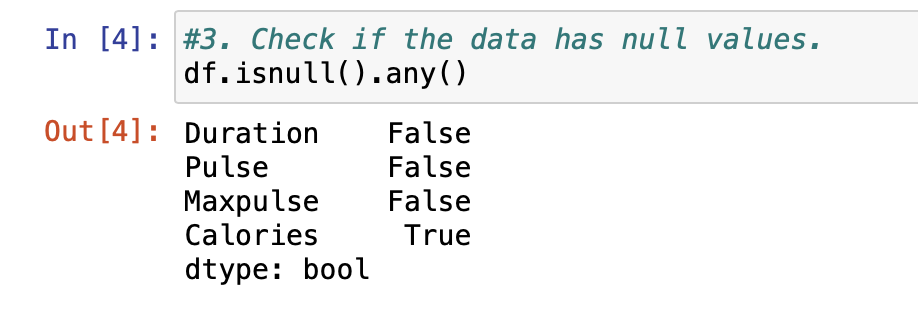
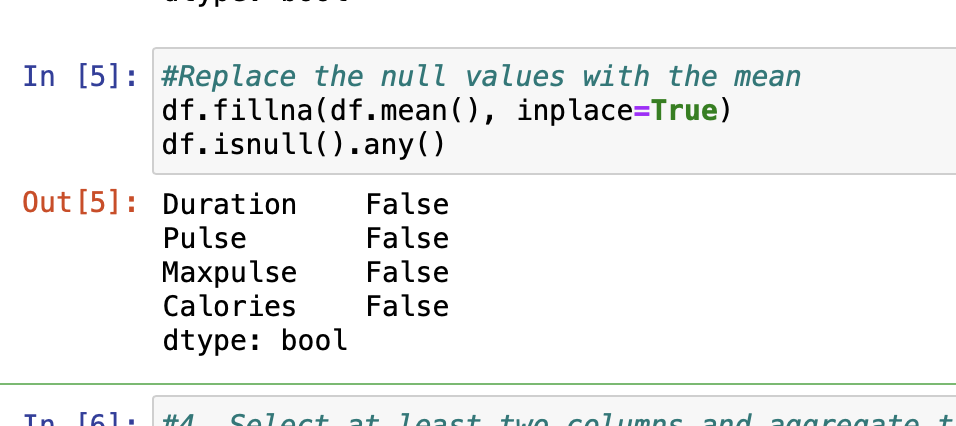
**Solution :**



**Explanation:** The above output show the description of the data.

* Check if the data has null values.
  1. Replace the null values with the mean

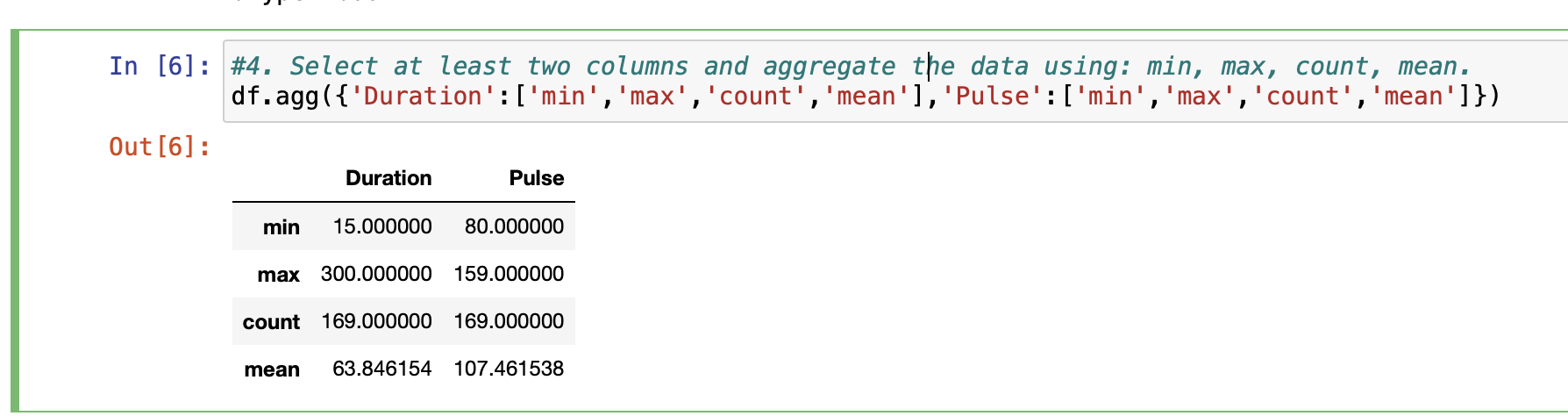
**Solution :**



**Explanation:** Checking if the data has the null values, replacing those null values.

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

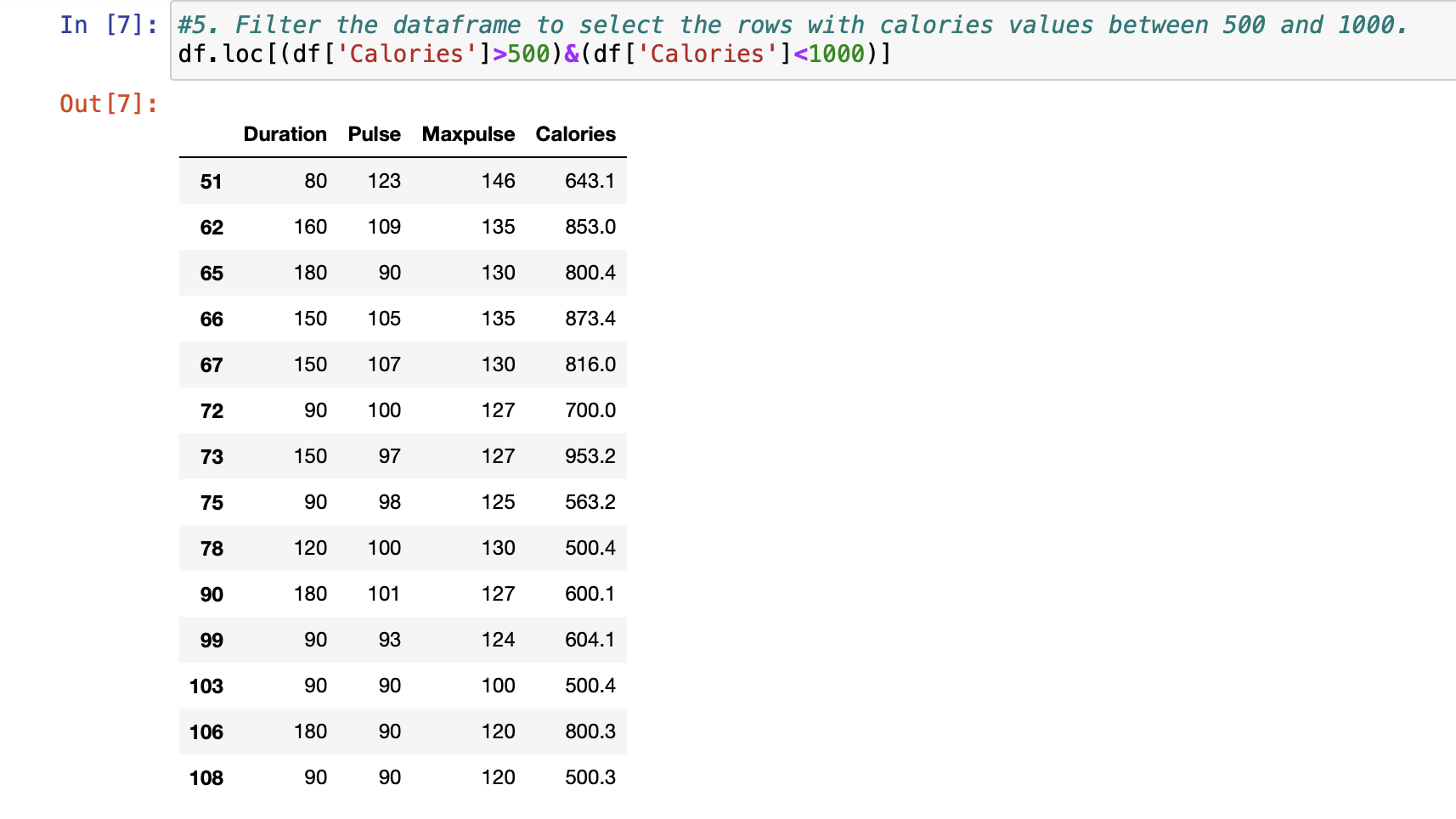
**Solution :**



**Explanation:** The two operation I have choosen are the duration and the pulse and aggregated the values.

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

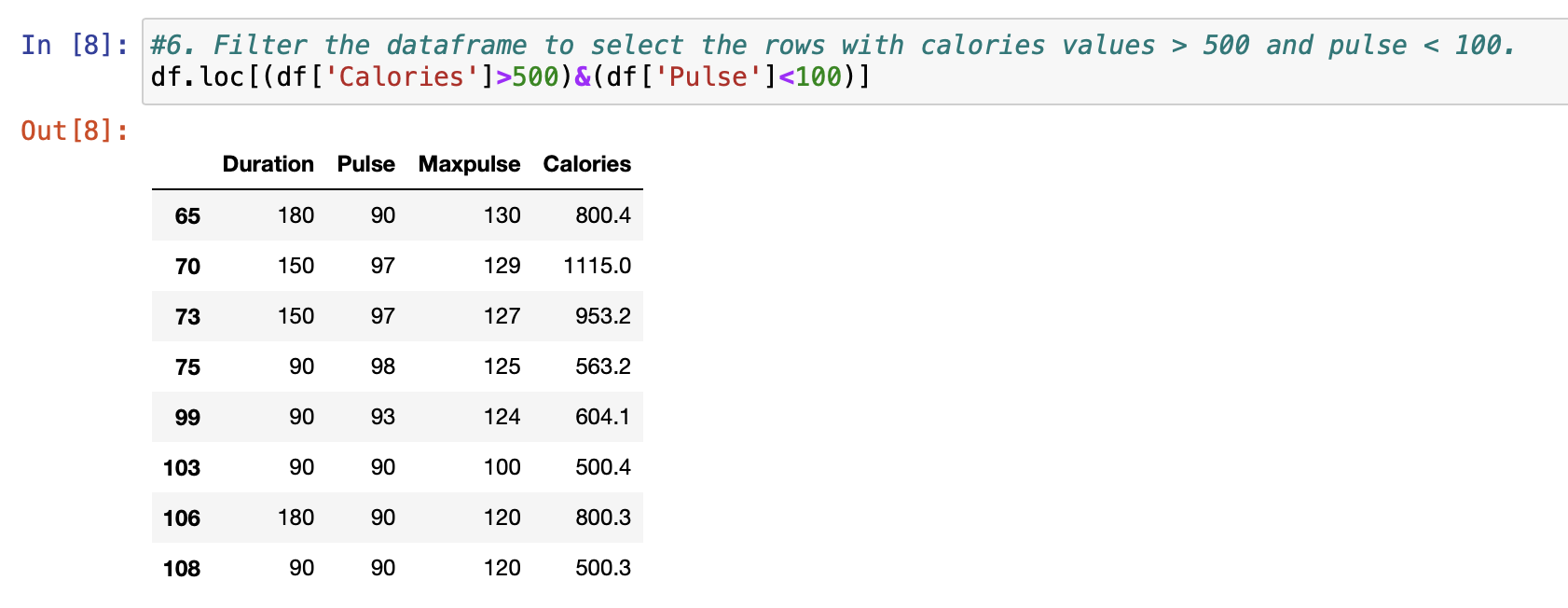
**Solution :**



**Explanation:**The calories data between 500 and 1000

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

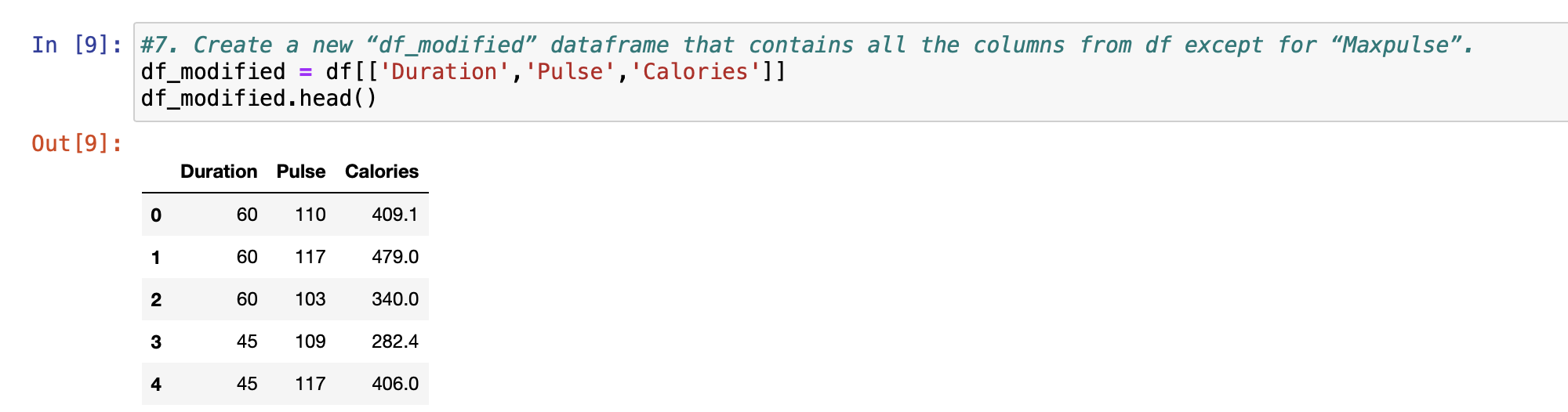
**Solution :**



**Explanation**: The calories data between 500 and 100

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

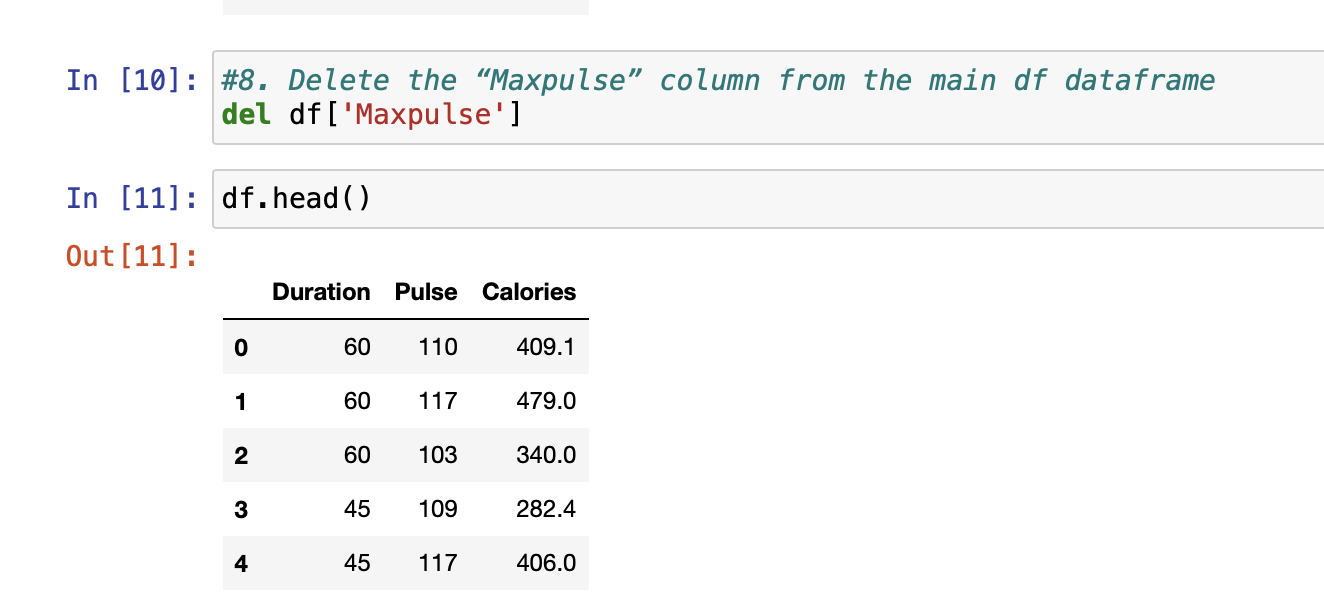
**Solution :**



**Explanation:** Created a data except the Maxpulse

* Delete the “Maxpulse” column from the main df dataframe

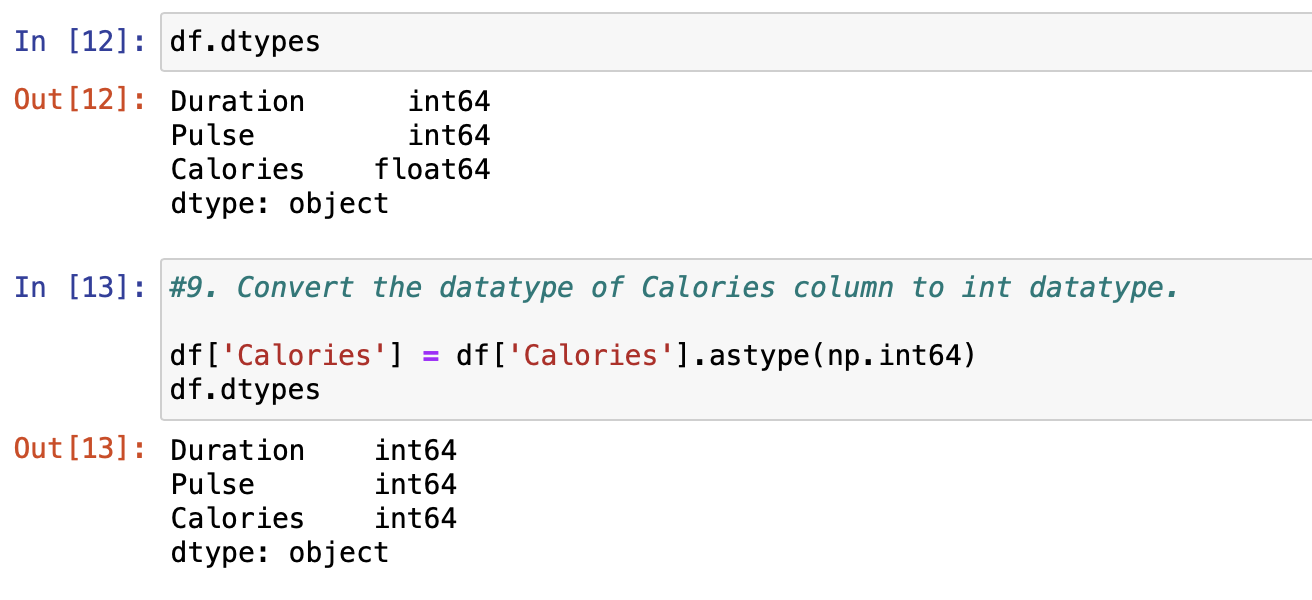
**Solution :**



**Explanation:** del deletes the Maxpulse.

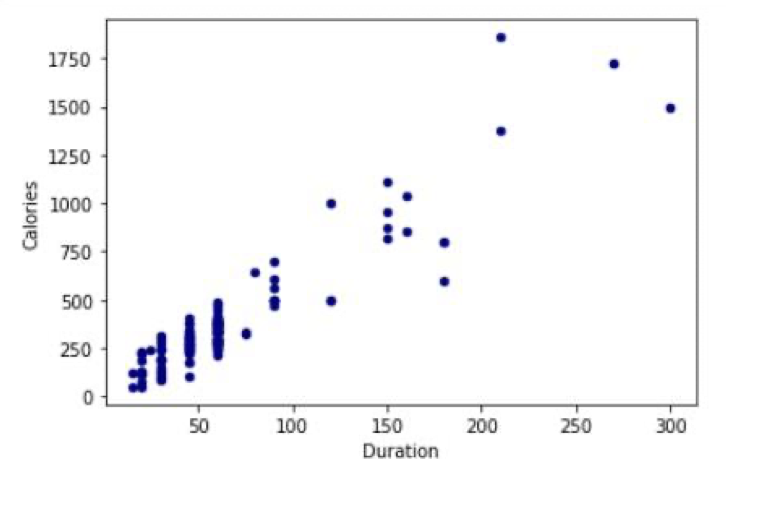
* Convert the datatype of Calories column to int datatype.

**Solution :**

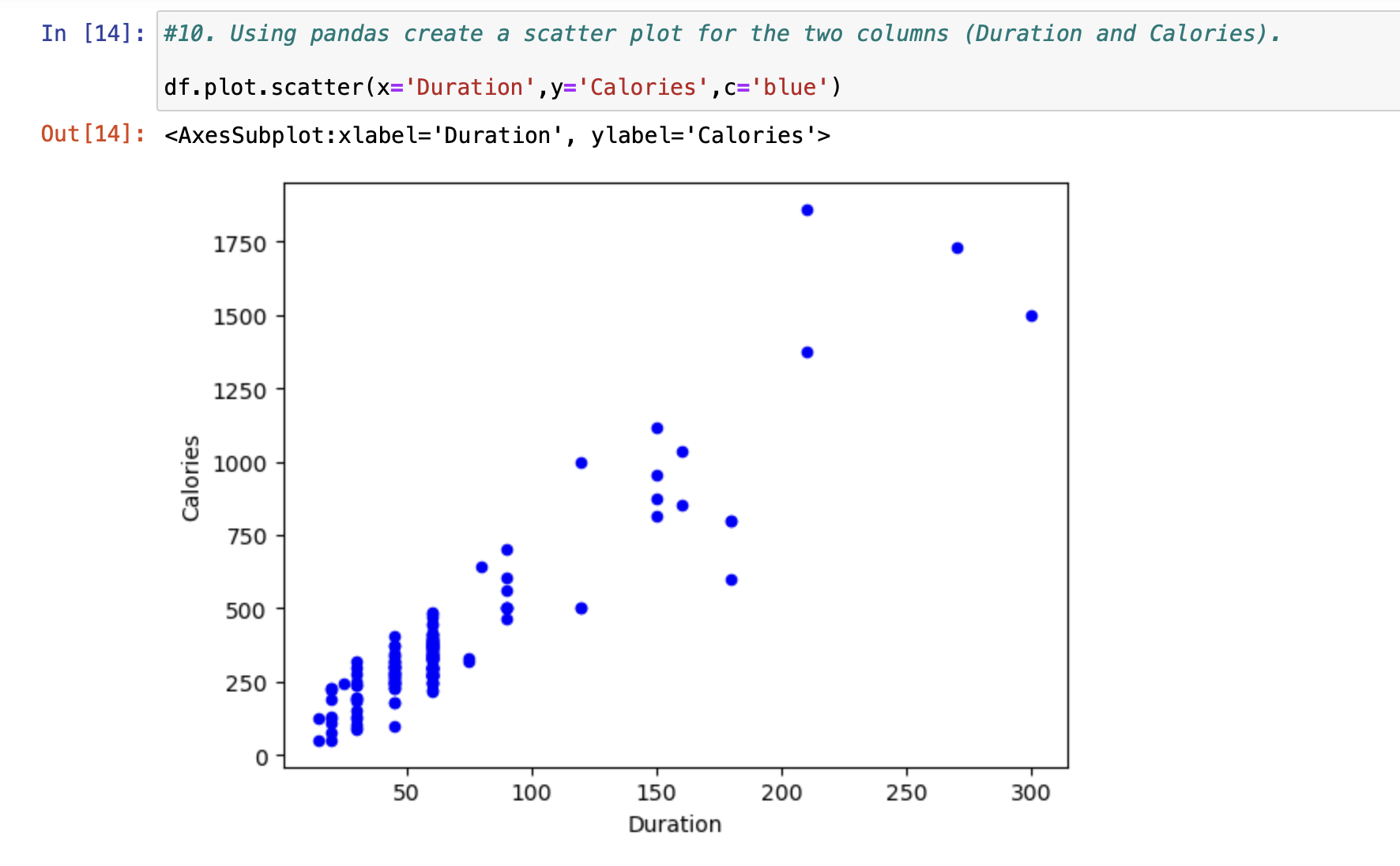


**Explanation:**Converting the datatype to int from float

* Using pandas create a scatter plot for the two columns (Duration and Calories).
  1. Example:



**Solution :**



**Explanation:**It is the scatter plot for two columns duration and the calories

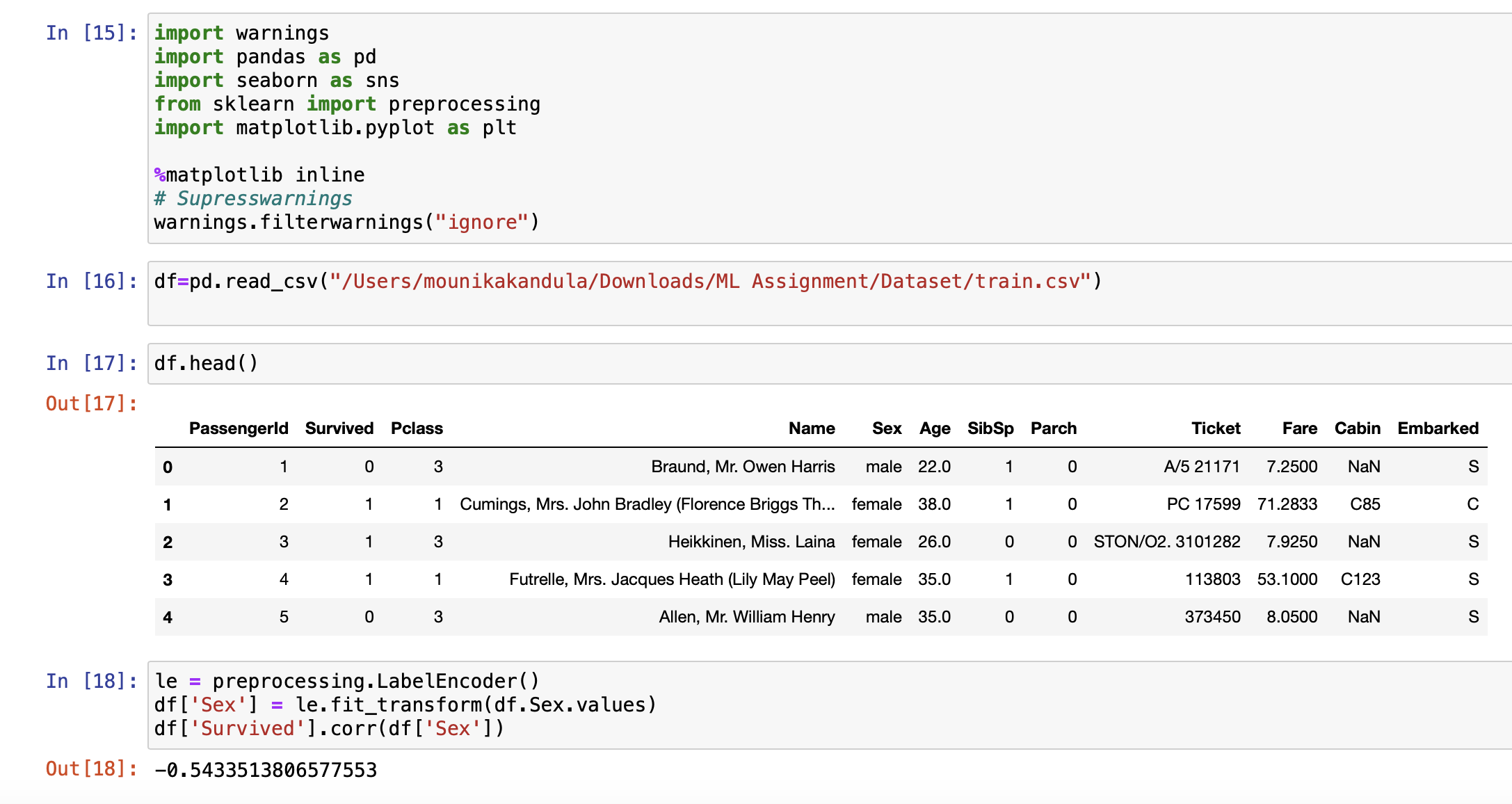
Question 2:

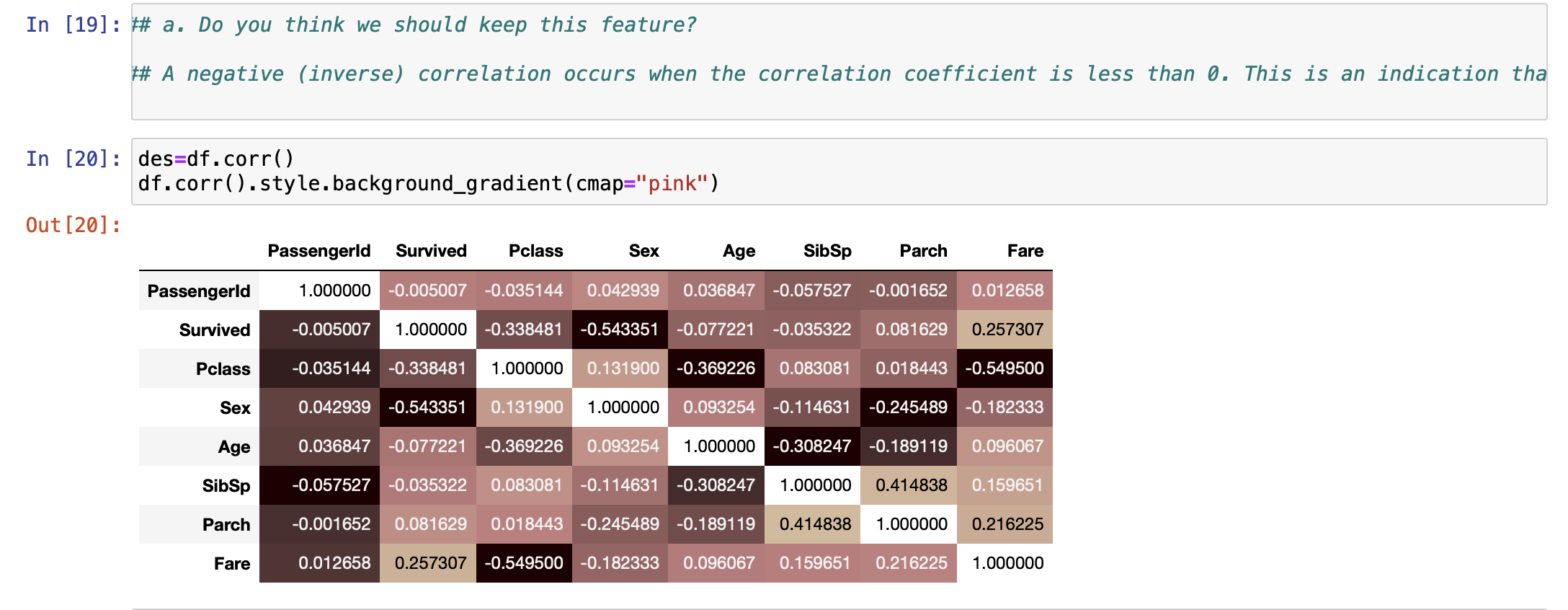
**(**Titanic Dataset**)**

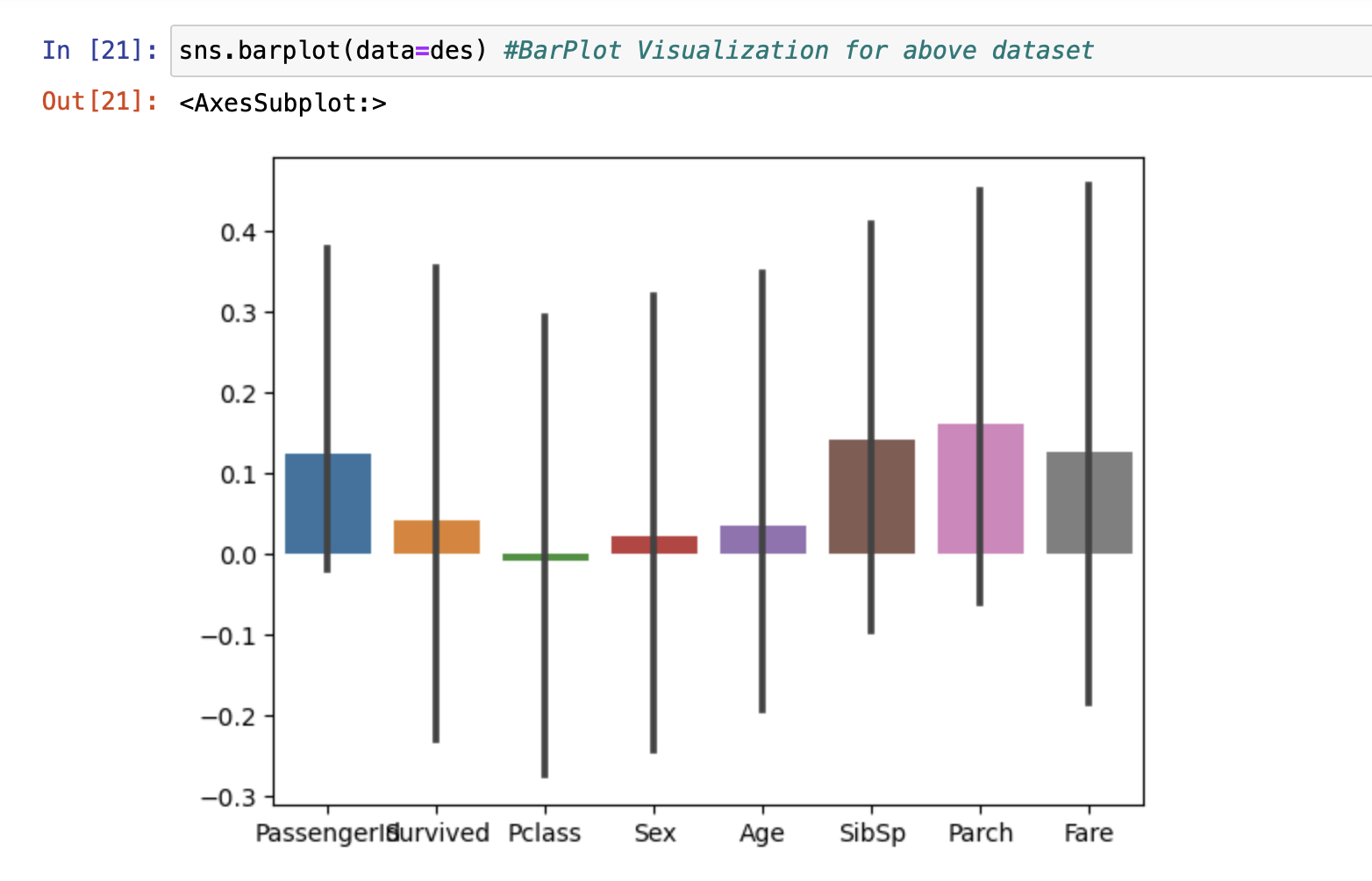
* Find the correlation between ‘survived’ (target column) and ‘sex’ column for the Titanic use case in class.

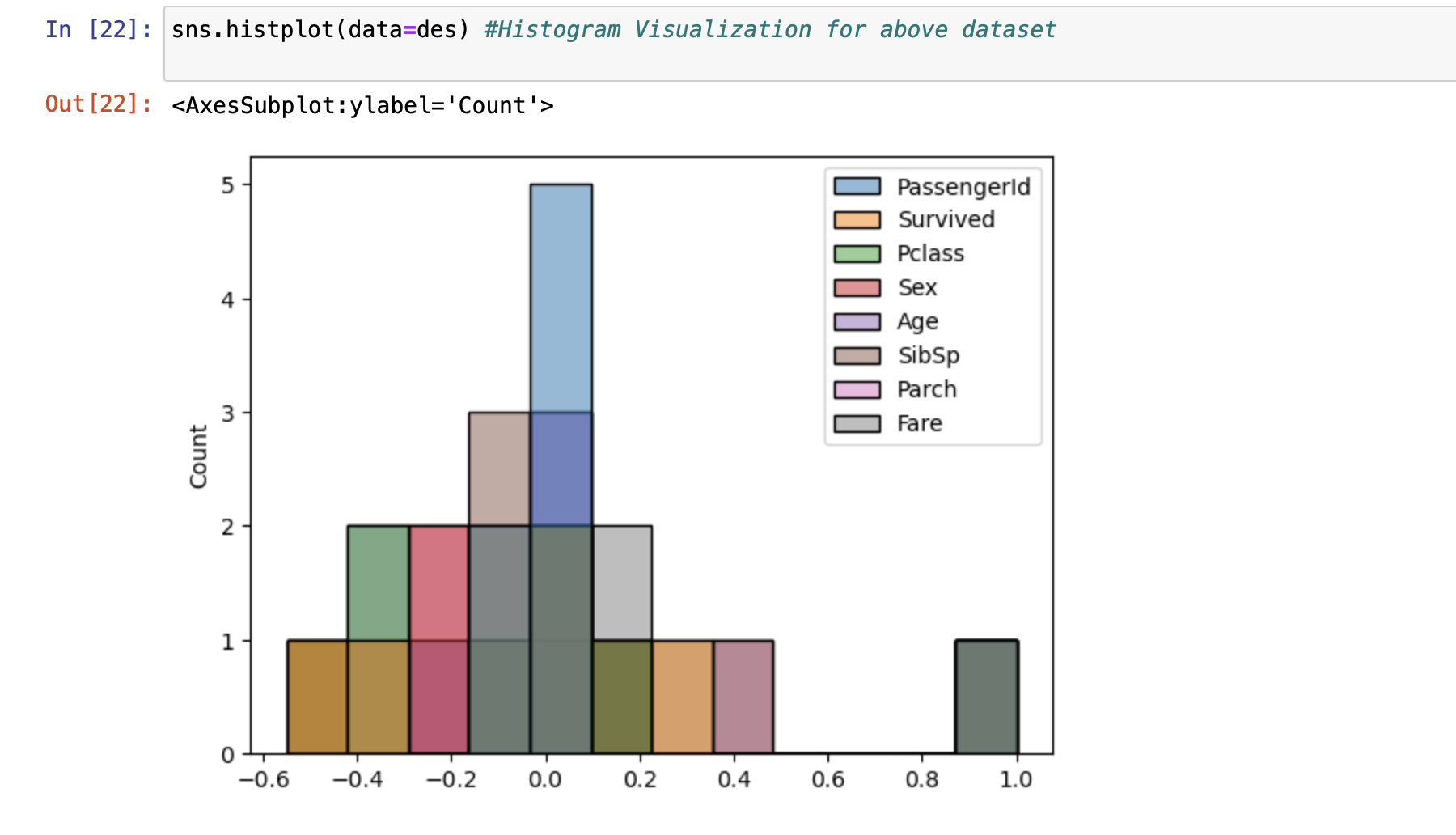
a. Do you think we should keep this feature?

* Do at least two visualizations to describe or show correlations.
* Implement Naïve Bayes method using scikit-learn library and report the accuracy.

**Solution :**  


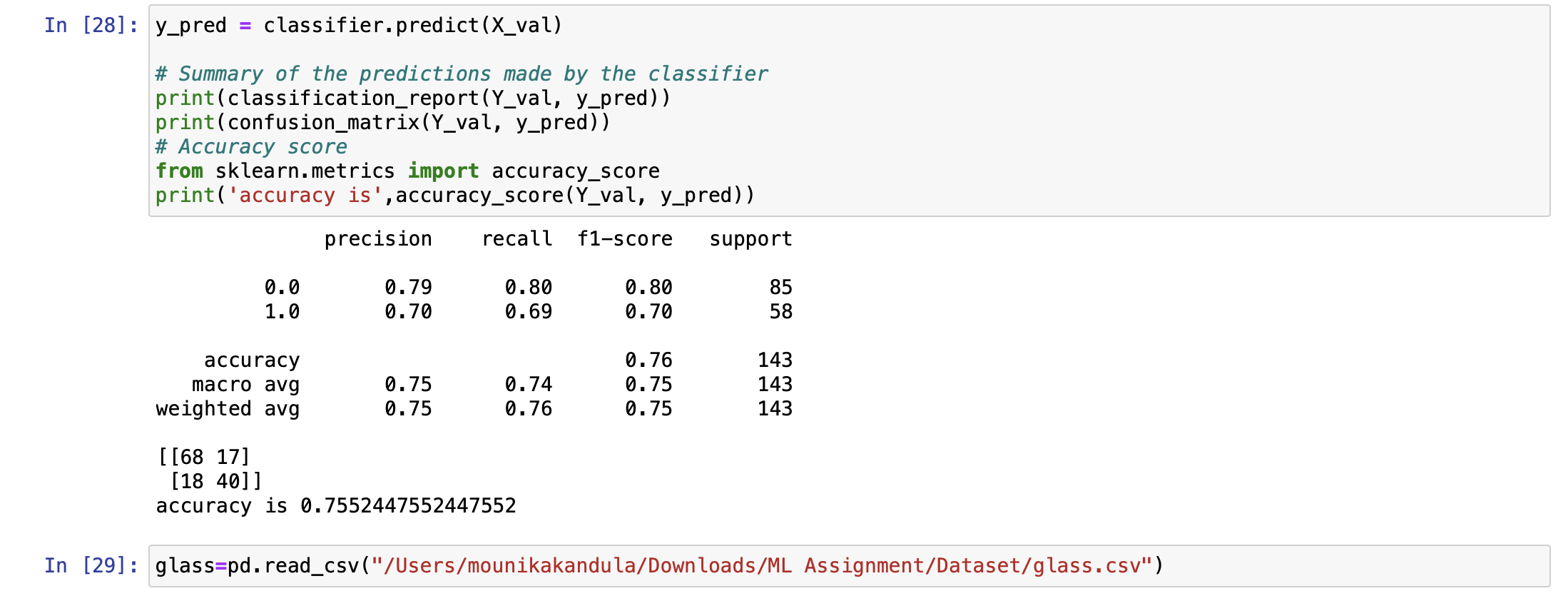








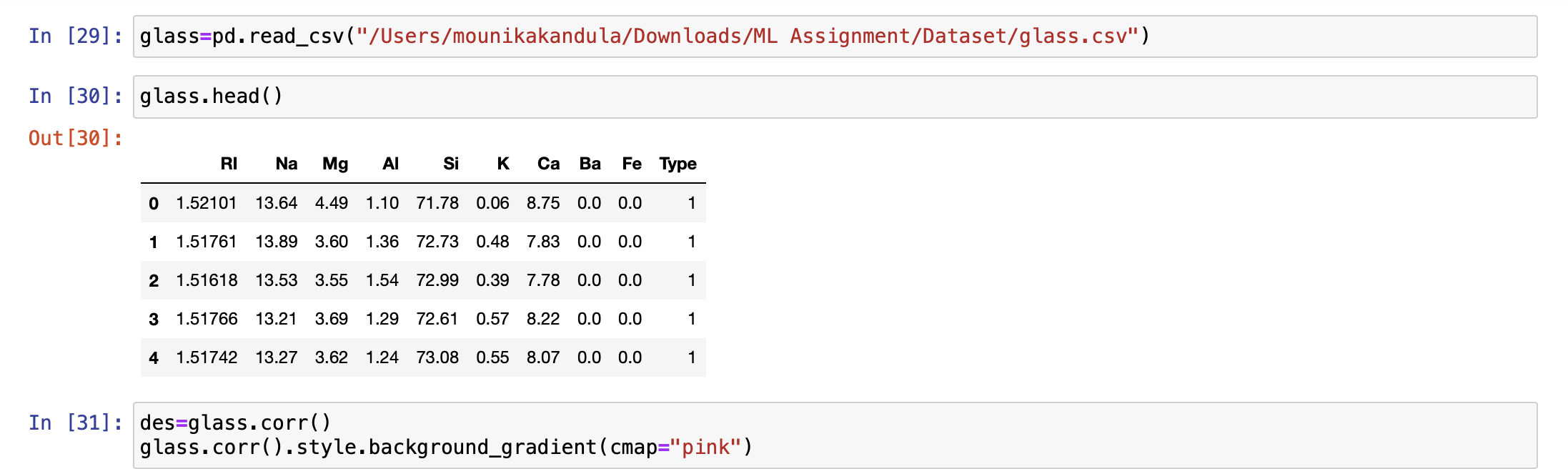


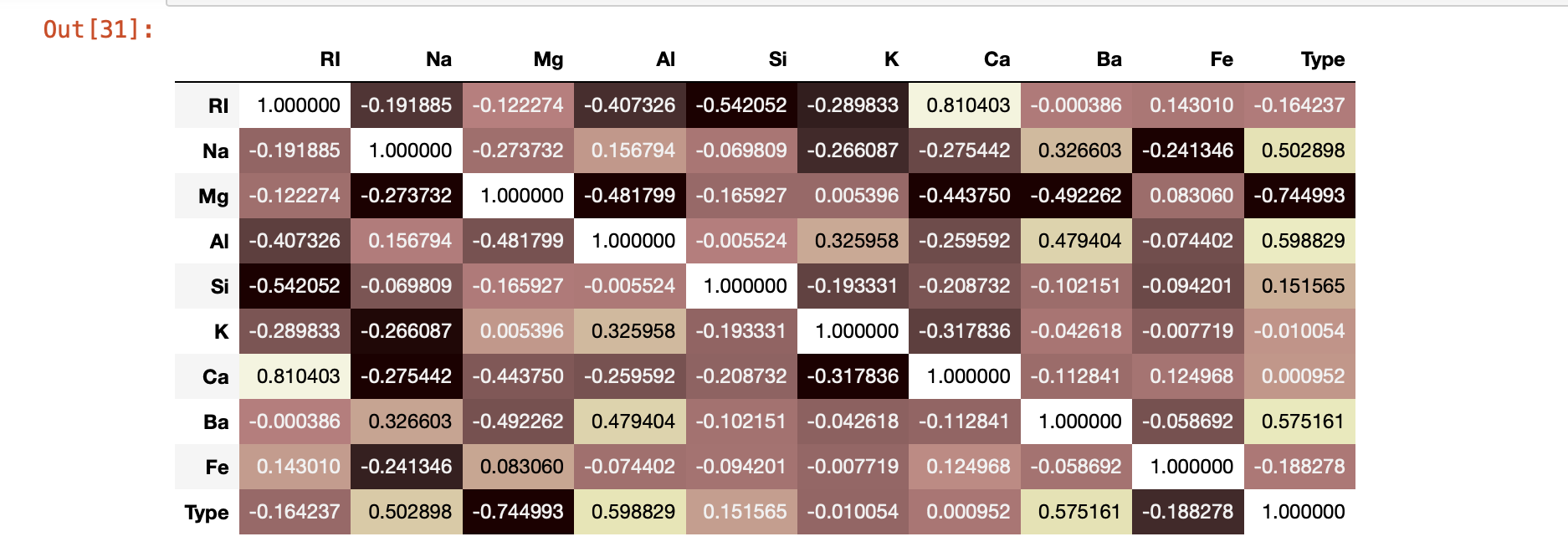


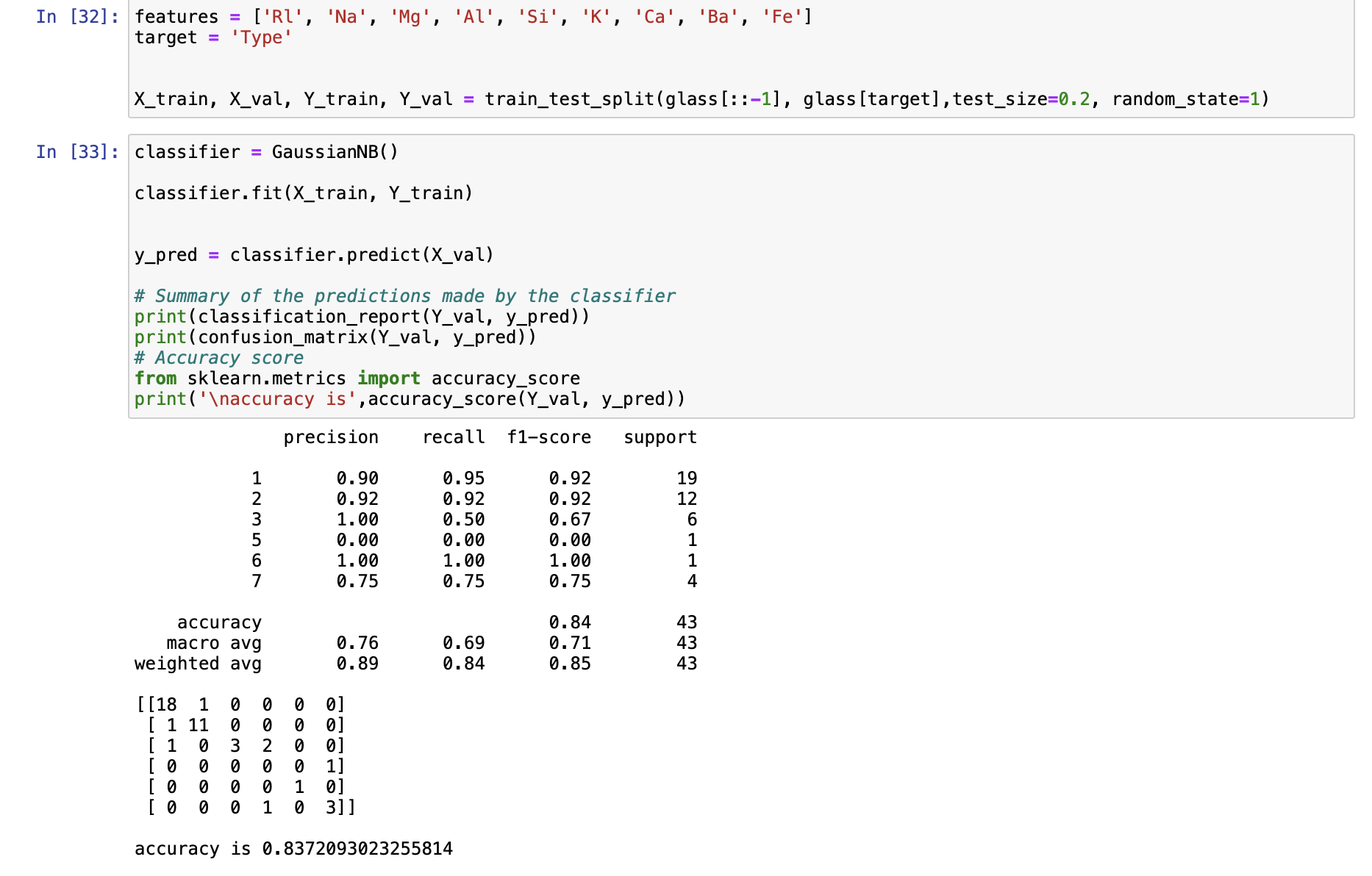
**(**Glass Dataset**)**

1. Implement Naïve Bayes method using scikit-learn library.
   1. Use the glass dataset available in Link also provided in your assignment.
   2. Use **train\_test\_split** to create training and testing part.
2. Evaluate the model on testing part using score and

**Solution :**





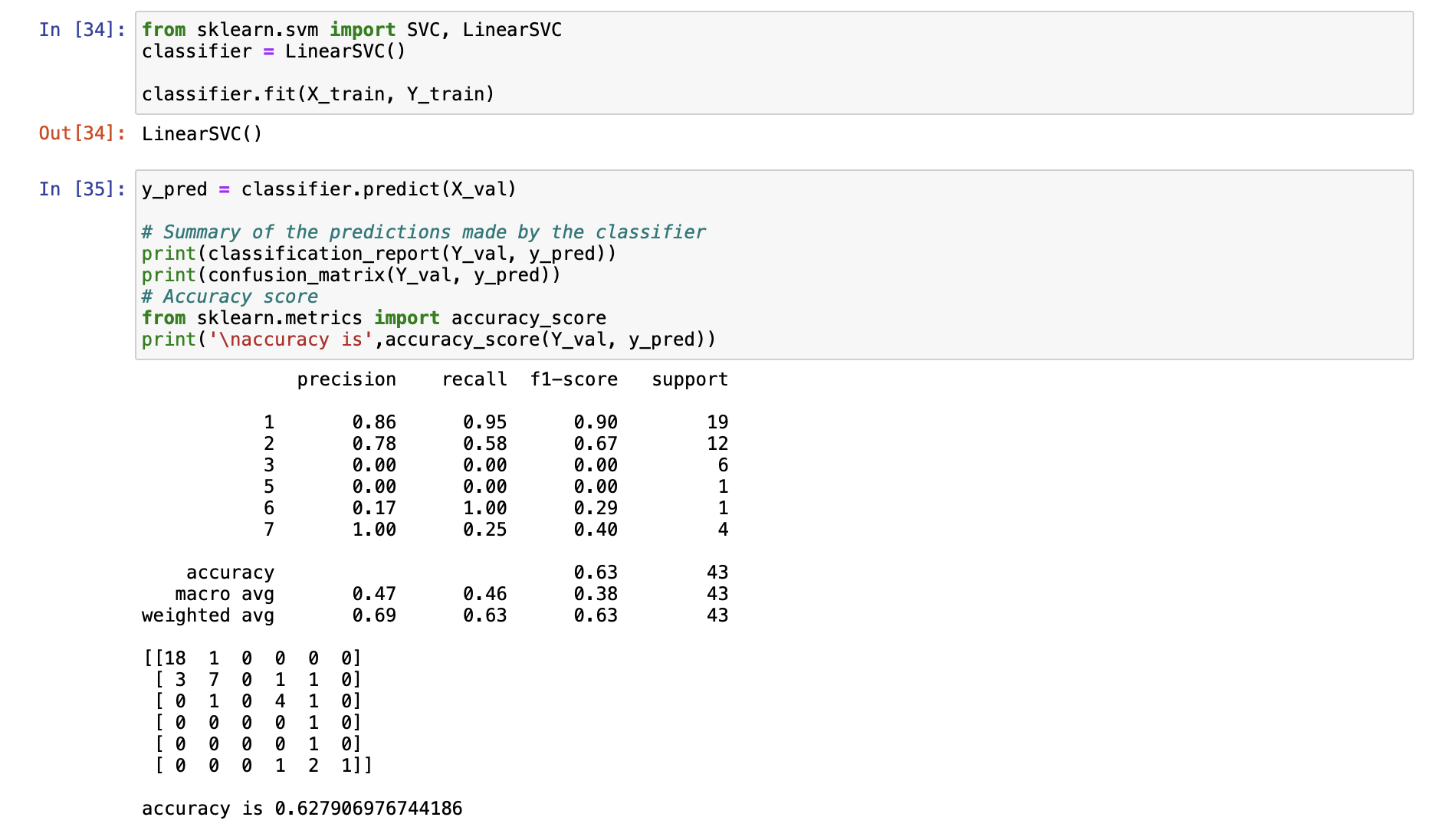


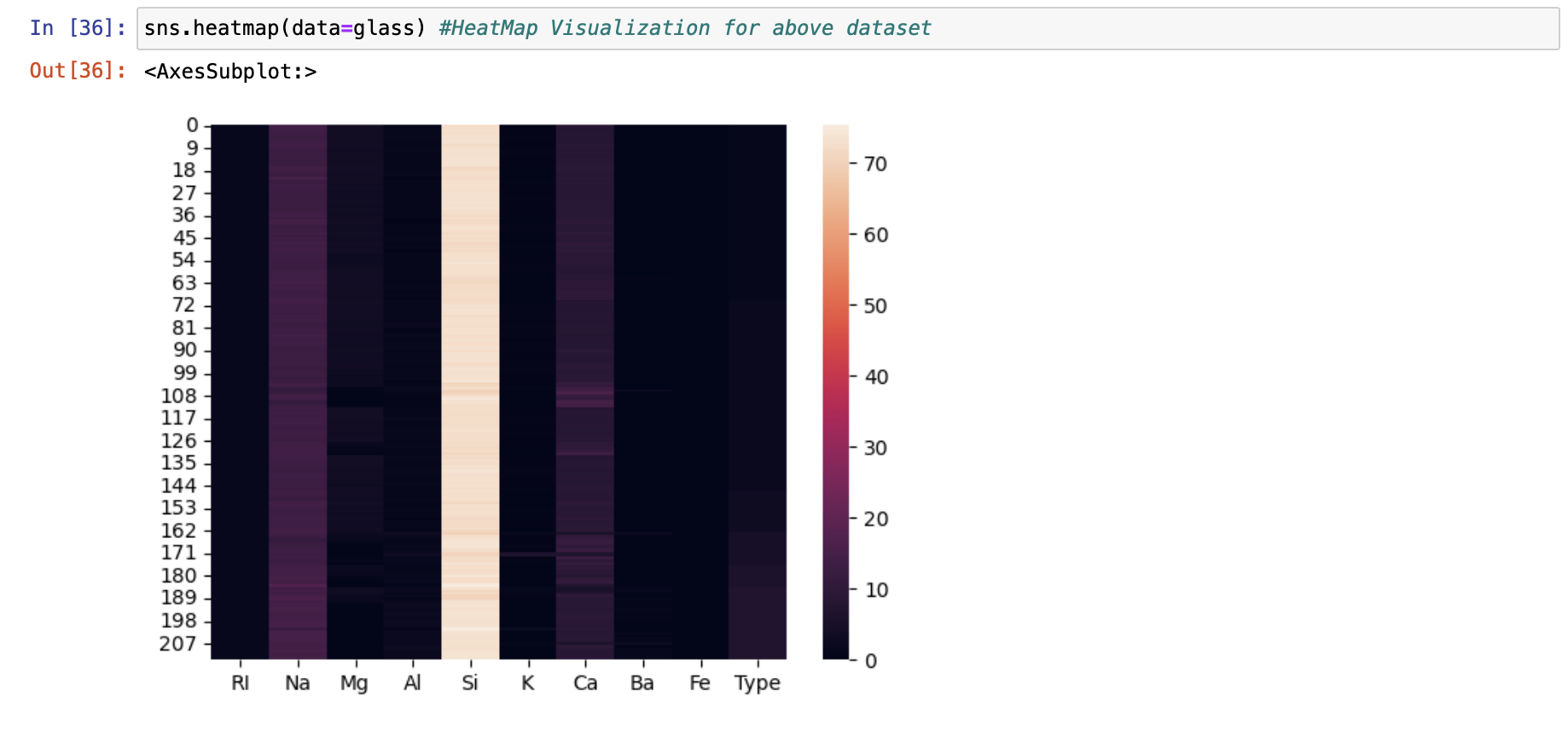
classification\_report(y\_true, y\_pred)

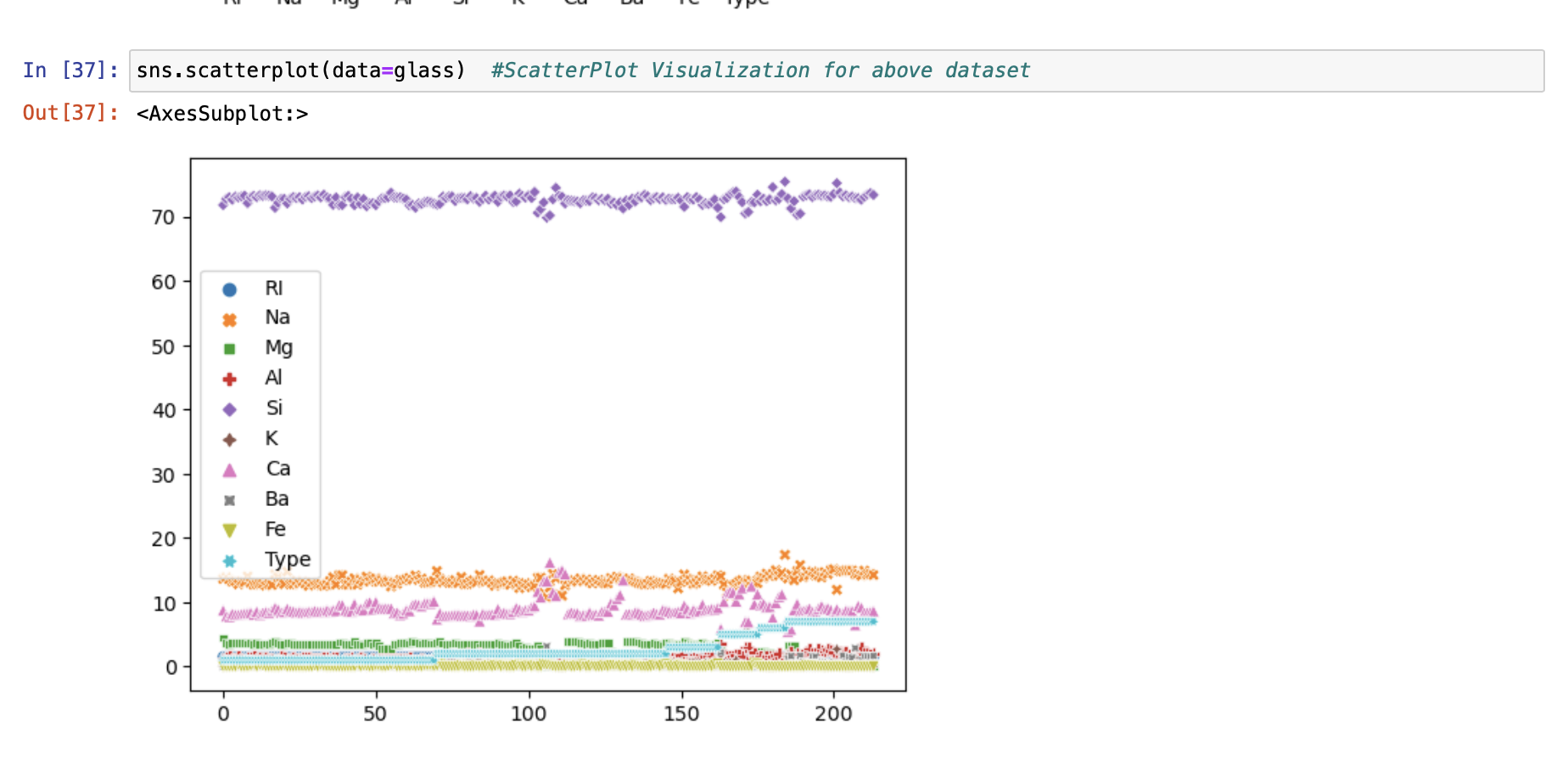
1. Implement linear SVM method using scikit library
   1. Use the glass dataset available in Link also provided in your assignment.
   2. Use **train\_test\_split** to create training and testing part.
2. Evaluate the model on testing part using score and

classification\_report(y\_true, y\_pred)

**Solution :**







Do at least two visualizations to describe or show correlations in the Glass Dataset. Which algorithm you got better accuracy? Can you justify why?

# Justification:

We got better accuracy for Naïve Bayes method which is 0.8372093023255814. Naive Bayes analysis works well with probabilistic concepts where as Linear SVM works better with linear regression logics. But to perform more accurately SVM requires large amounts of data to train and test the data. So, due to the amount of data Naive Bayes algorith gives better accuracy compared to Linear SVM.