# LAB1 – Deep Learning with Python CS5590 Module 1

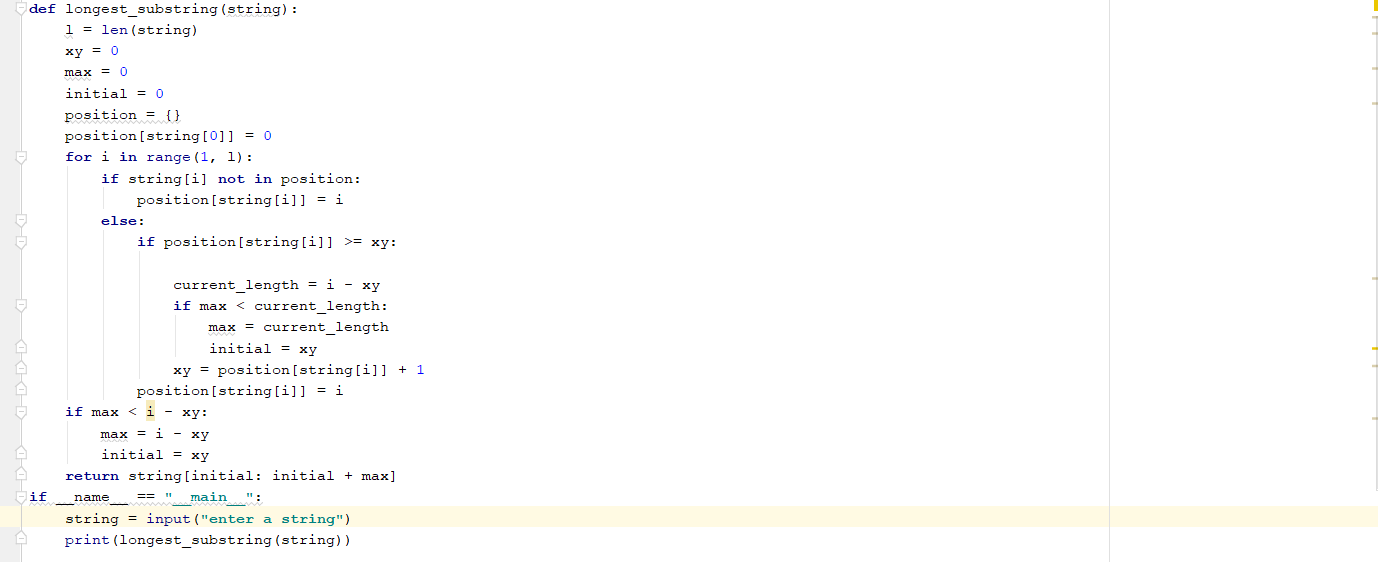
## Team Members: Ahmed Siddiqui, Maseerah Muradabadi, Vamsi Draksharam

1. Objective
2. Features
3. Steps (input, output, explain code)
4. Configurations
5. Limitations
6. References

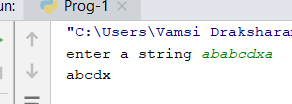
**Question 1: Write a program to find a longest substring without repeating characters from a given string input from the console.Sample Input: ‘ababcdxa’Sample Output: abcdx.**

Solution:

**Code Steps:**

****

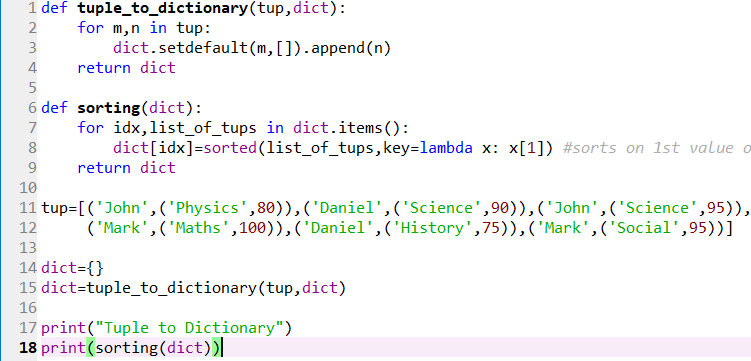
**Output:**

****

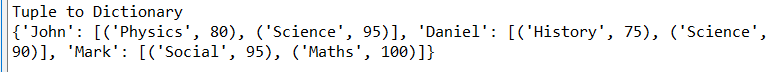
**Question 2: Suppose you have a list of tuples as follows:[( ‘John’, (‘Physics’, 80)) , (‘ Daniel’, (‘Science’, 90)), (‘John’, (‘Science’, 95)), (‘Mark’,(‘Maths’, 100)), (‘Daniel’, (’History’, 75)), (‘Mark’, (‘Social’, 95))]Create a dictionary with keys as names and values as list of (subjects, marks) in sorted order.{John : [(‘Physics’, 80), (‘Science’, 95)]Daniel : [ (’History’, 75), (‘Science’, 90)]Mark : [ (‘Maths’, 100), (‘Social’, 95)]}**

Solution:

**Code Steps:**

****

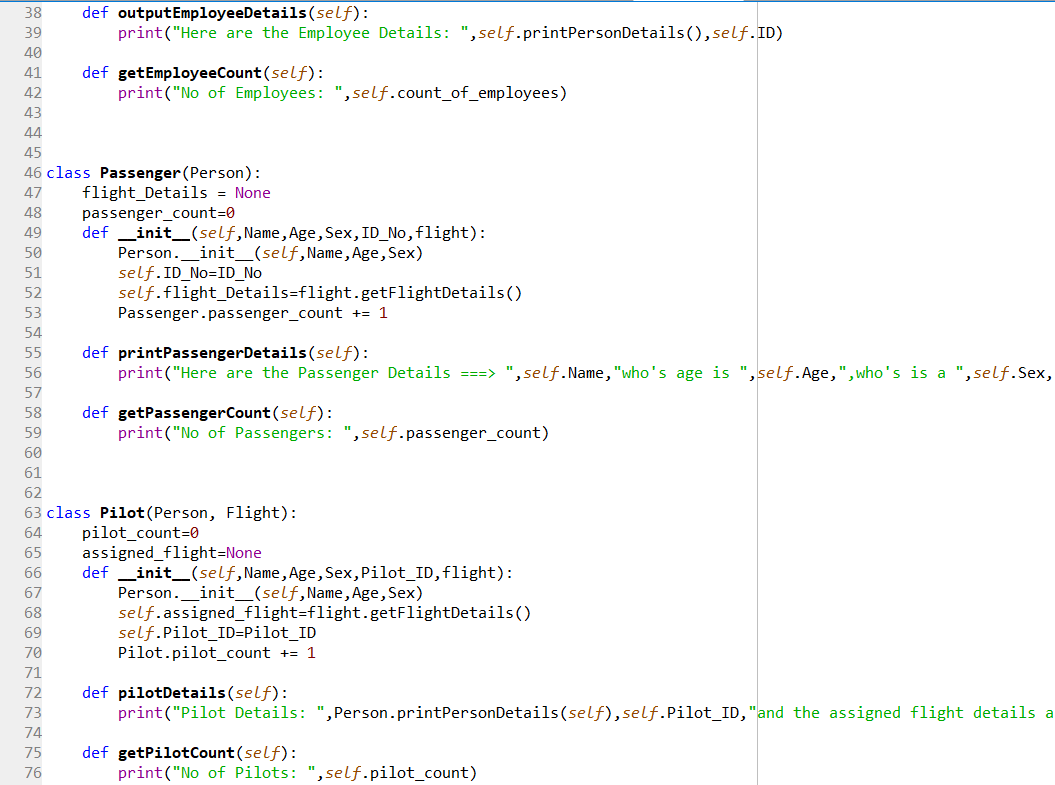
**Output:**

****

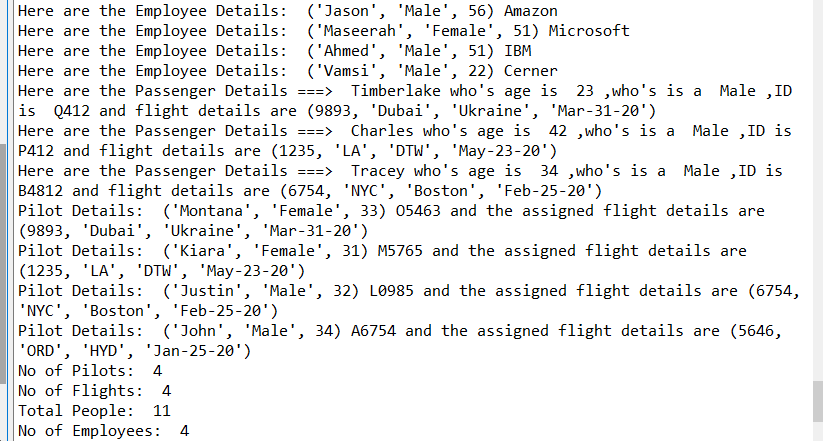
**Question 3: Write a python program to create any one of the following management systems. 1. Airline Booking Reservation System (e.g. classes Flight, Person, Employee, Passenger etc.)2. Library Management System(eg: Student, Book, Faculty, Department etc.).**

Solution:

**Code Steps:**

**** ****   

**Output:**

****

**Question 5: Pick any dataset from the dataset sheet in the class sheet or online which includes both numeric and non-numeric features a. Perform exploratory data analysis on the data set (like Handling null values, removing the features not correlated to the target class, encoding the categorical features, ...) b. Apply the three classification algorithms Naïve Bayes, SVM and KNN on the chosen data set and report which classifier gives better result.**

Solution:

**Objective:** Perform EDA and fit classification algorithms like Naïve Bayes, KNN and SVM

**Features:**  We are using the titanic.csv ass our data set, our target is ‘Survived’ and we have features like PassengerId, Name, Age, Sex etc.

**Code Steps:**

Step 1: Import libraries that are needed

Step 2: Read the data

Step 3: Sum up al the null values from all columns and print it

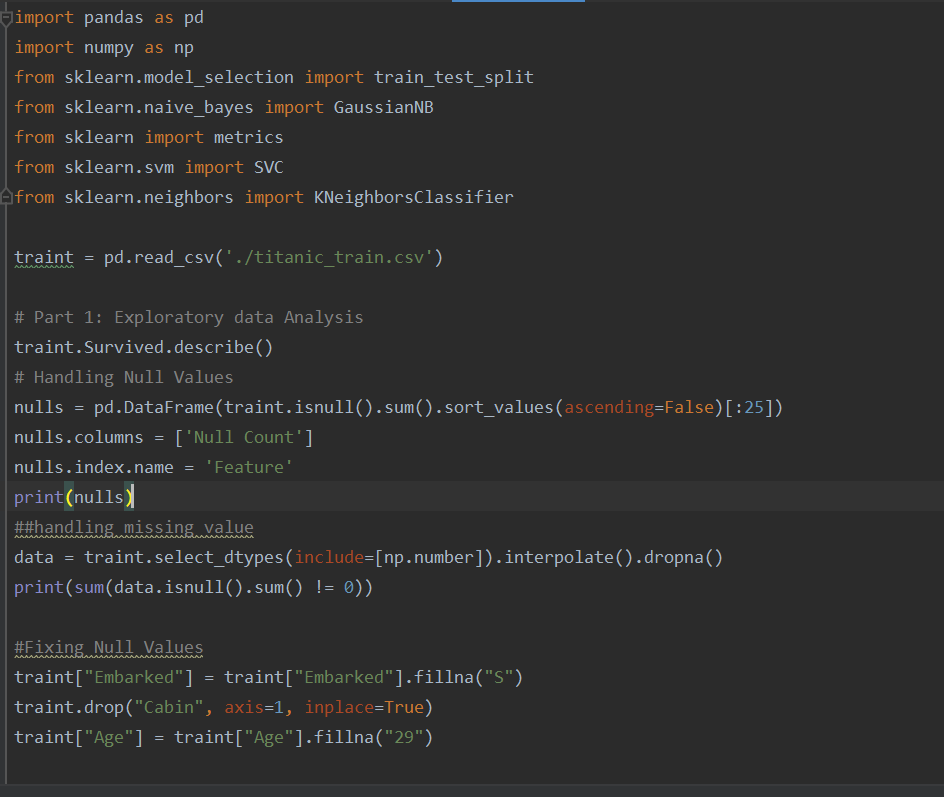
Step 4: Drop all null values

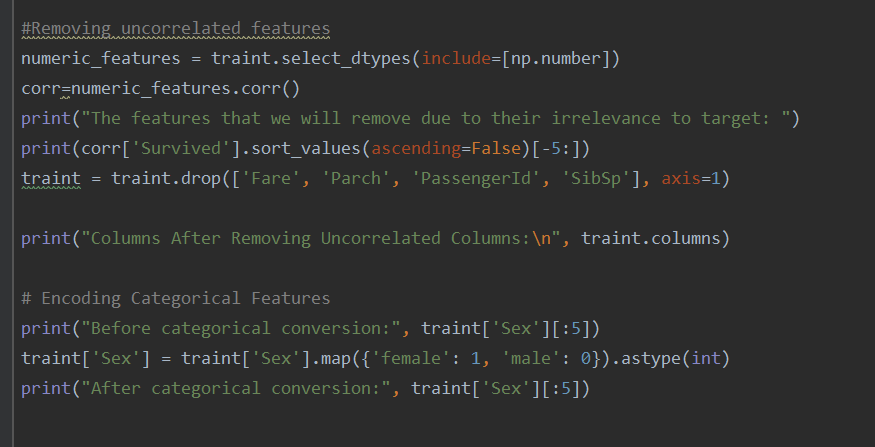
Step 5: Since columns Embarked, Cabin and Age have null values we can manipulate Embarked and Age by filling missing values with a mean value and drop Cabin column

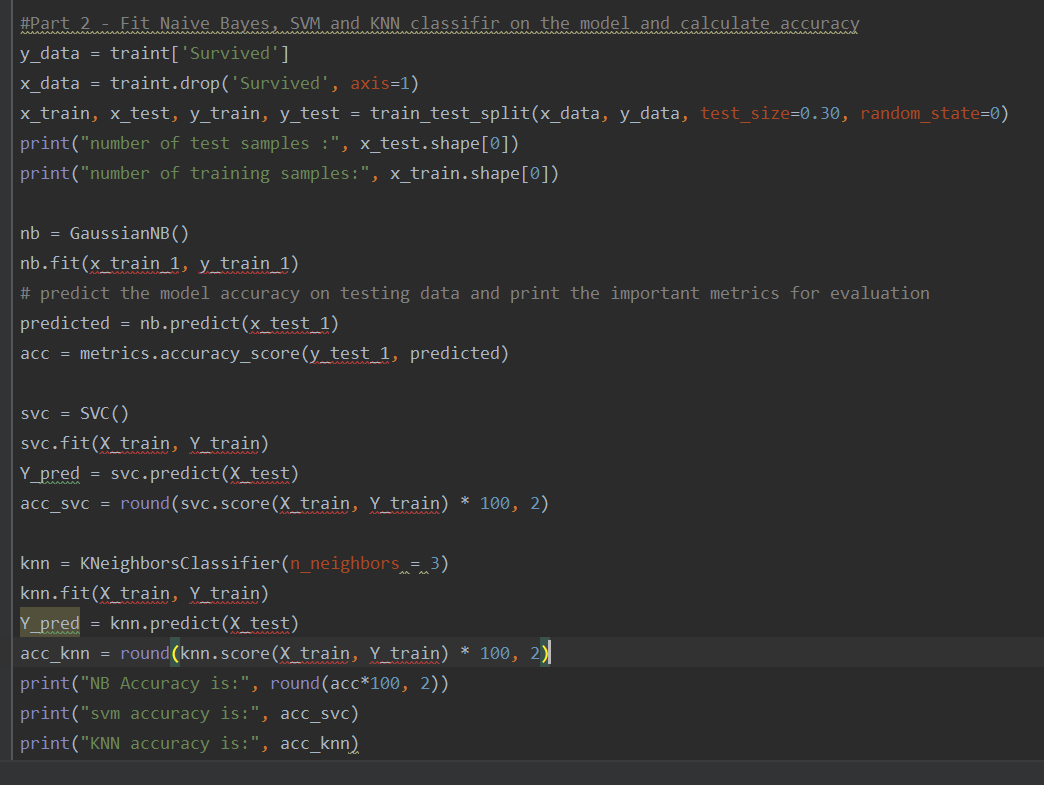
Step 6: Find the most and least correlated features to target ‘Survived’ and drop the least correlated.

Step 7: For column Sex convert categorical row values male and female to numeric int values 0 and 1

Step 8: Fit models naïve bayes, knn and svm by calling the model and then fitting it on traing data and predicted the model on testing data and finally calculating the accuracy using metrics.accuracy\_score()

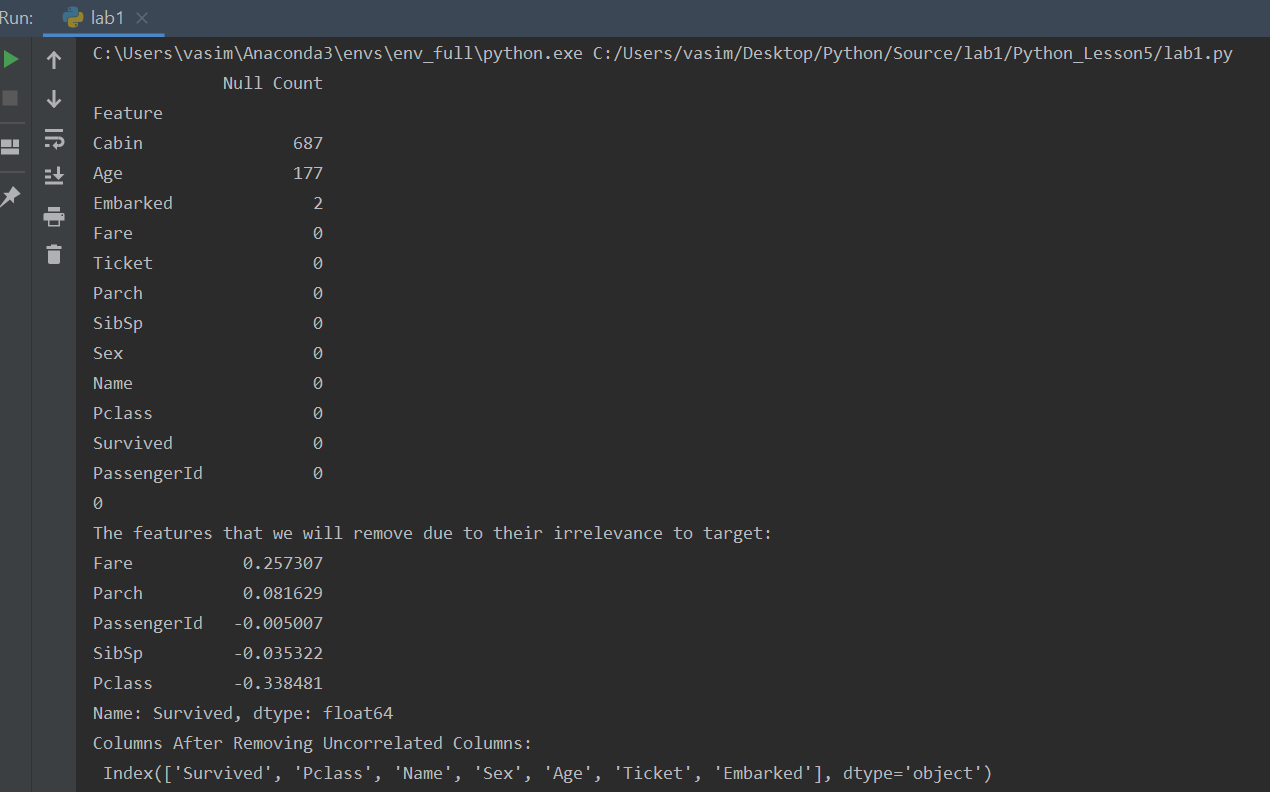




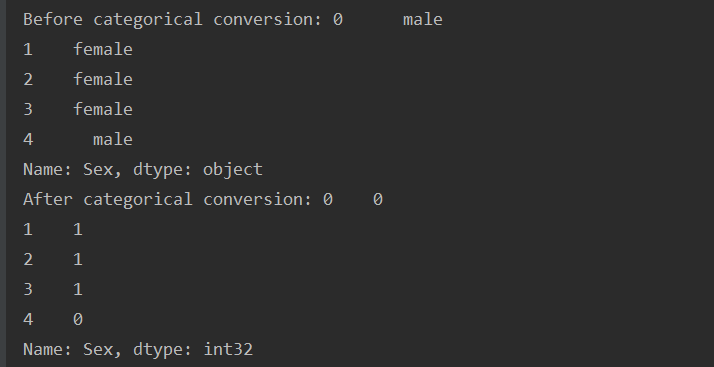


**Output:**

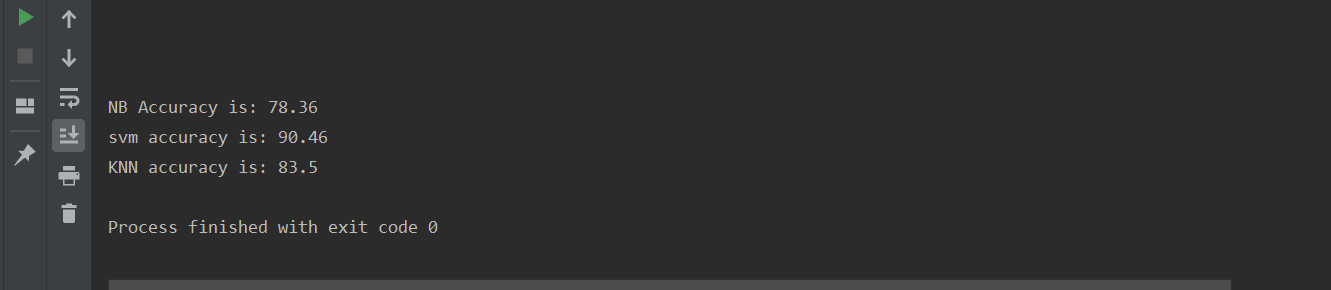
Showing the null sum count from each column



Values before and after conversion:



SVM outperformed NB and KNN



**Configurations:** Check data set after each step

**Limitations:** No limitations

**References:**

1. Class slides

2. In class programming code

**Question 6: Choose any dataset of your choice. Apply K-means on the dataset and visualize the clusters using matplotlib or seaborn.a. Report which K is the best using the elbow method.b. Evaluate with silhouette score or other scores relevant for unsupervised approaches (before applying clustering clean the data set with the EDA learned in the class)**

Solution:

**Objective:** To clean the data, perform K means clustering and visualize it

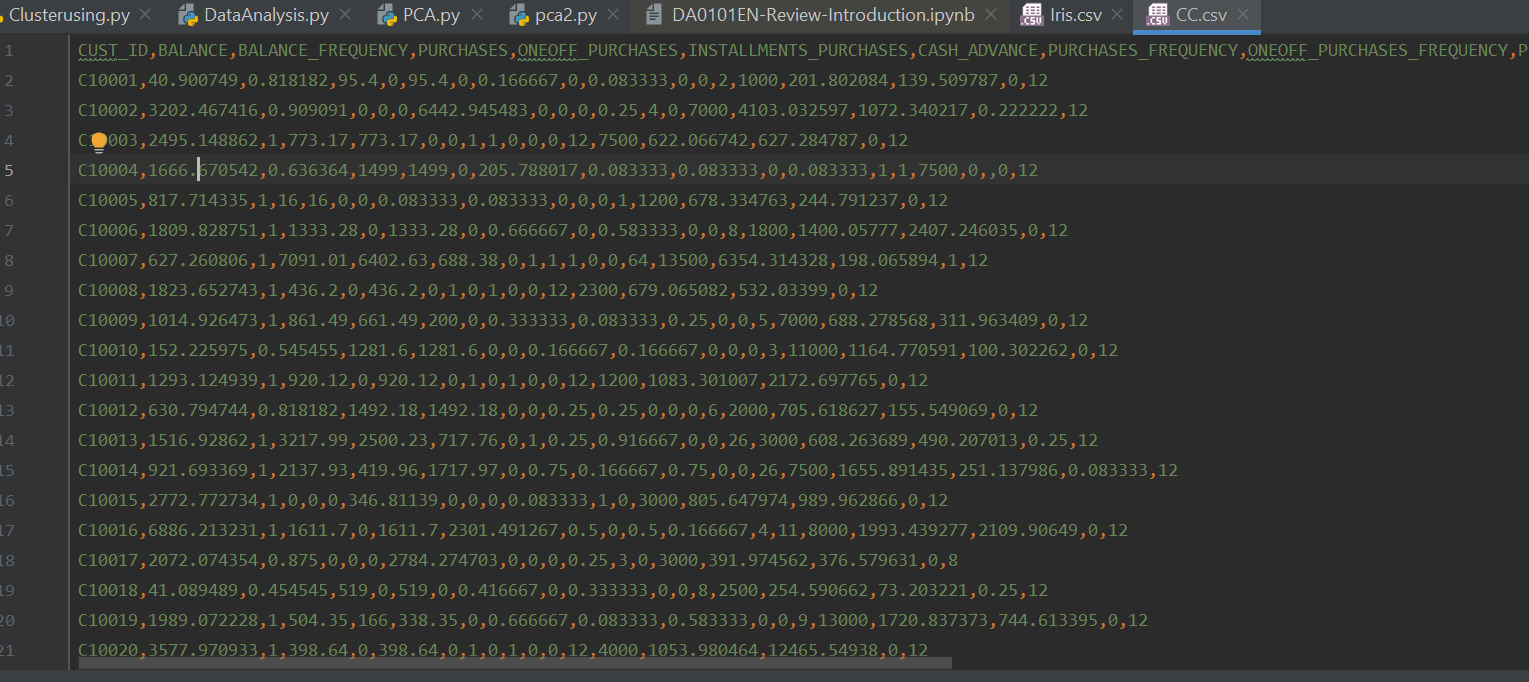
**Features:**

Let us understand the ‘CC.csv’ data set that was given in class lesson 6 for clustering problem

In the data we see 17 rows that are customer id, balance, purchases, installments purchases, cash advance, credit limit, purchase transactions etc

These columns are contributing some knowledge about the target ‘tenure’

Tenure generally has many meanings but according to this data given it means ‘reverse mortgage’



Explaining our data:

Say an old customer of age 62 has a property of x dollars, now he will apply for a monthly loan by keeping his property in the bank. More technically: a reverse mortgage is a loan. A homeowner who is 62 or older and has considerable home equity can borrow against the value of their home and receive funds as a lump sum, fixed monthly payment or line of credit. Unlike a forward mortgage—the type used to buy a home—a reverse mortgage doesn’t require the homeowner to make any loan payments.

**Code Steps:**

Step 1: Import the needed libraries

Step 2: Read the data from csv file

Step 3: Using index by row function we take data into x\_train variable

Step 4: EDA – Count null values from all columns and print their sum

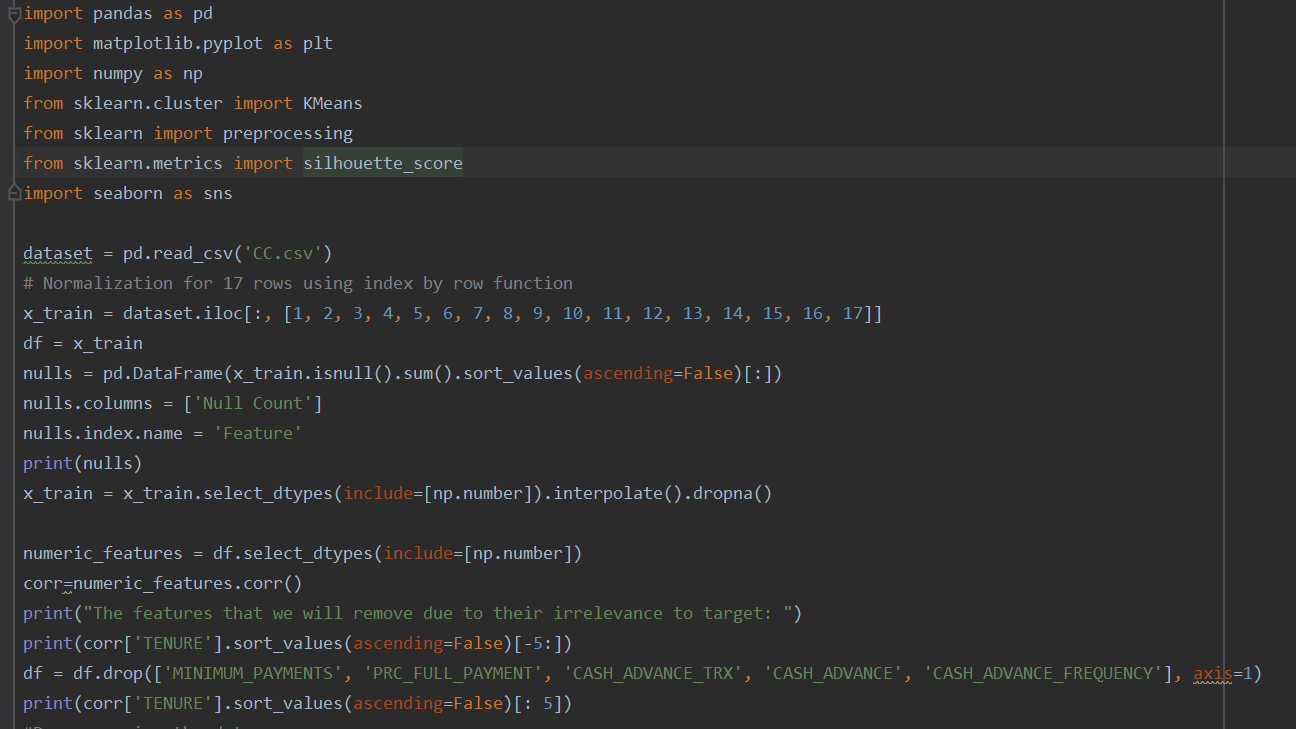
Step 5: Drop all rows that have null values

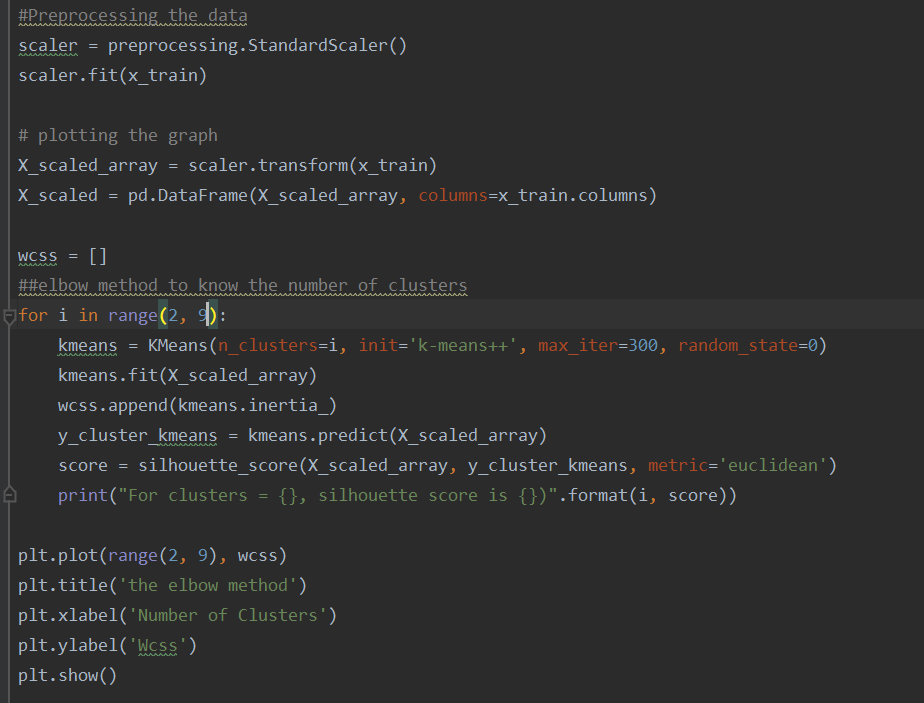
Step 6: Find the most and least correlated features to target ‘Tenure’ and drop the least correlated.

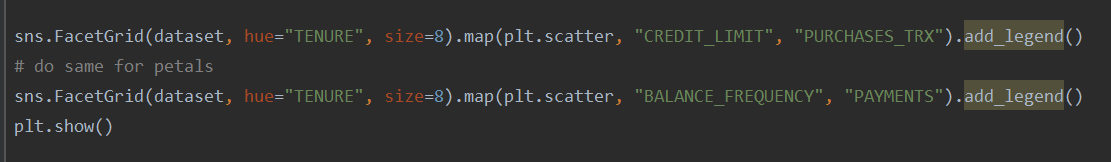
Step 7: Do feature scaling so that all features are in the same range and larger value features don’t dominate the rest using StandardScaler() and transform our data into a dataframe that is feature vectored already

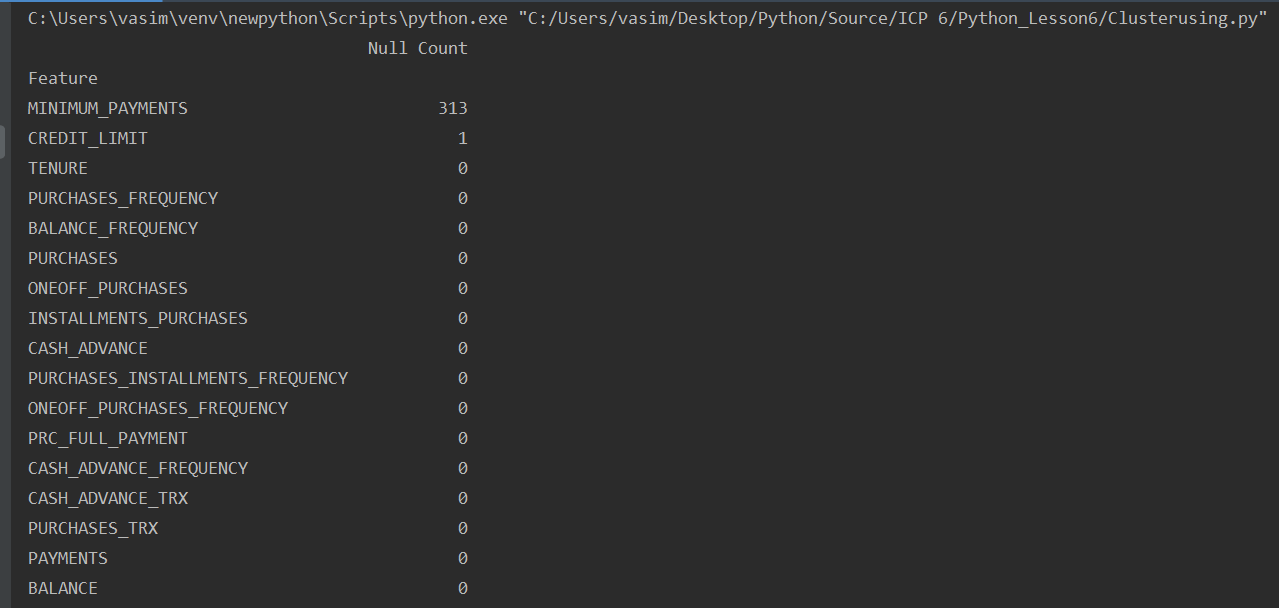
Step 8: Using list comprehension, loop through the the number of K clusters and calculate the Silhoutte Score, print them and plot the elbow method that shows the optimal k value

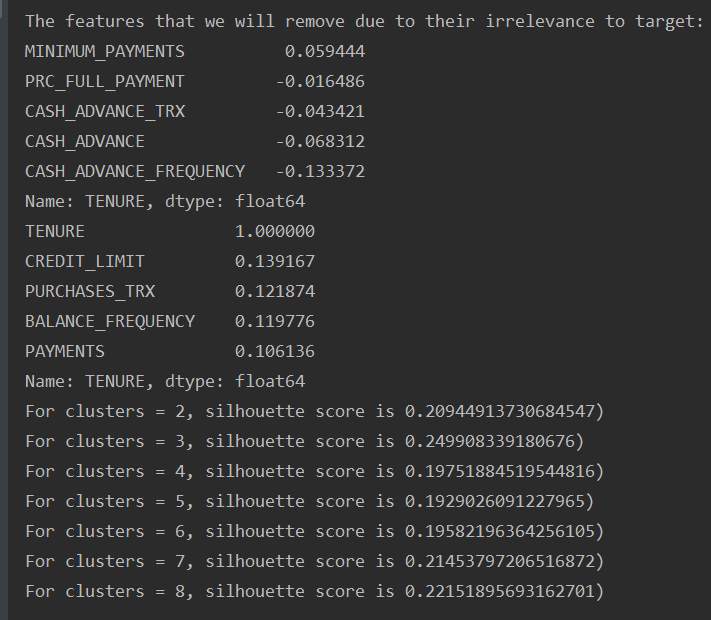
Step 9: using seaborn libraries use the top 4 correlated features and plot two graphs to get insights on the data. Pick the one that defines the clusters clearly and analyze it

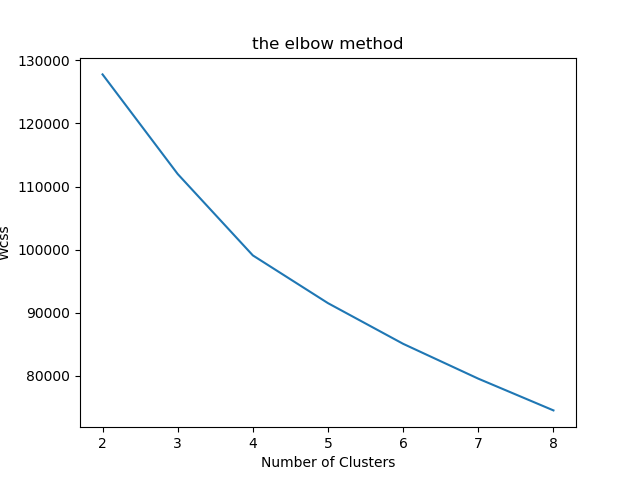




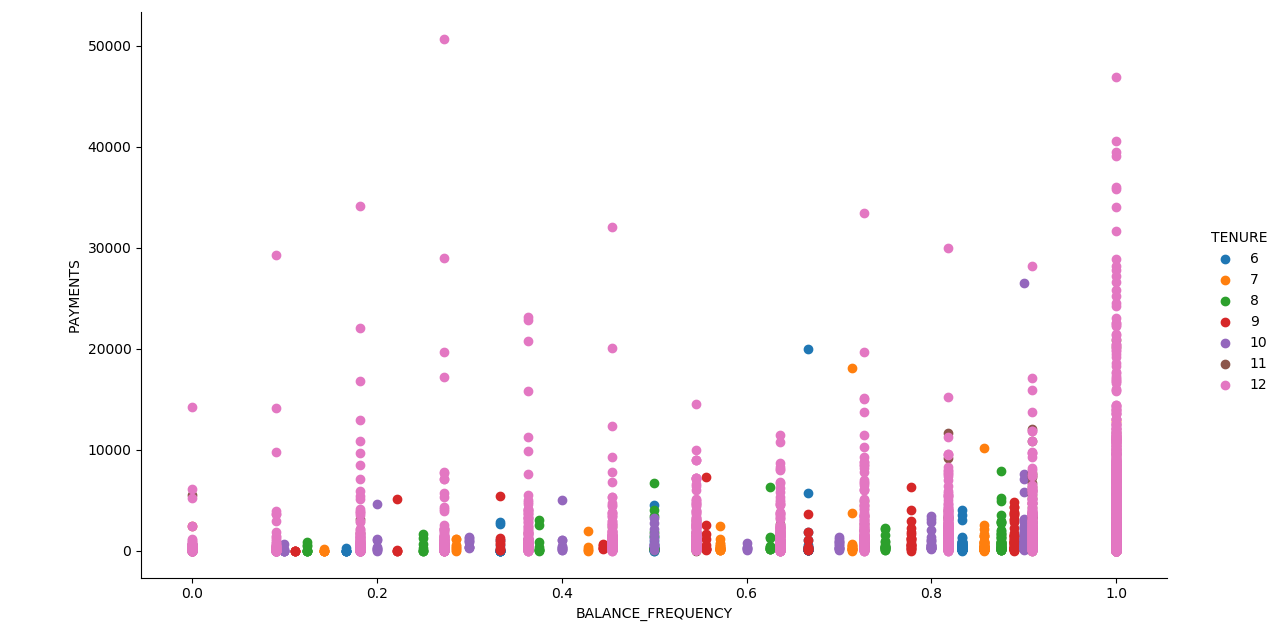


Output: 

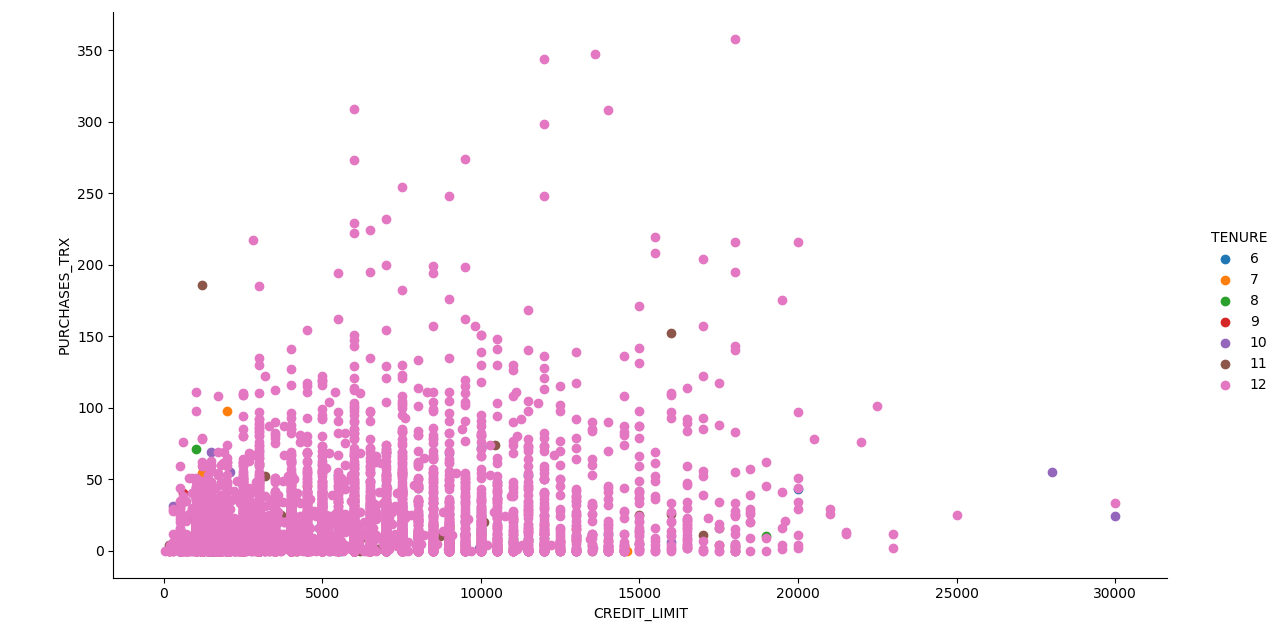




The optimal number of clusters is k=8



The above plot does not separate the data properly giving any meaning information about tenure



The graph above clearly shows the reasons when tenure can be approved for customer xyz

Findings:

Now based on our data set we found in the EDA part that features like credit limit, purchase transactions and so on contribute a lot towards grouping or basically understanding the data spread

Decision:

The first seaborn plot contributing to our target ‘Tenure’ is a scatter plot of ‘Credit Limit’ vs ‘Purchase Transactions’, these features are also the most correlated, we can observe that to achieve a tenure of 12 the data clusters between ‘$0 -$1500’ of credit limit the most but only if the purchase transactions are less than a 100.

On point: That just says that if we have a customer looking for tenure payment plan then he must show a valid bank account with credit limit of anywhere from $0 to $1500 and his purchase transactions should be less than 100.

**Configurations:** Install Anaconda and the libraries needed, analyze your data thoroughly

**Limitations:** No limitations were found

**References:**

1. <https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html>

2. <https://www.investopedia.com/mortgage/reverse-mortgage/>

**Question 7 : Write a program in which take an Input file, use the simple approach below to summarize a text file:Link to input file:** [**https://umkc.box.com/s/7by0f4540cdbdp3pm60h5fxxffefsvrw**](https://umkc.box.com/s/7by0f4540cdbdp3pm60h5fxxffefsvrw)

**a. Read the data from a file**

**b. Tokenize the text into words and apply lemmatization technique on each word.**

**c. Find all the trigrams for the words.**

**d. Extract the top 10 of the most repeated trigrams based on their count.**

**e. Go through the text in the file**

**f. Find all the sentences with the most repeated tri-grams**

**g. Extract those sentences and concatenate**

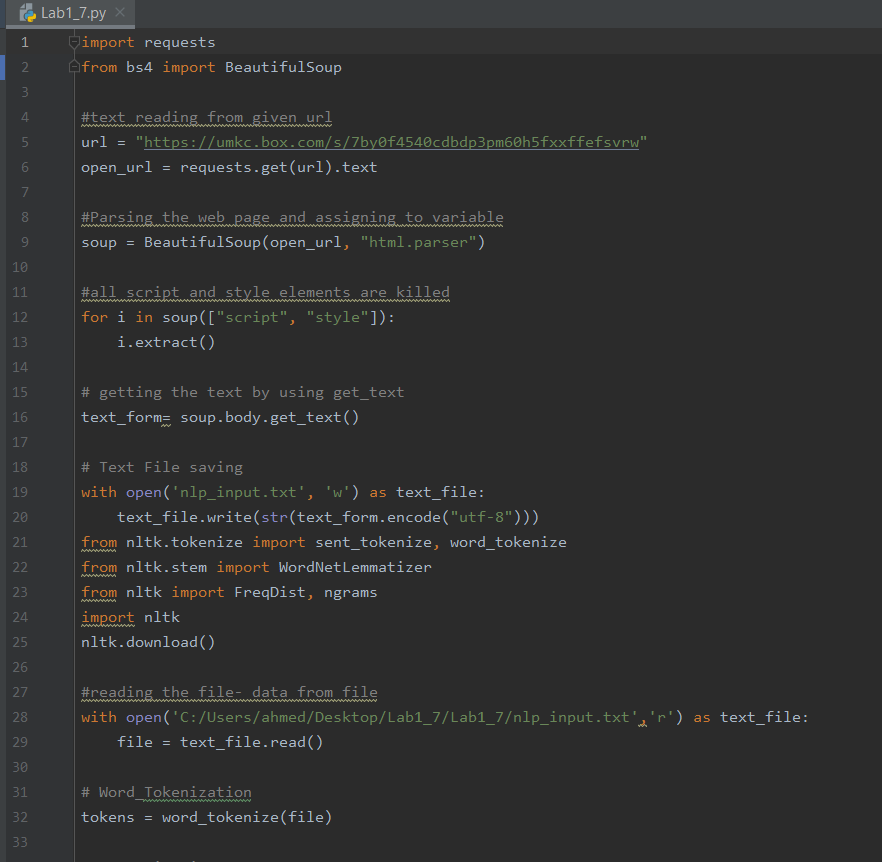
**h. Print the concatenated result**

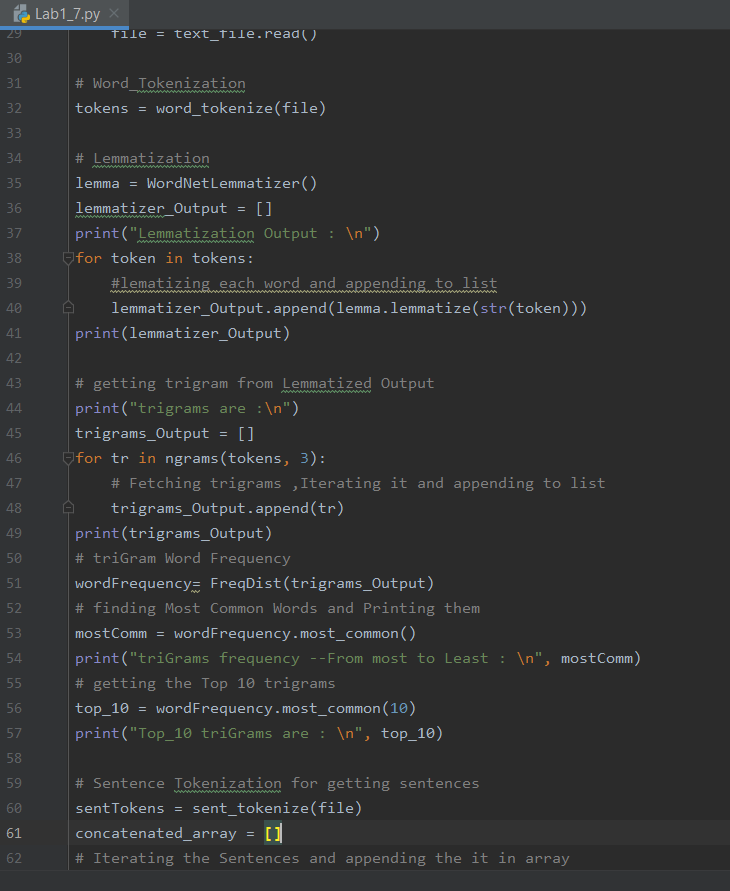
**Solution:**

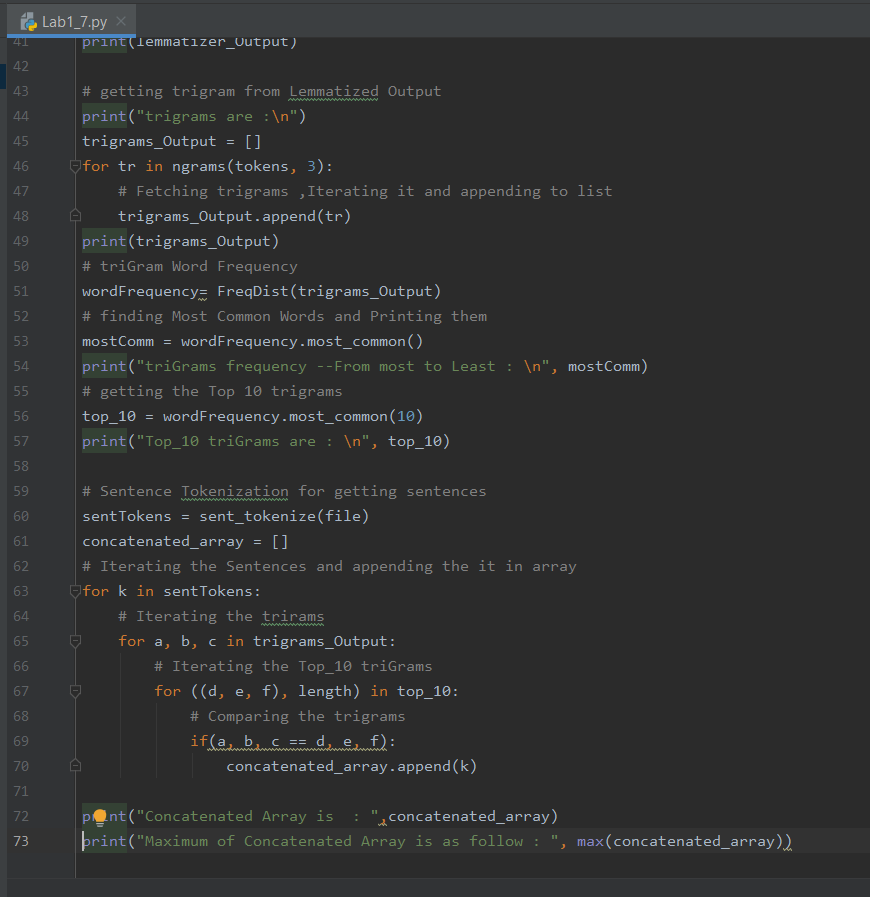
**Objective:**

**Features:**

**Code Steps:**







**Output:**



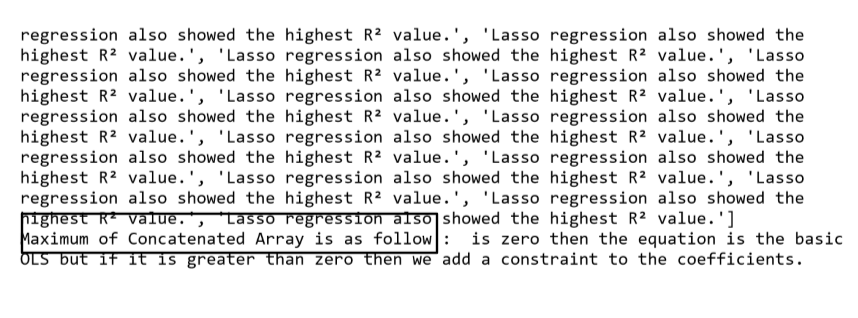












**Configurations:**

**Limitations:** No limitations

**References:**

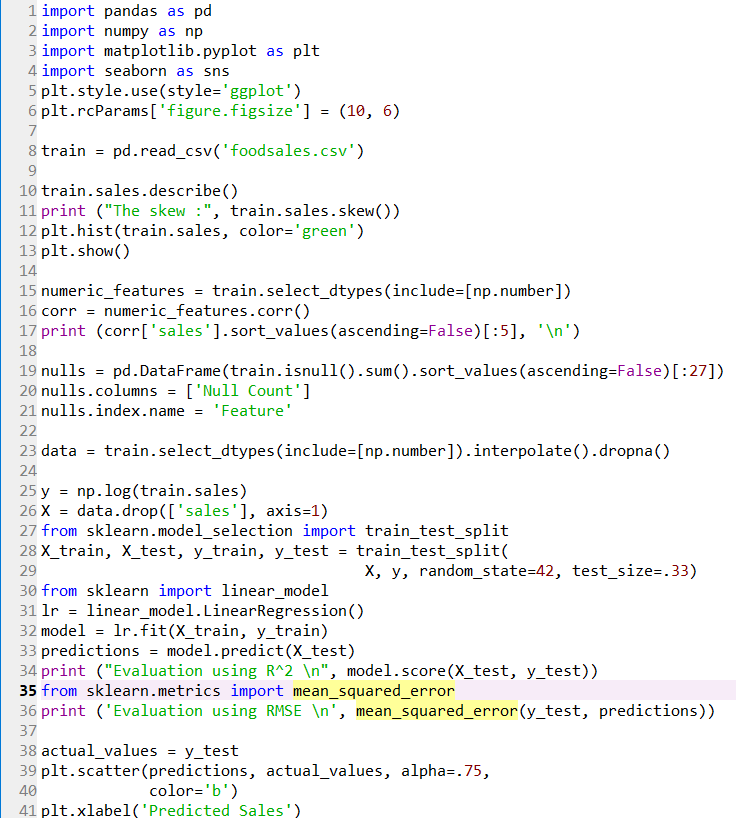
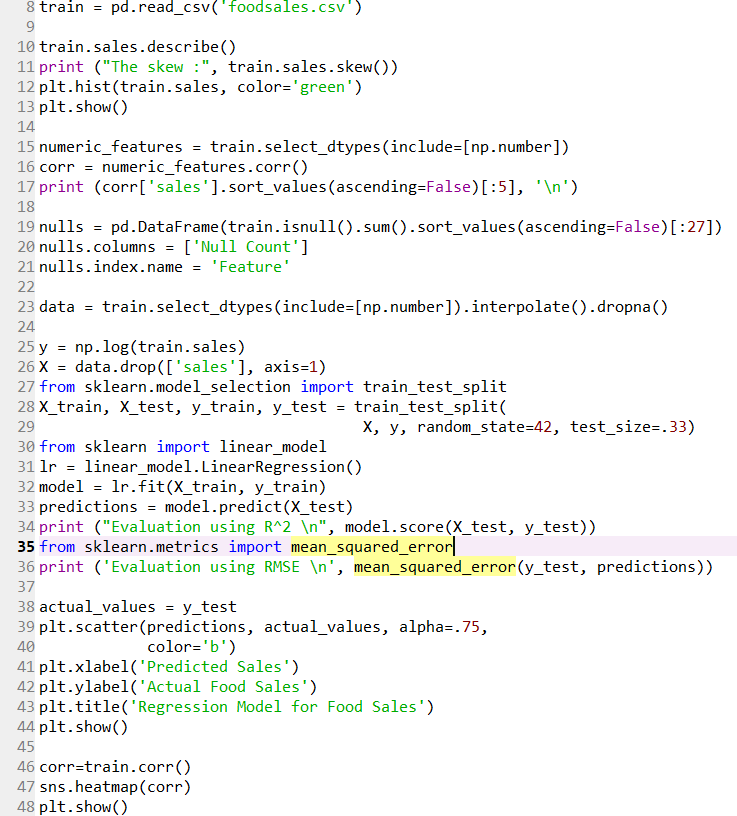
1. Class slides

2. In class programming code

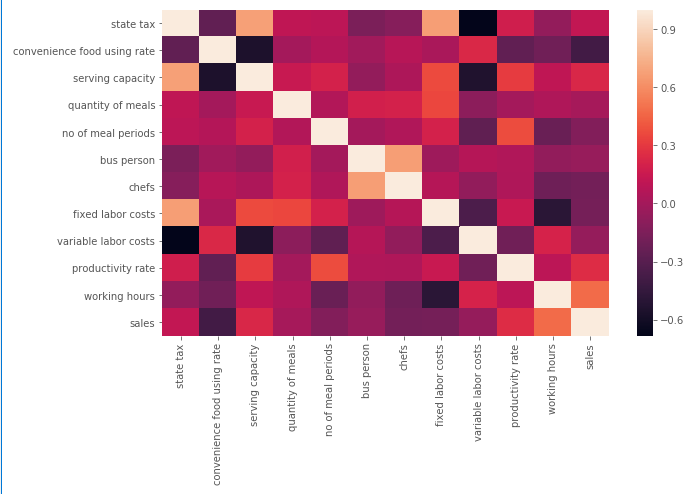
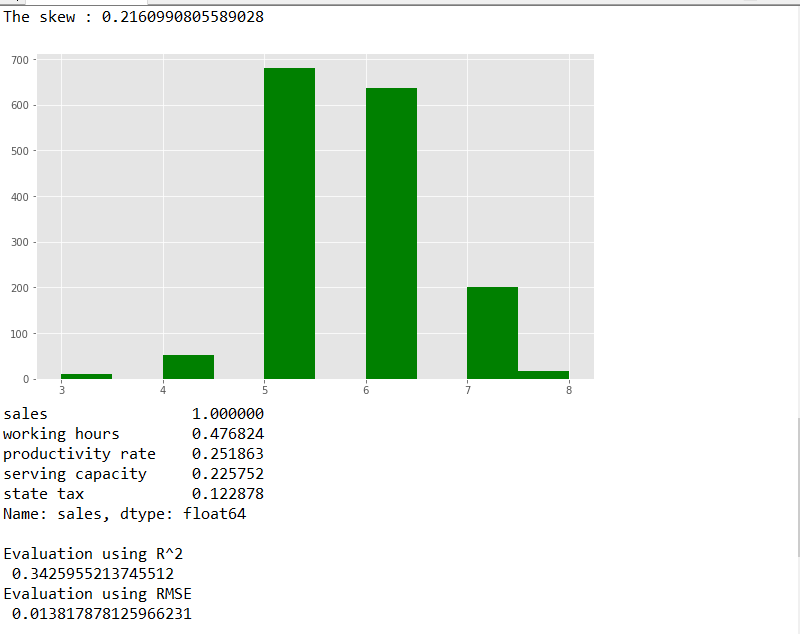
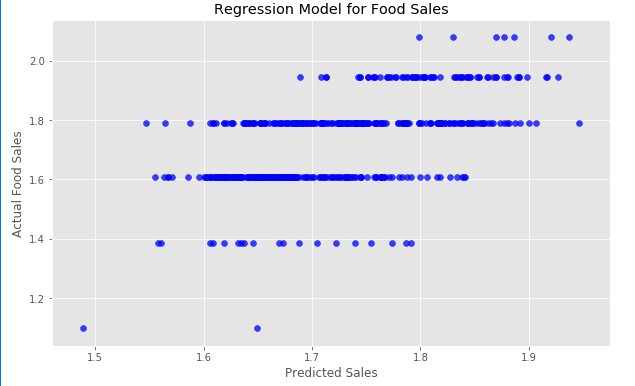
**Question 8: Create Multiple Regression by choosing a dataset of your choice (again before evaluating, clean the data set with the EDA learned in the class). Evaluate the model using RMSE and R2 and also report if you saw any improvement before and after the EDA.**

Solution:

**Code Steps:**

**Output:**

****

**Configurations:**

**Limitations:** No limitations

**References:**

1. Class slides

2. In class programming code