LAB ASSIGNMENT 2

NAME: HARISH CHANDRA JYOSHI

CLASS ID:12

UMKC MAIL ID:hjddh@mail.umkc.edu

NAME:VENKATA AKHILA KRISHNA ATLURI

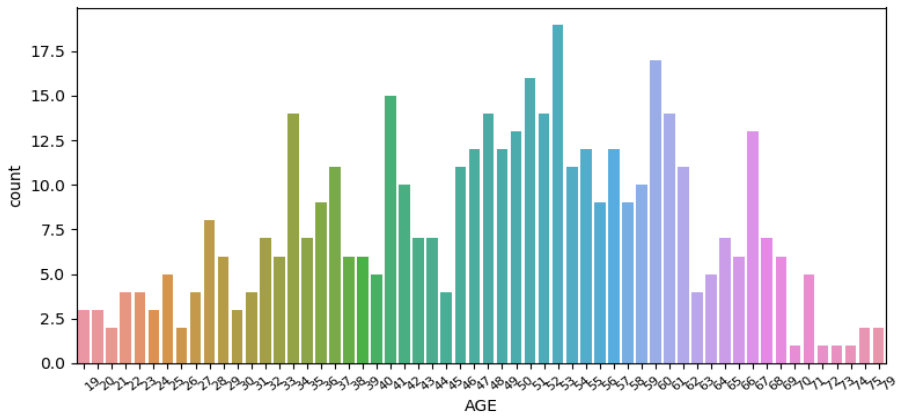
CLASS ID:04

UMKC MAIL ID:vagq2@mail.umkc.edu

TASK1:

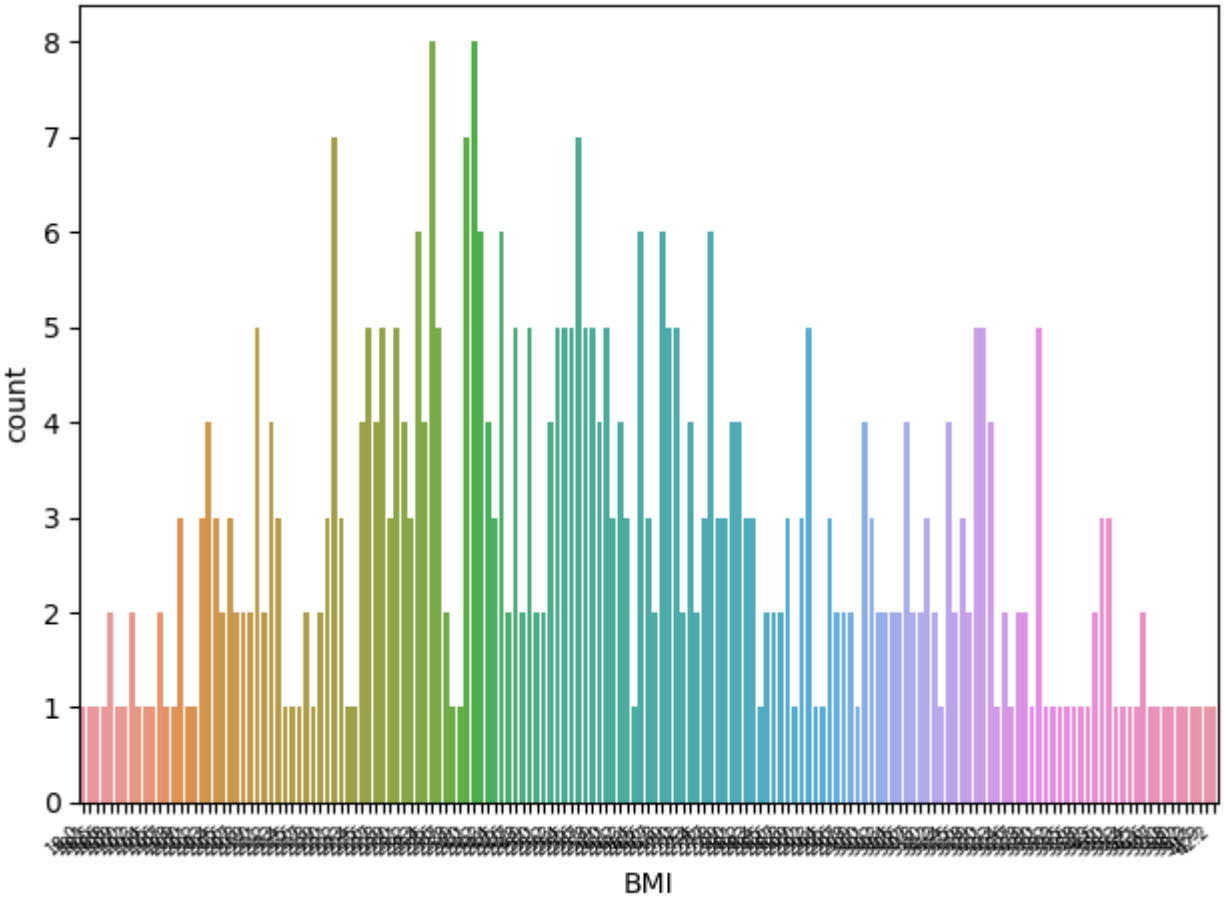
a) Choosing any dataset and plotting each category in dataset.

we have choose diabetes as our dataset and plotted each of the category in the dataset. we made use of seaborn library and matplotlib to plot the each category in the dataset.

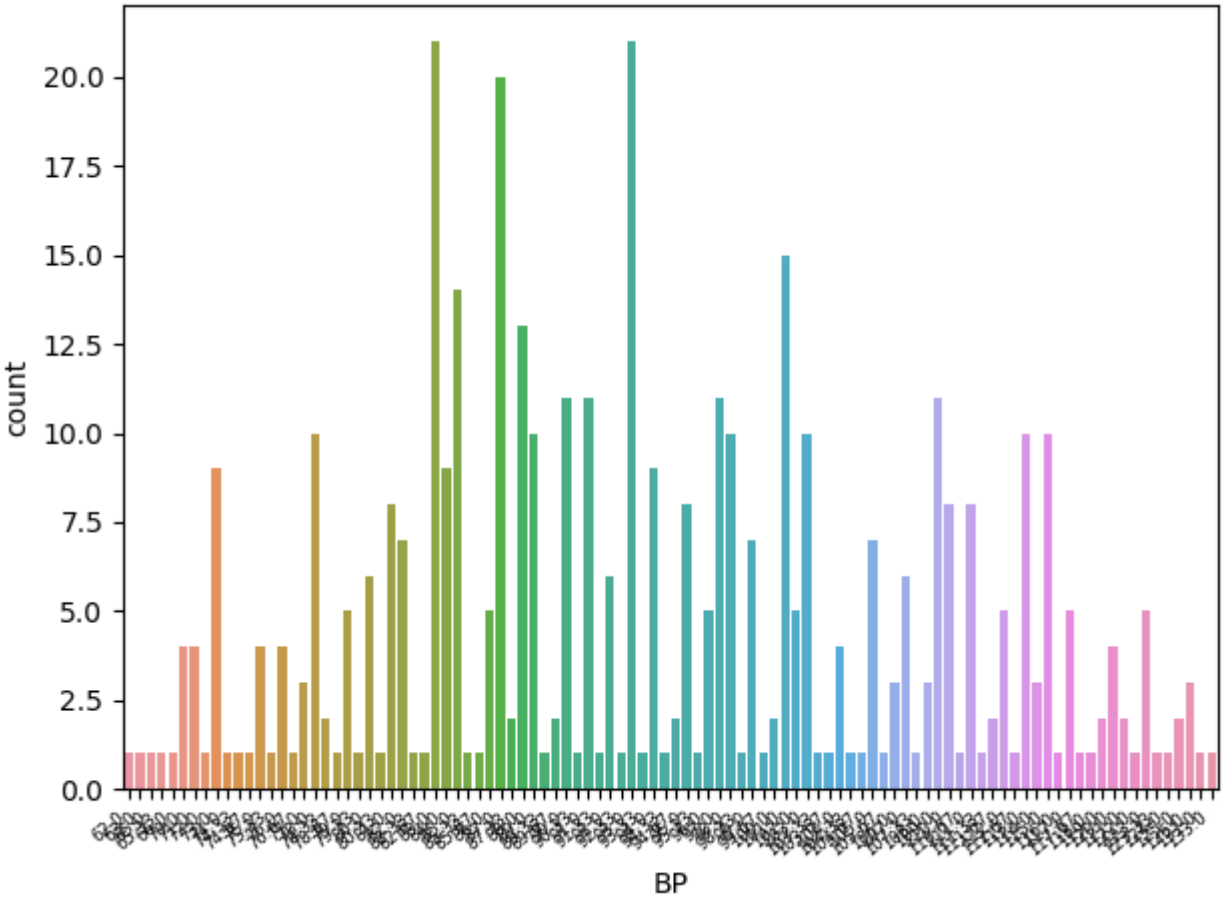


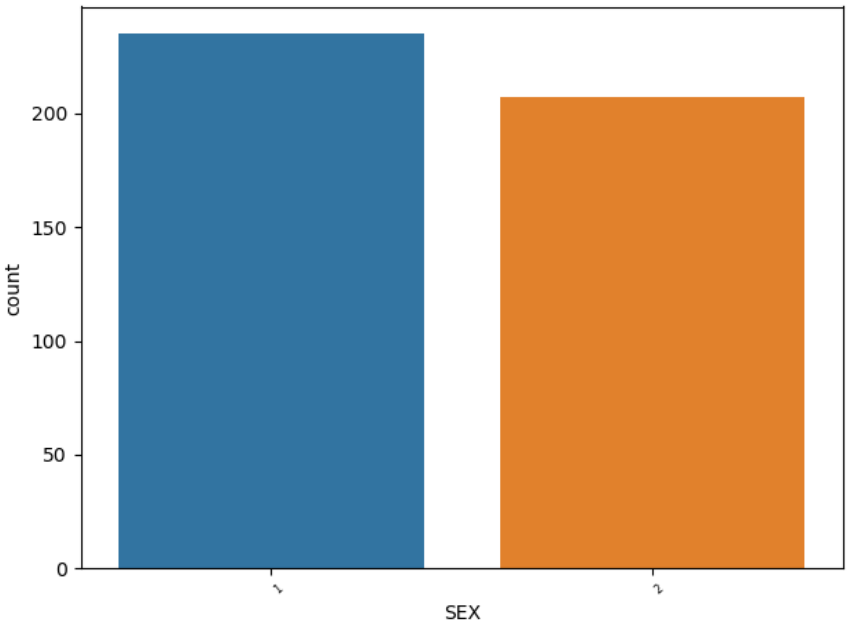
AGE

BMI

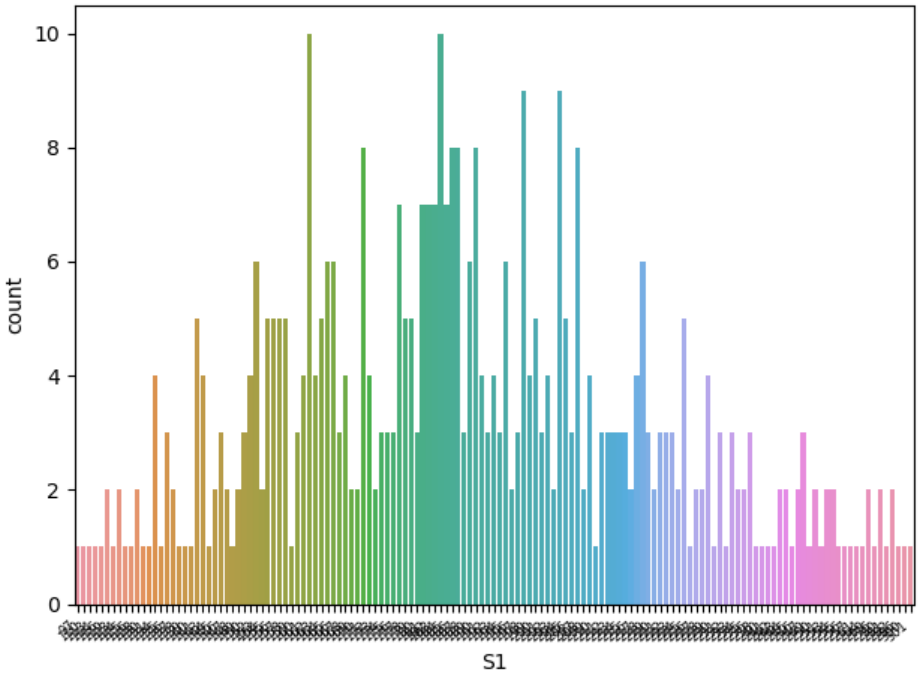


BP


20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
count
BP

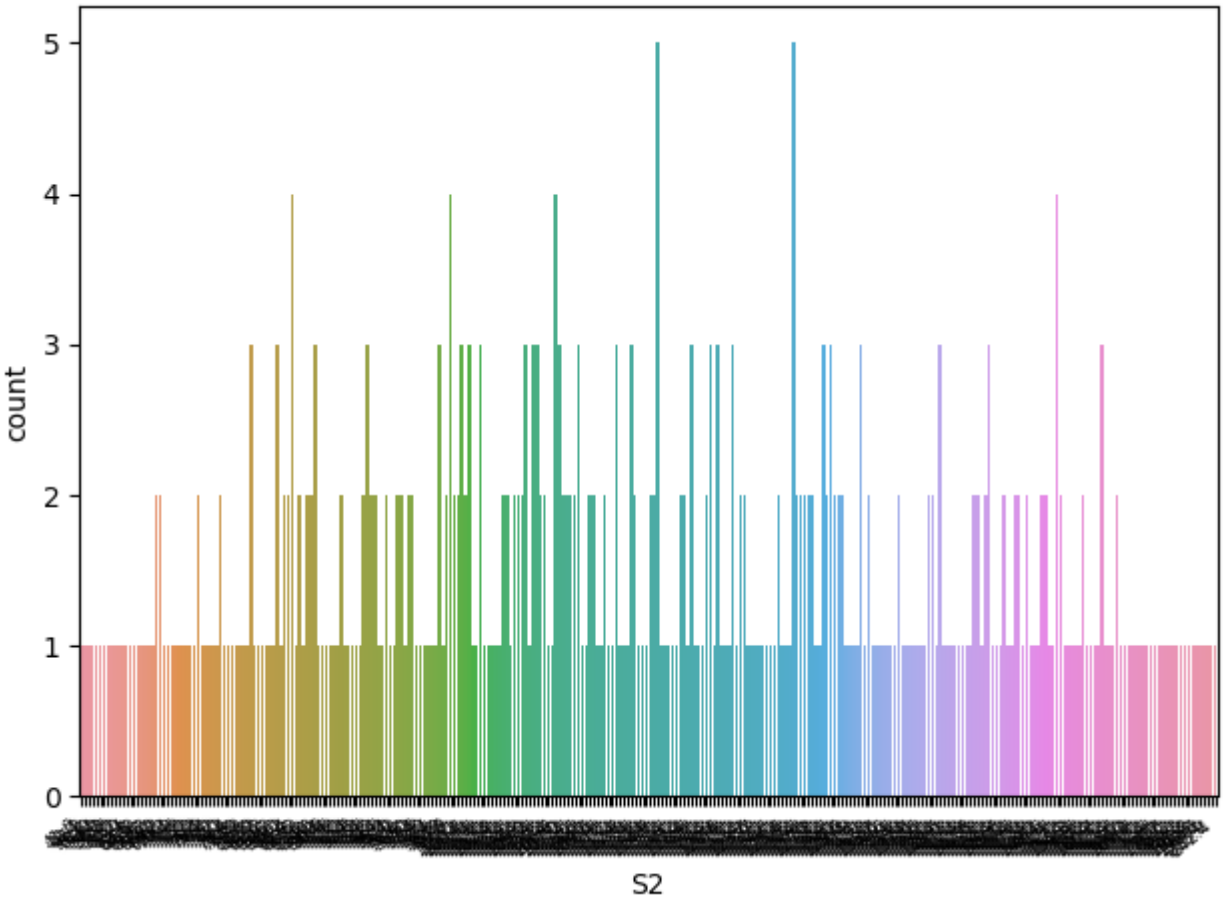

200
150
100
50
0
count
1
2
SEX

SEX

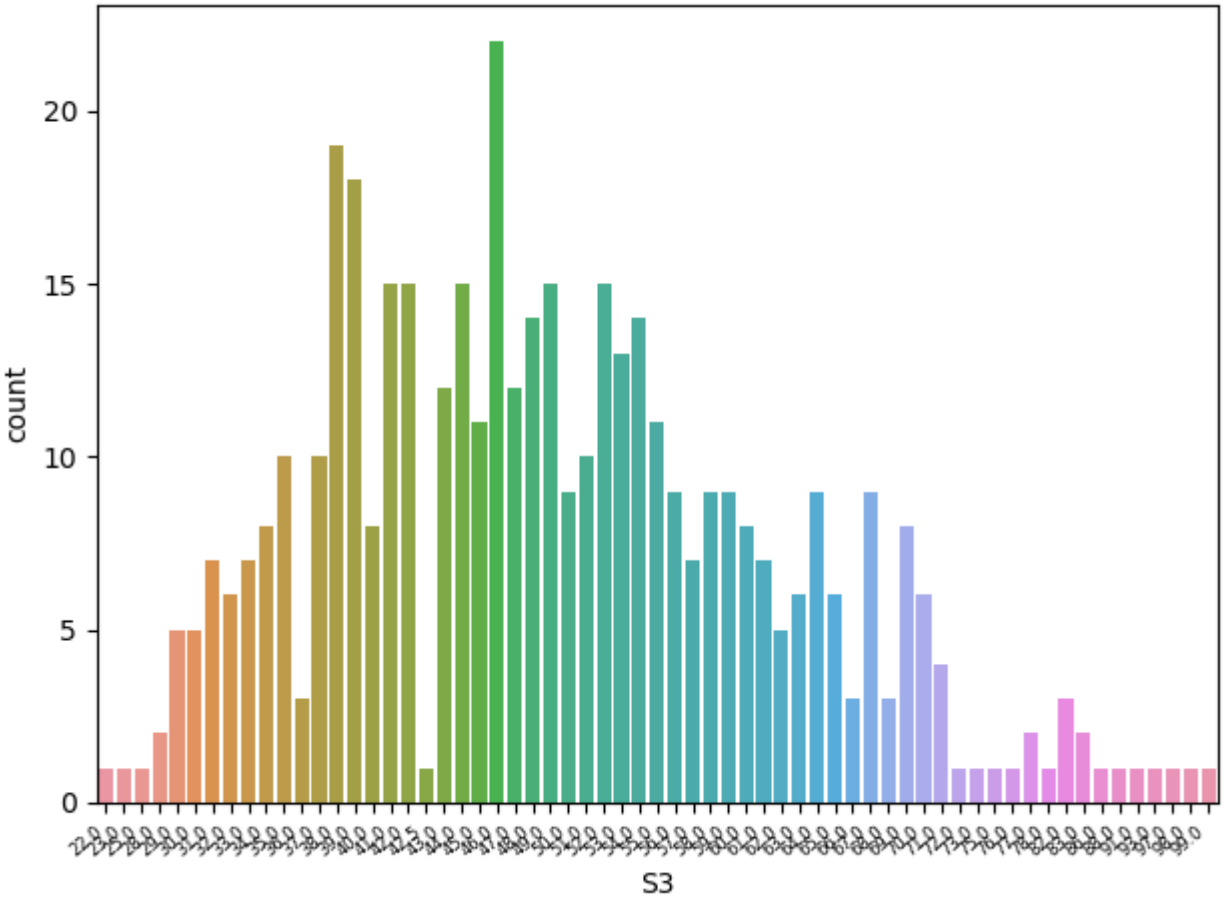


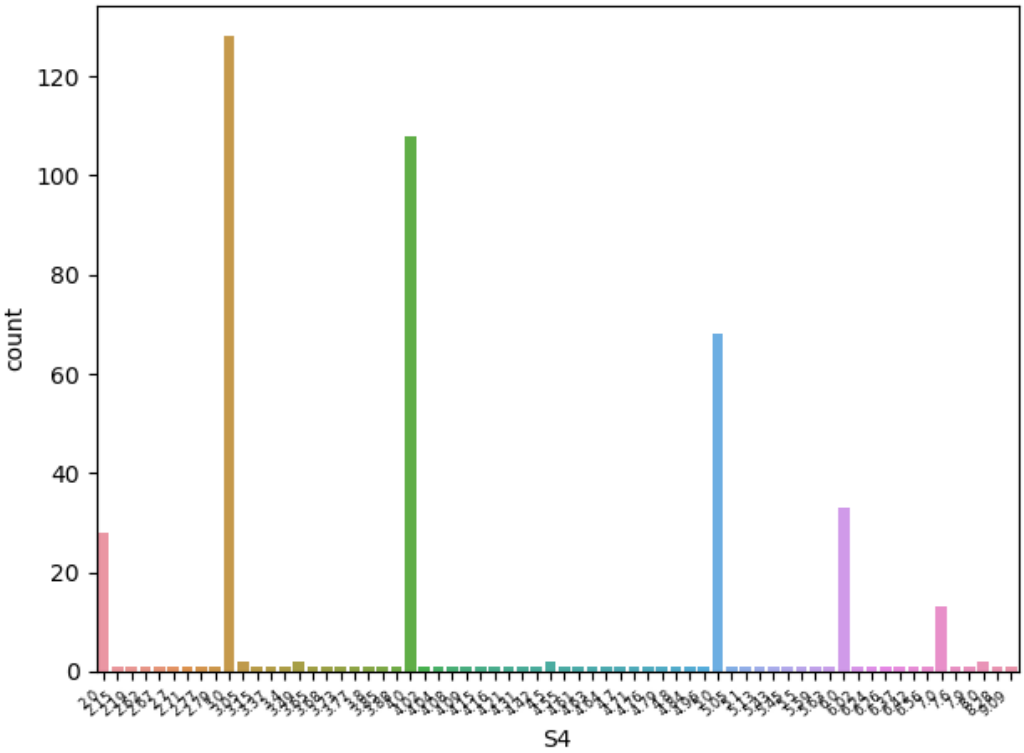
S1

s2



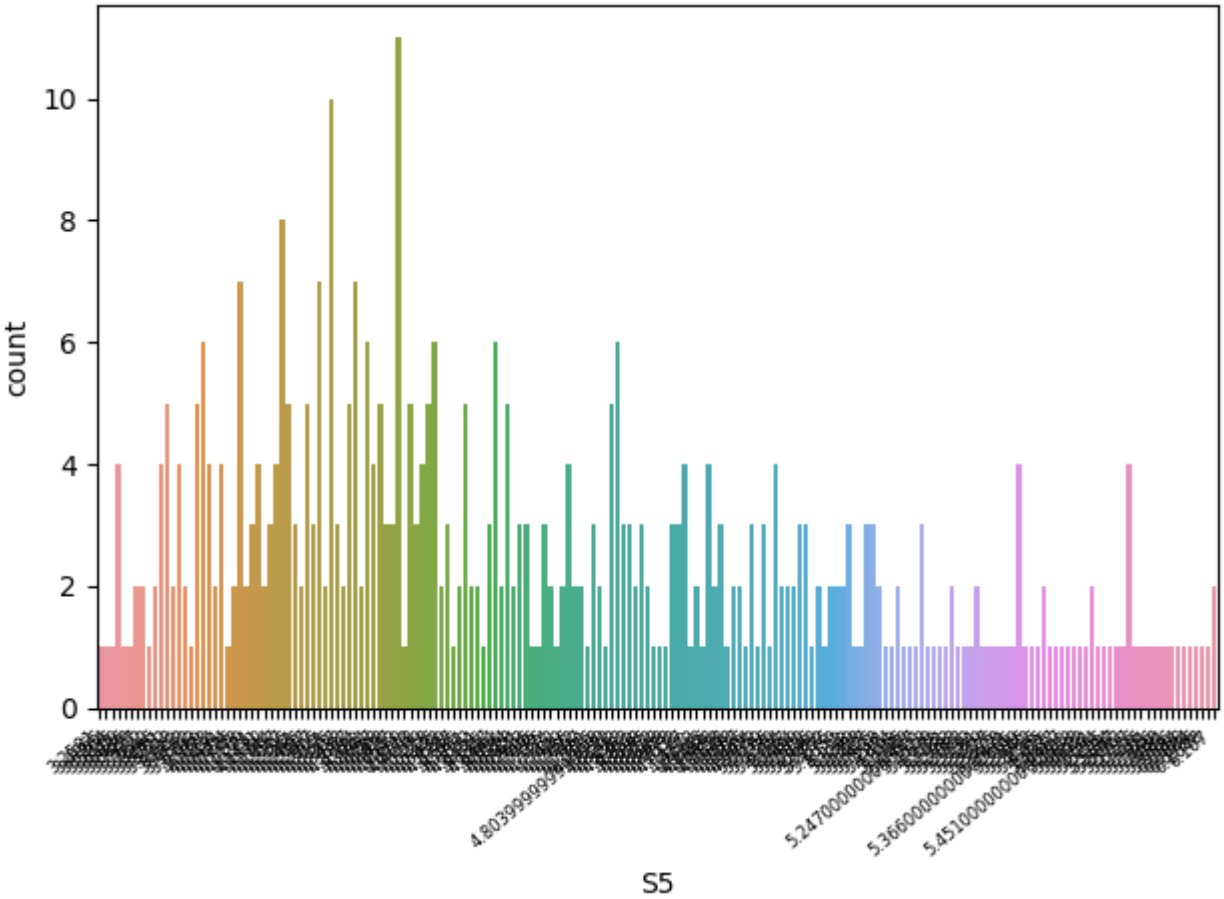
s3

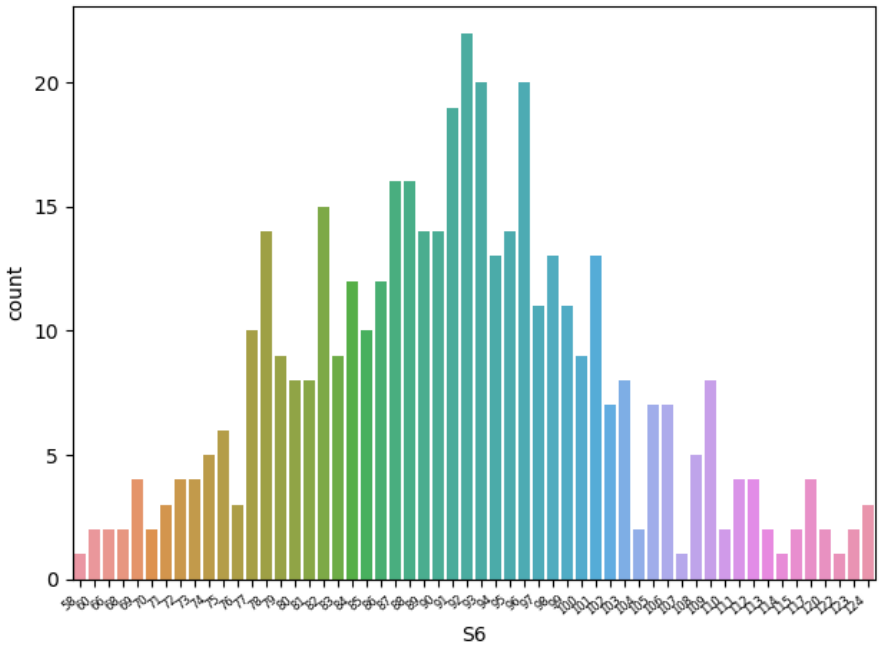




s4

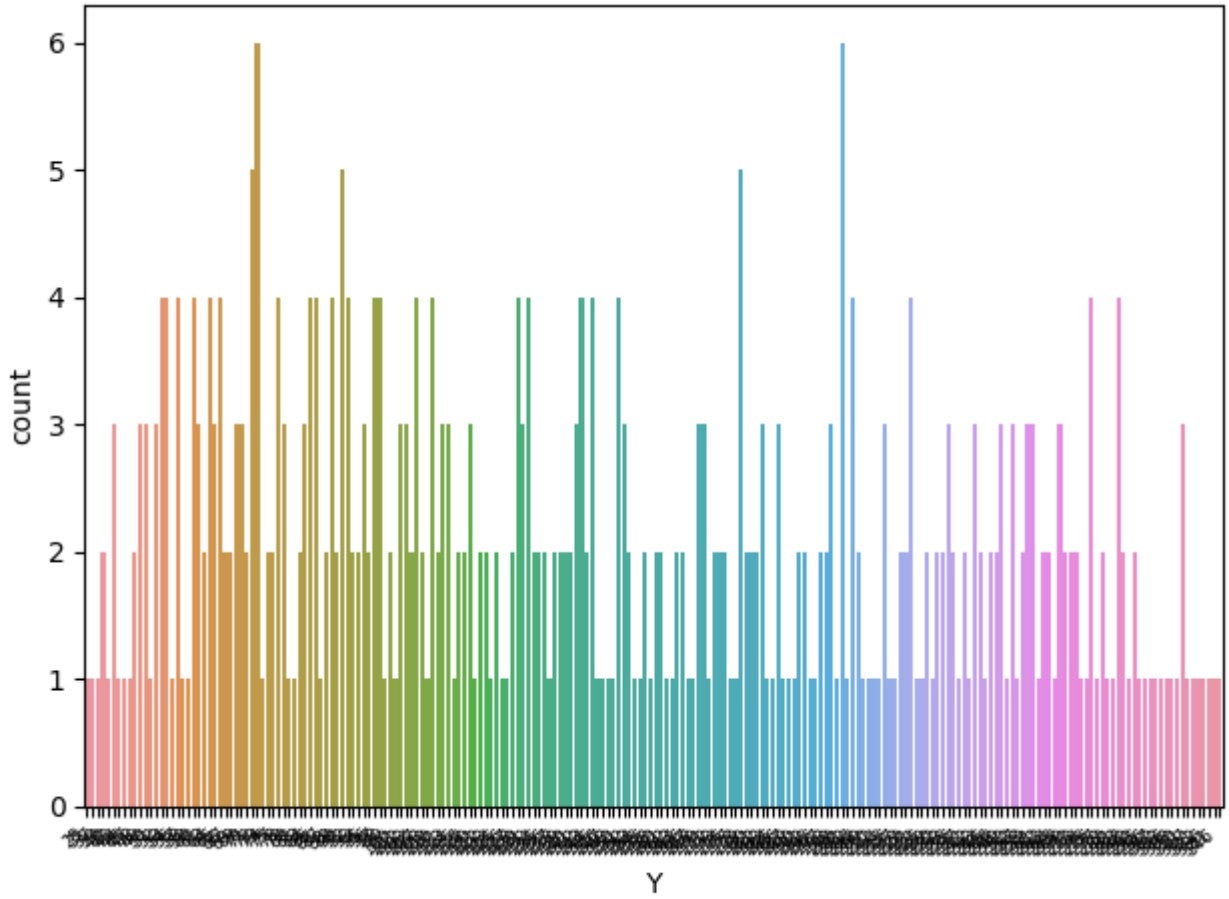
s5


count
0
2
4
6
8
10
4.803999999
5.247000000
5.366000000
5.451000000
S5


count
0
5
10
15
20
S6

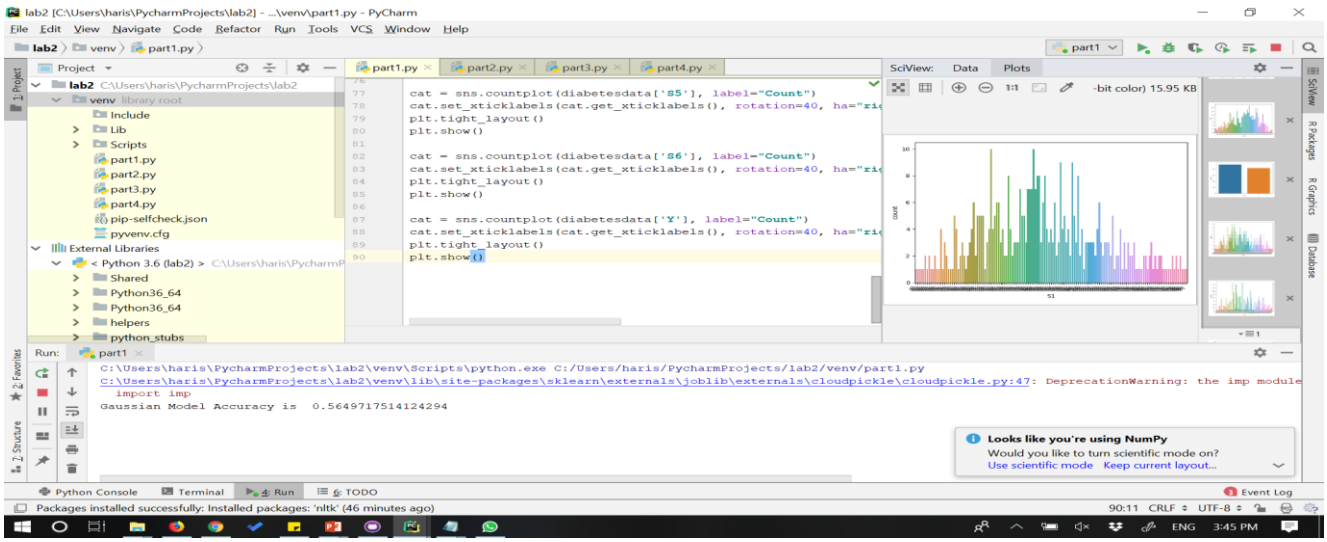
s6

Y



b) creating and evaluating one model based on naive bayes classification

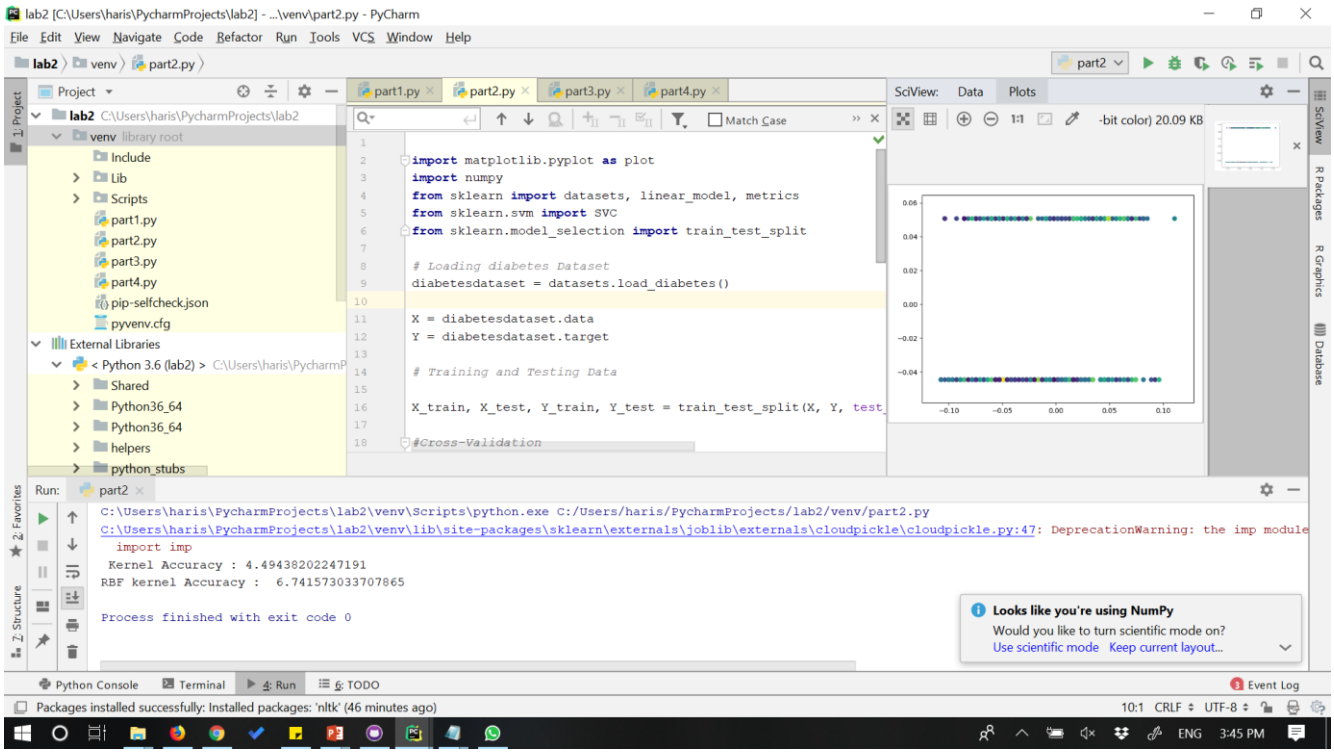
we have choose diabetes as our dataset and performed naive bayes classification and evaluated our model and printing the test accuracy of the model.



TASK2:

Implementing Support Vector Machine classification:

we have choose diabetes as our dataset and implemented "poly" kernel with degree=4 and "rbf" kernel and compared the accuracy between two models.

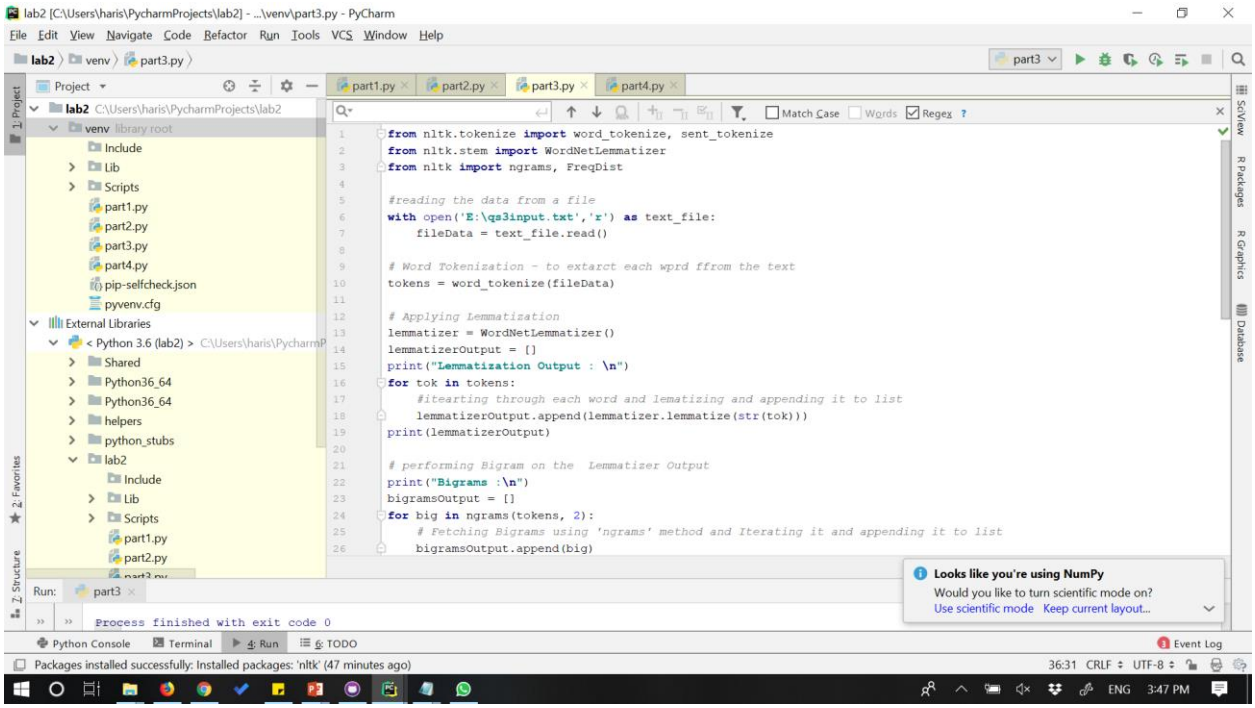


comparing the results we found out that rbf kernel has the greater accuracy in the diabetes dataset. Generally the accuracy depends on the c and gamma parameters. More the random state, greater the accuracy.

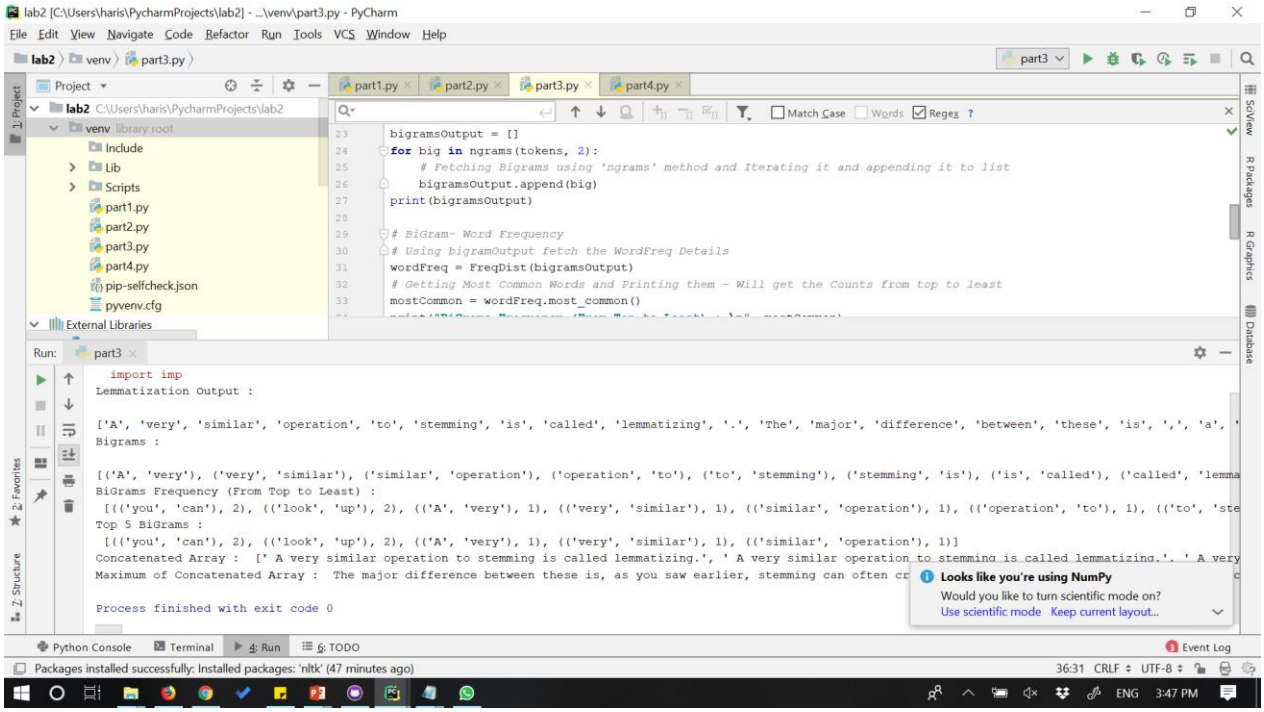
TASK3:

Performing summarization on the text data:

we have taken a text file and performed lemmatization. we have implemented word tokenization initially and implemented lemmatization and bigrams on the tokens and later implemented sentence tokenization, extracted them and concatinated the sentences.



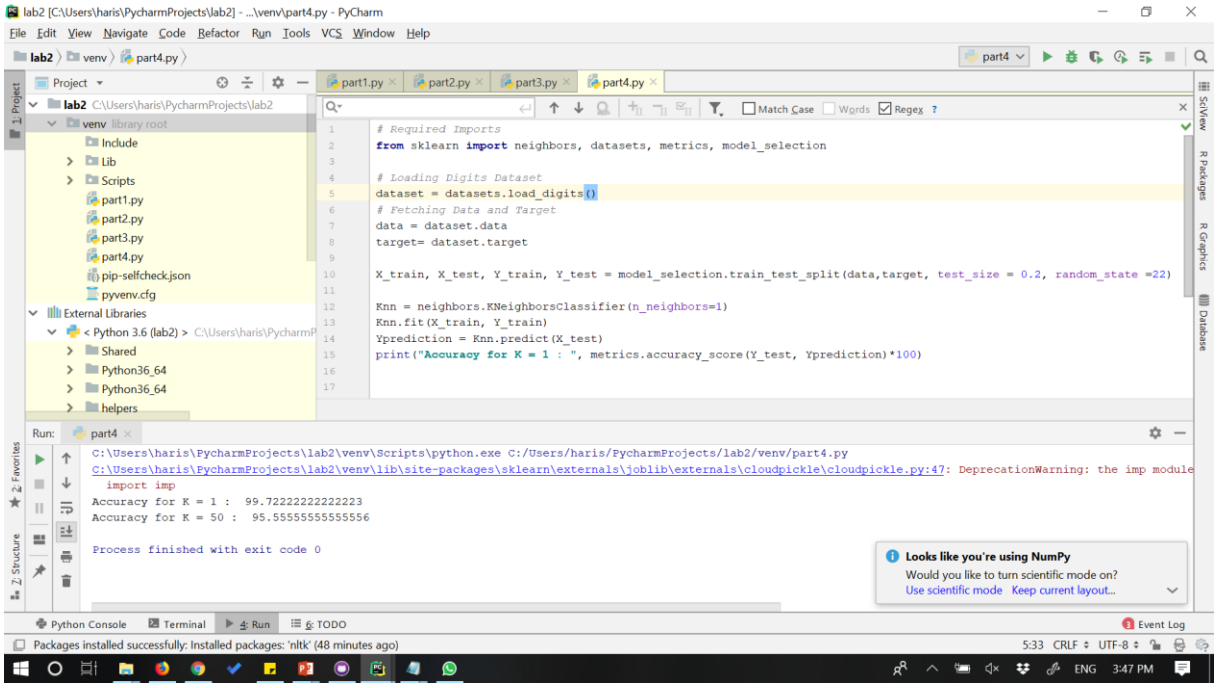
Output:



TASK4:

Reporting your views on the k nearest neighbor algorithm when we change the K

Here we have taken digits dataset and implemented k nearest neighbour algorithm on this. divided the data into training and testing data. Calculated the accuarcy scores by changing the K values.



When we provide a smaller value for k, suppose k=1, it provides more natural fit and leads to less variance, and when we provide a greater value for k, like k=50, it overfits the model and leads to high variance.

we got a accuracy of 99.7% when k=1 and accuracy of 95.5 when k=50.

YOUTUBE LINK: