

## LAB ASSIGNMENT 2

NAME: HARISH CHANDRA JYOSHI

CLASS ID:12

UMKC MAIL ID:hjddh@mail.umkc.edu

NAME:VENKATA AKHILA KRISHNA ATLURI

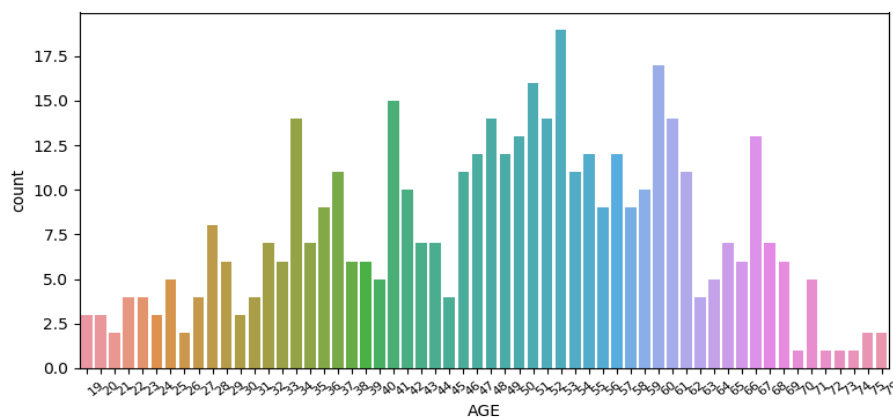
CLASS ID:04

UMKC MAIL ID:vagq2@mail.umkc.edu

### TASK1:

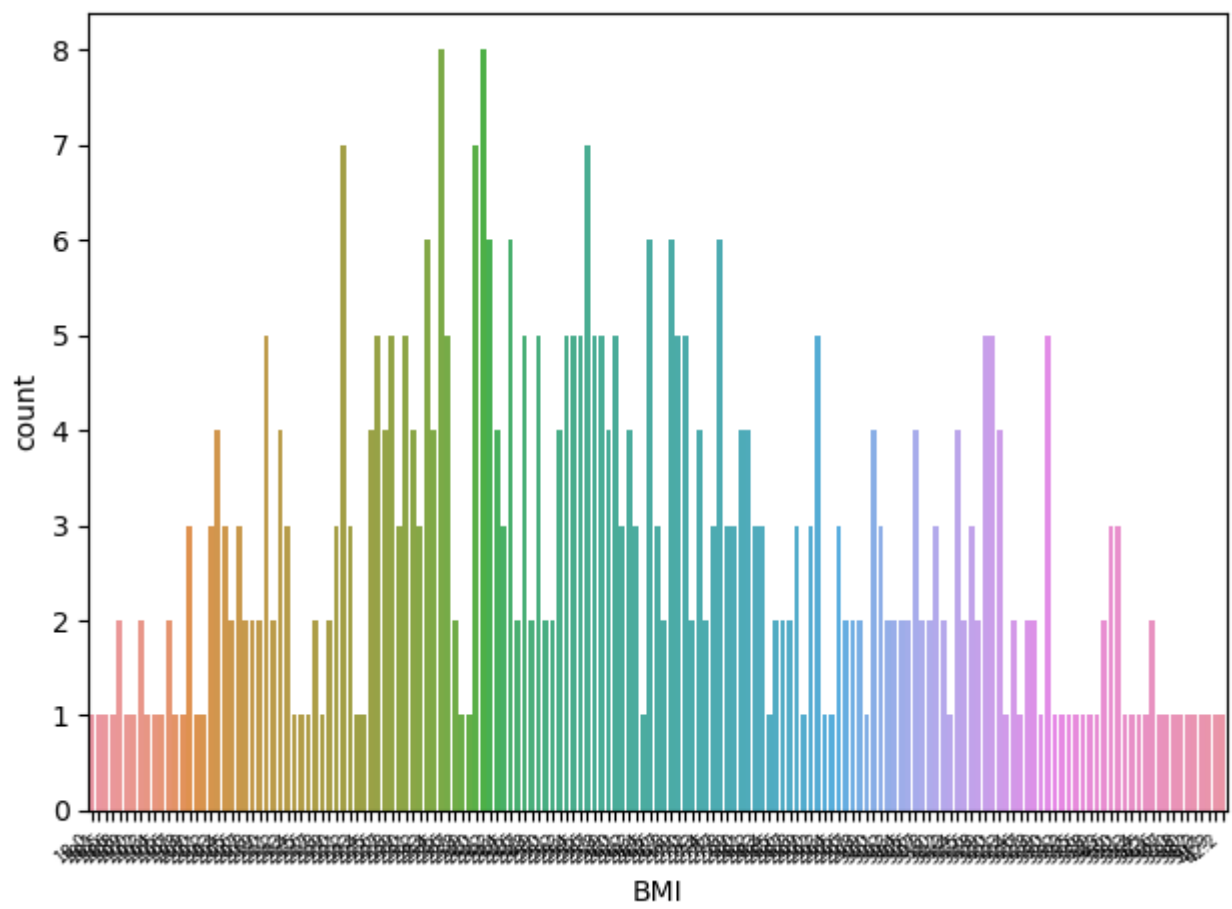
#### a) Choosing any dataset and plotting each category in dataset.

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.

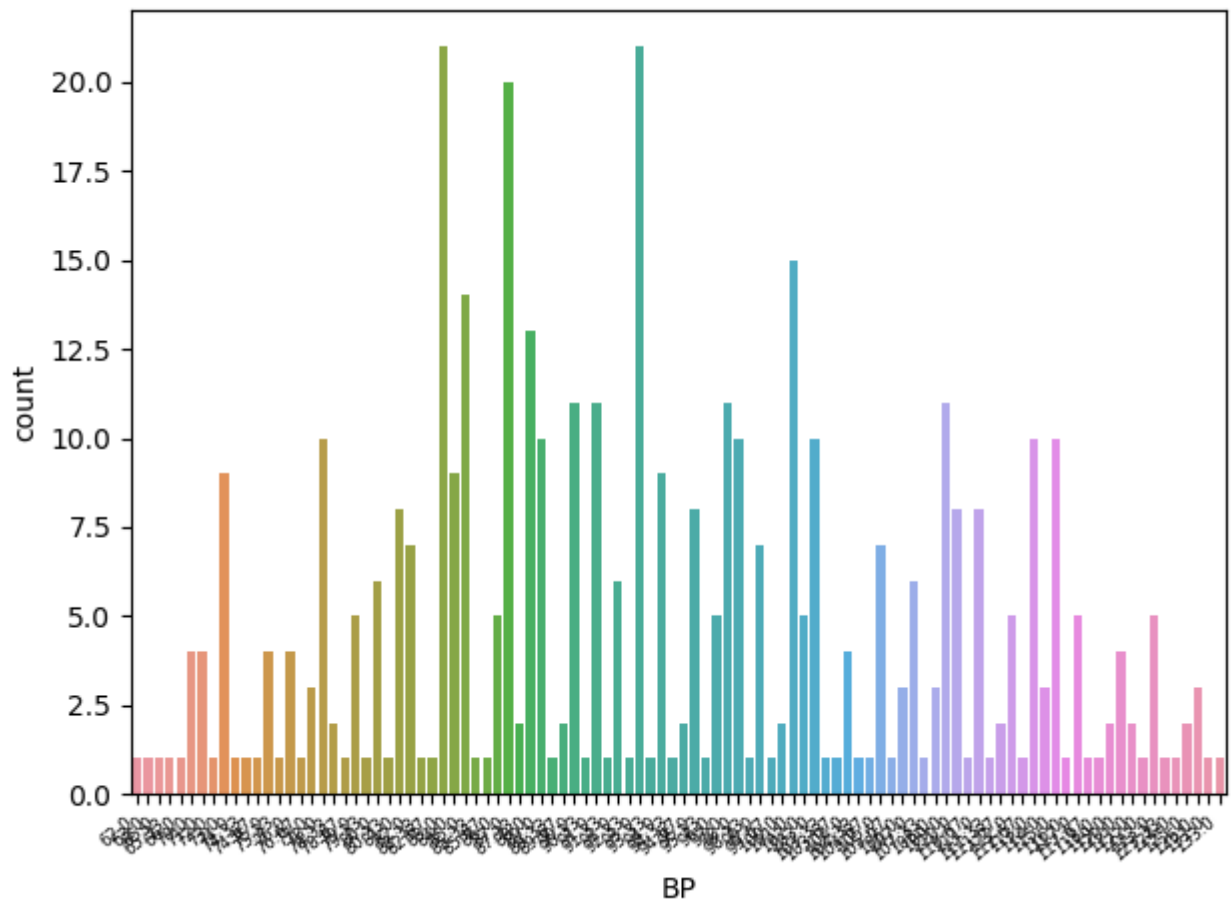


AGE

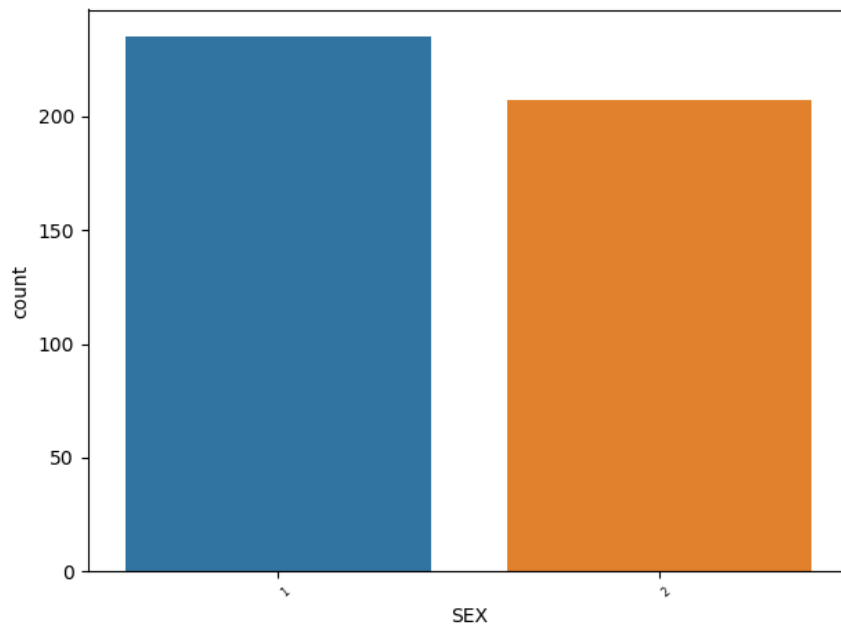
## BMI

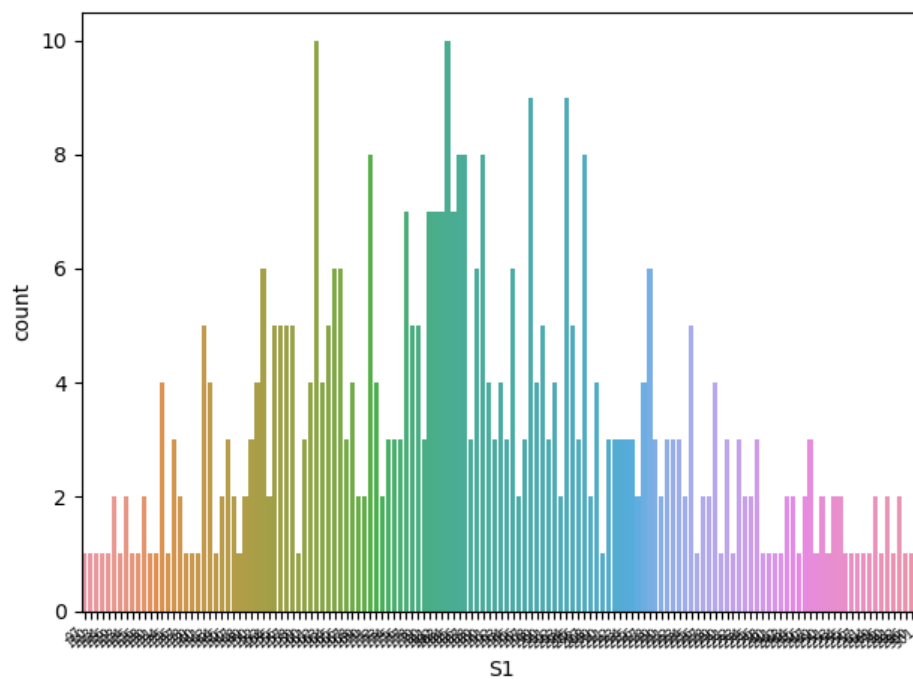


**BP**



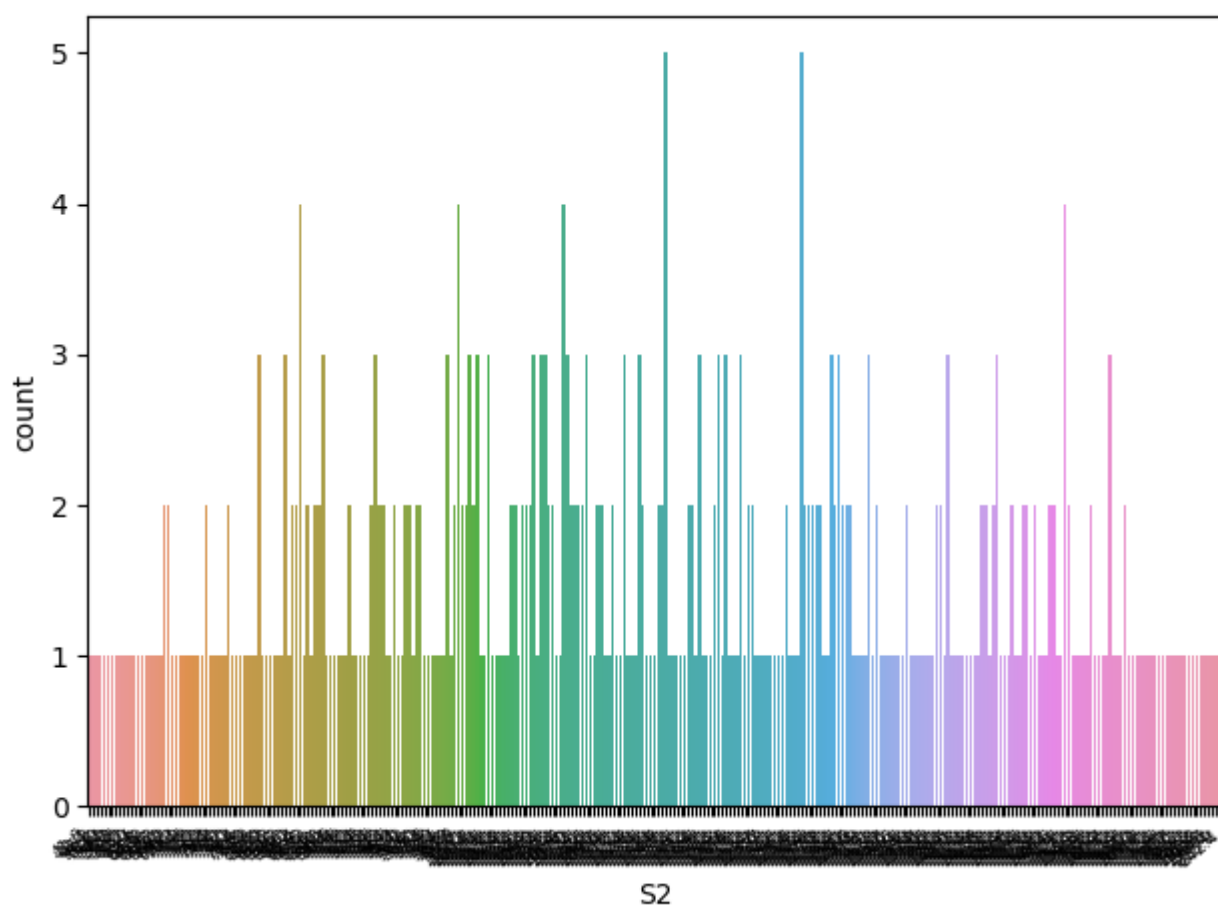
**SEX**



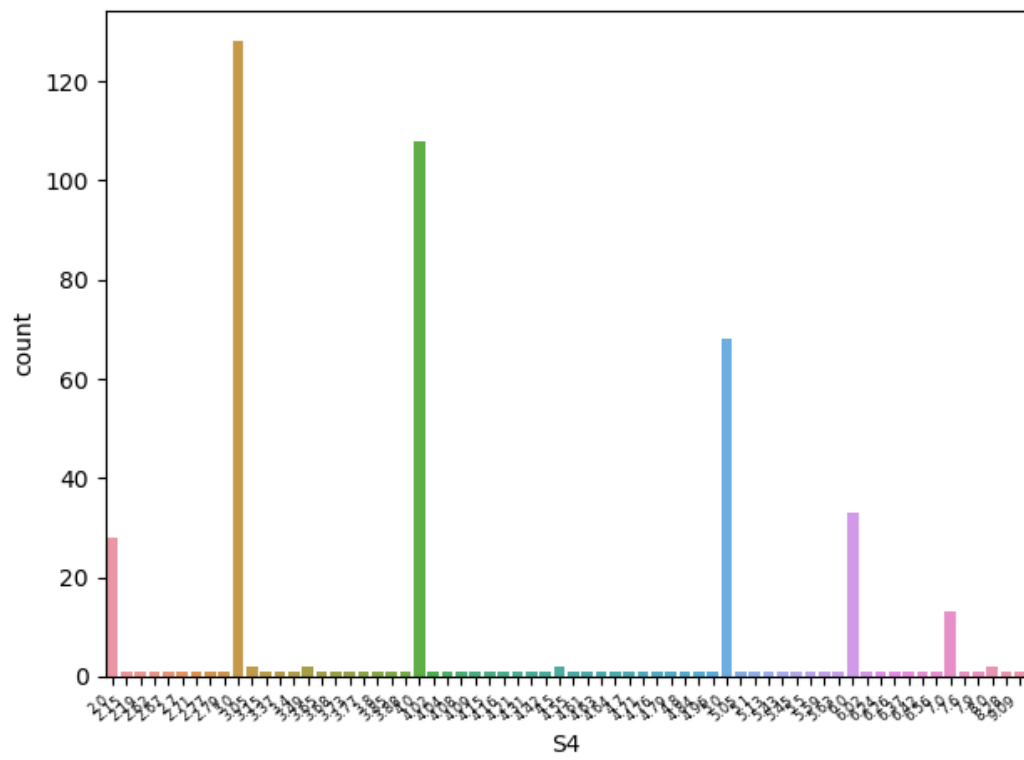
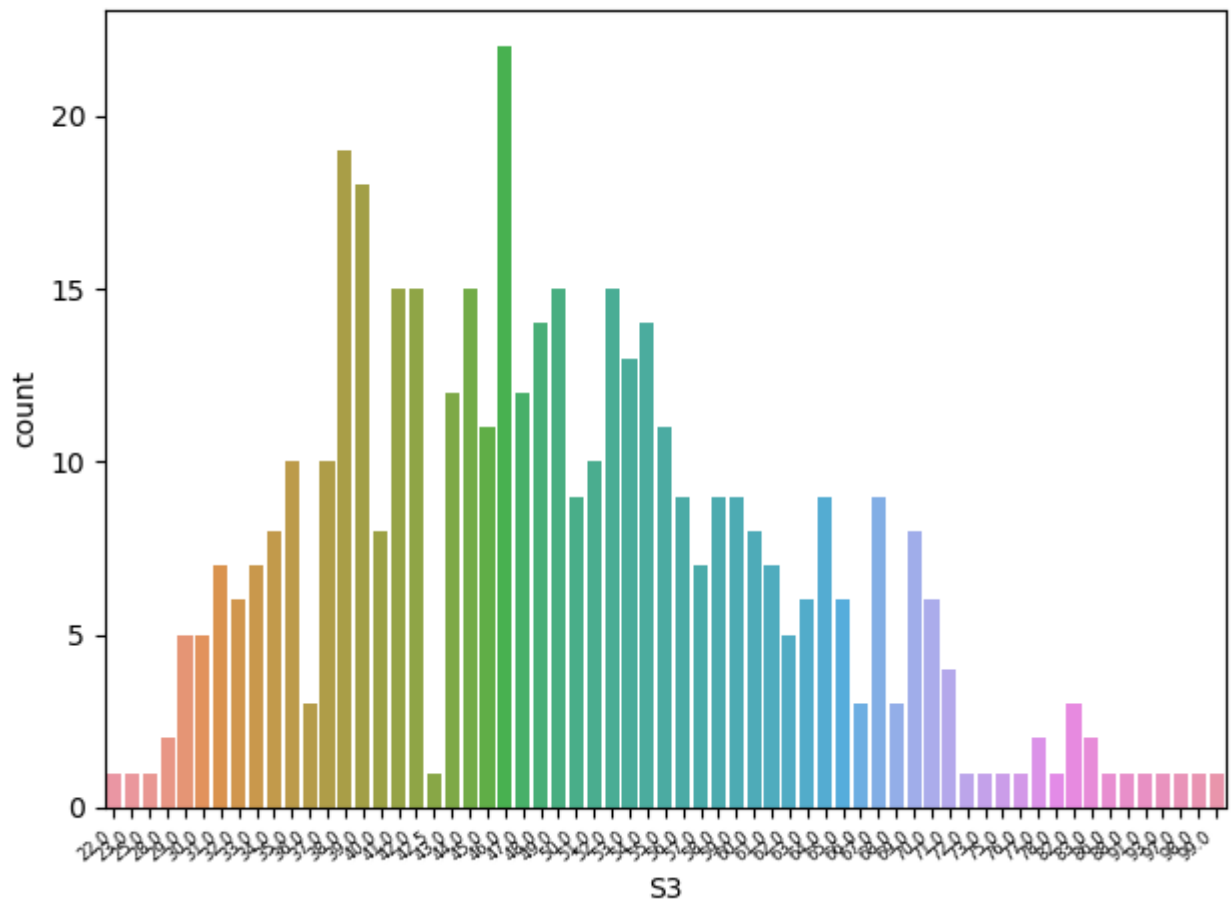


**S1**

**s2**

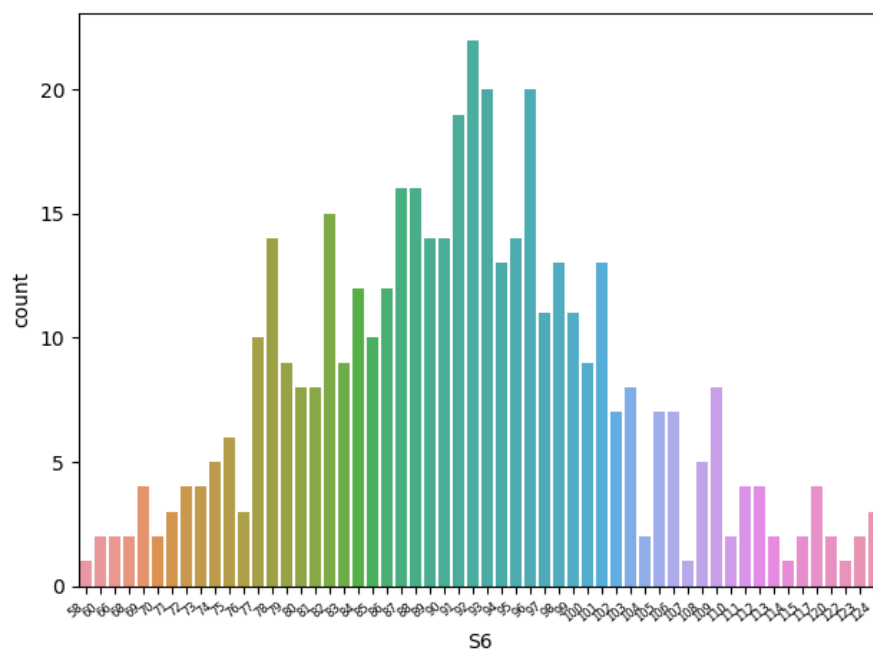
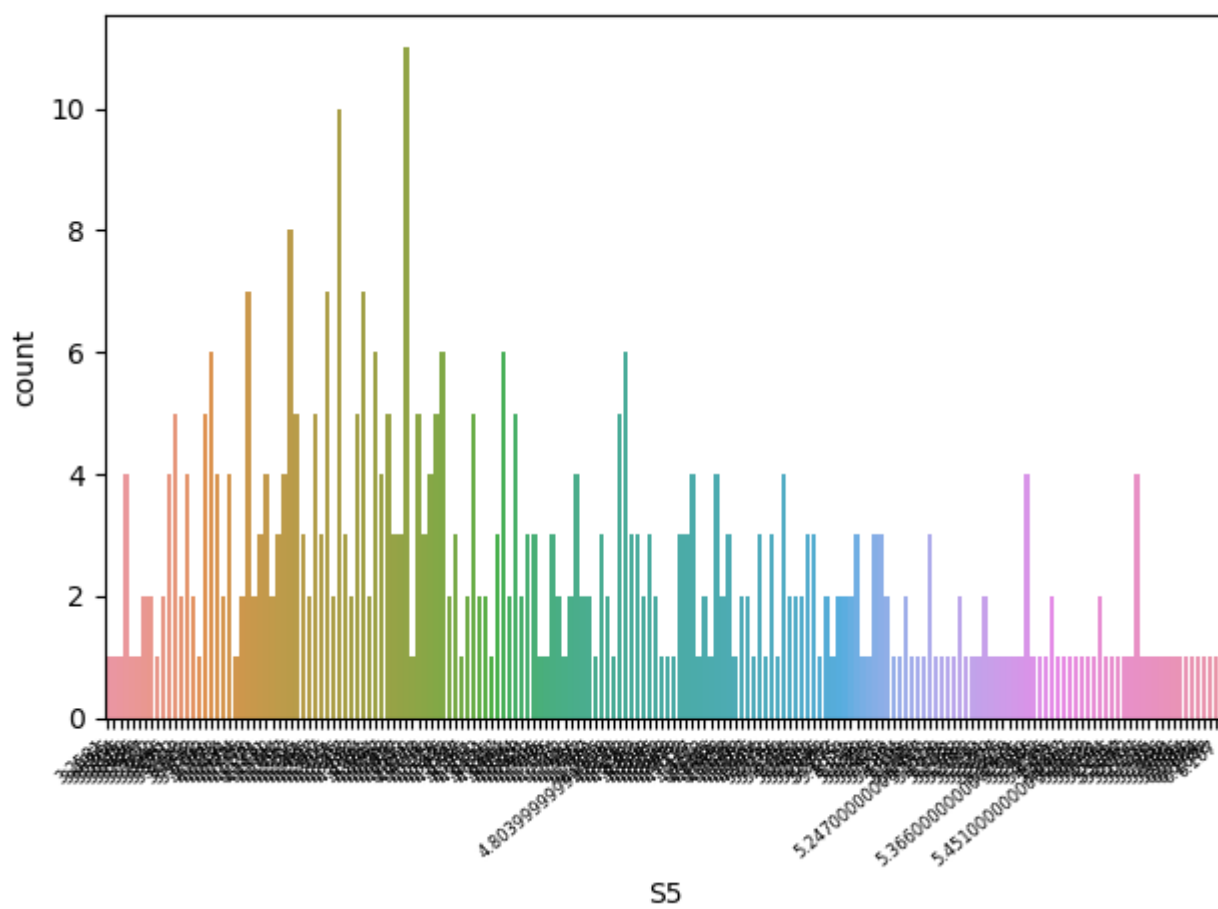


s3



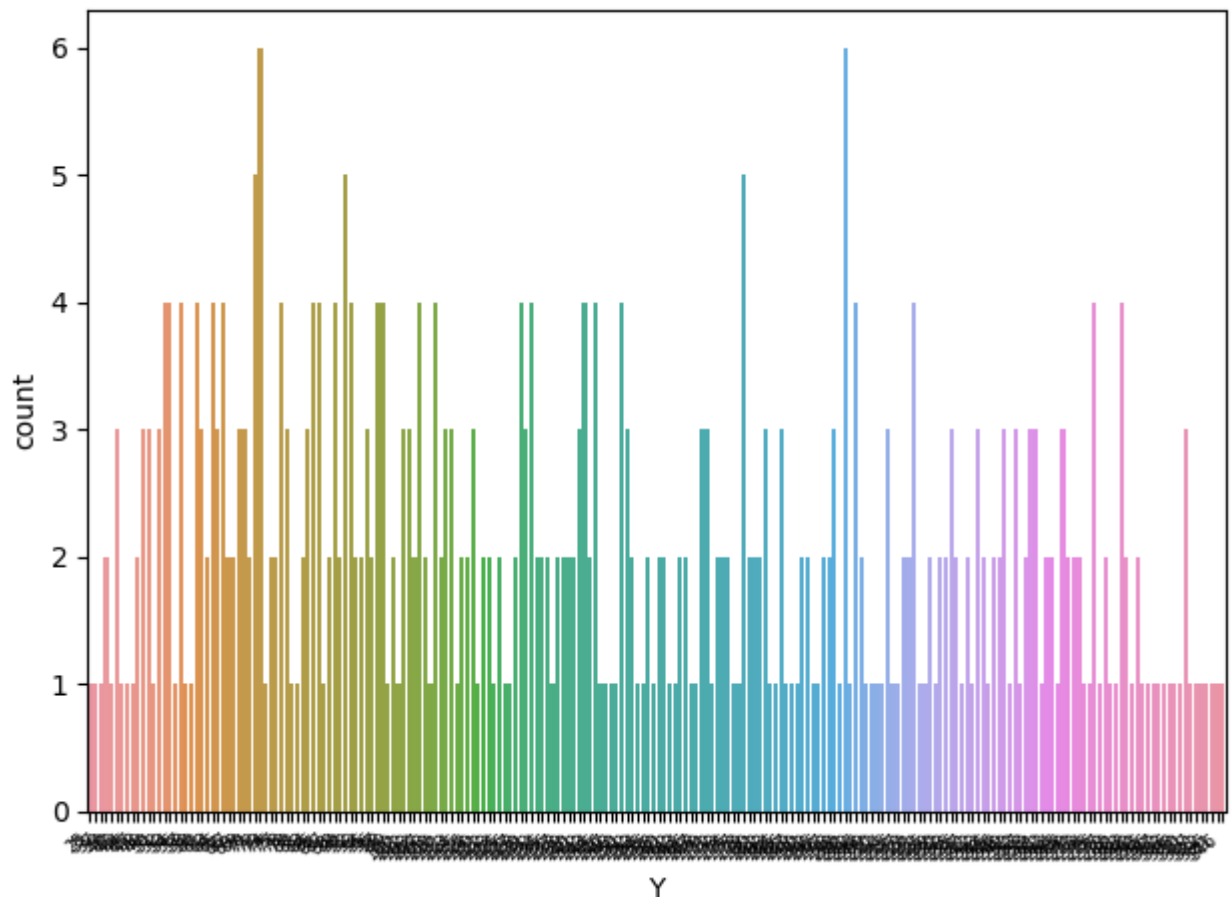
s4

s5



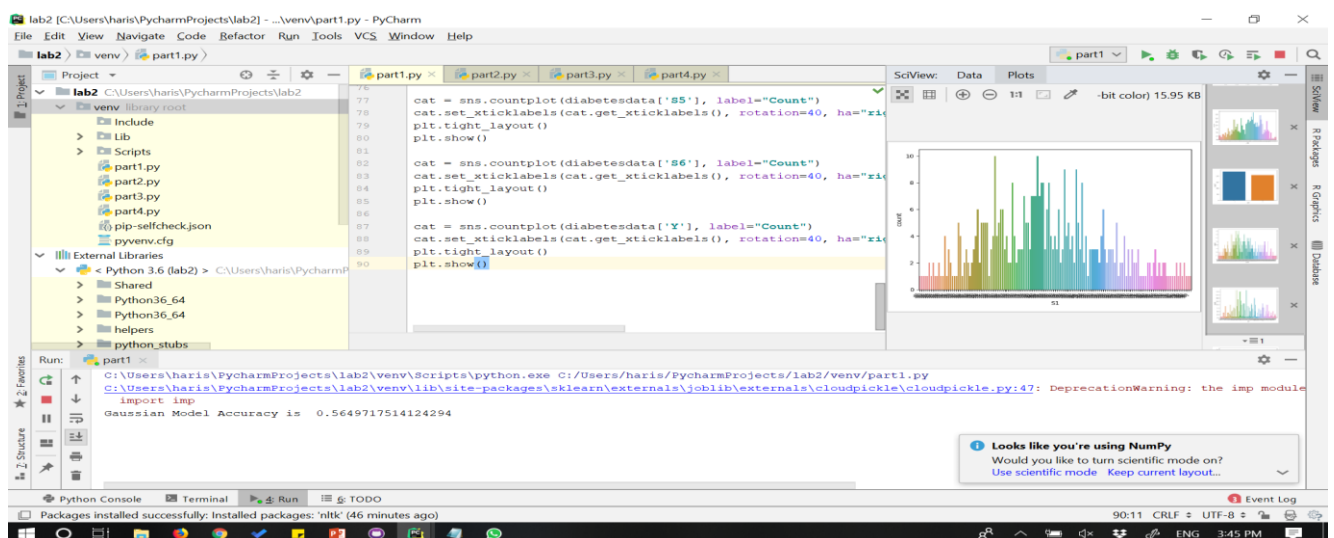
s6

Y



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

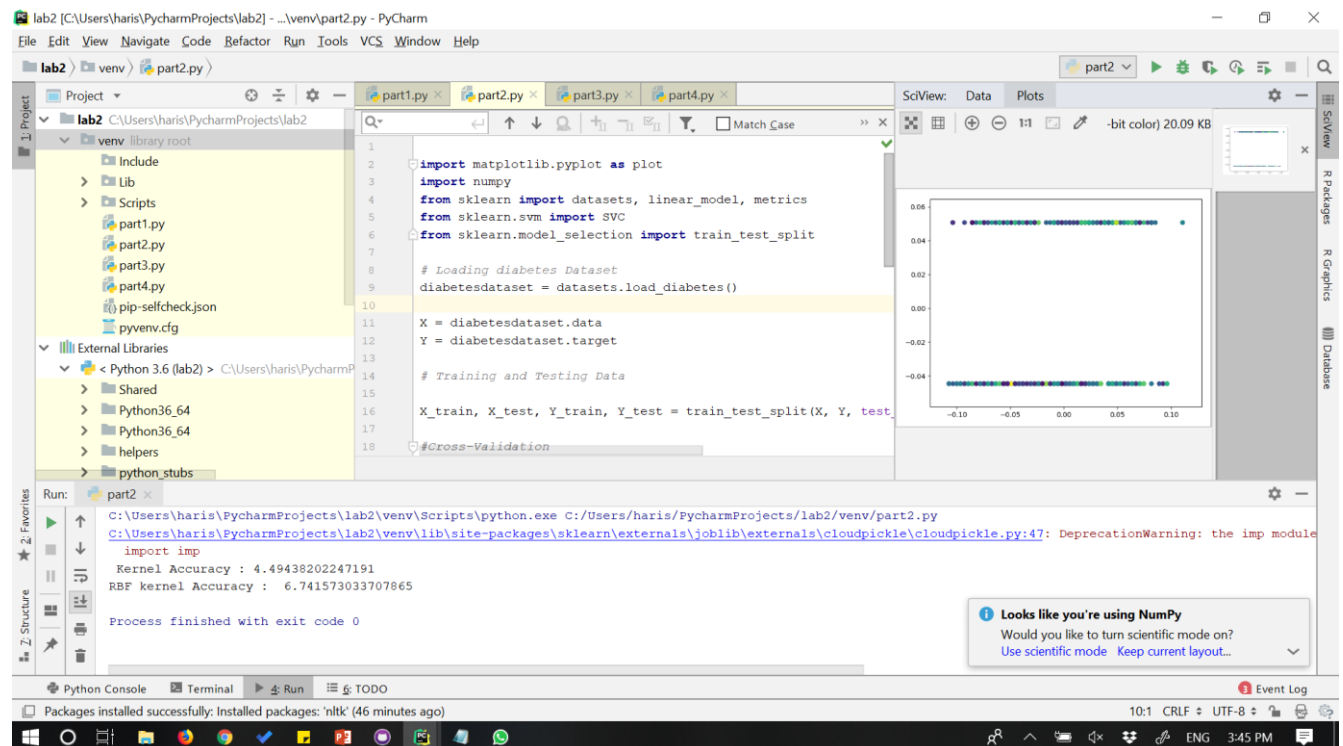
we have choose diabetes as our dataset and performed naive bayes classification and evaluated our model and printing the test accuracy of the model.



## 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:

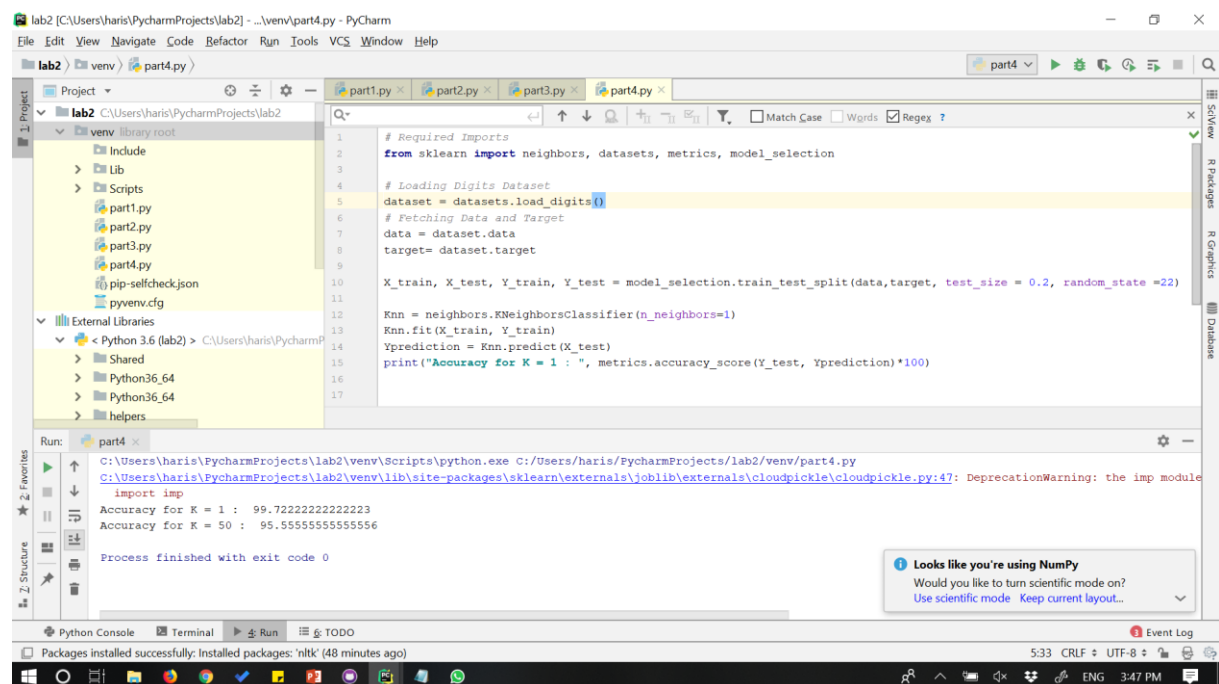
### Performing summarization on the text data:

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 concatenated the sentences.





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



```
1 # Required Imports
2 from sklearn import neighbors, datasets, metrics, model_selection
3
4 # Loading Digits Dataset
5 dataset = datasets.load_digits()
6
7 # Fetching Data and Target
8 data = dataset.data
9 target = dataset.target
10
11 X_train, X_test, Y_train, Y_test = model_selection.train_test_split(data, target, test_size = 0.2, random_state = 22)
12
13 Knn = neighbors.KNeighborsClassifier(n_neighbors=1)
14 Knn.fit(X_train, Y_train)
15 Yprediction = Knn.predict(X_test)
16 print("Accuracy for K = 1 : ", metrics.accuracy_score(Y_test, Yprediction)*100)
17
```

Run: part4

```
C:\Users\harris\PycharmProjects\lab2\venv\Scripts\python.exe C:/Users/harris/PycharmProjects/lab2/venv/part4.py
C:\Users\harris\PycharmProjects\lab2\venv\lib\site-packages\sklearn\externals\joblib\externals\cloudpickle\cloudpickle.py:47: DeprecationWarning: the imp module
import imp
Accuracy for K = 1 : 99.7222222222223
Accuracy for K = 50 : 95.5555555555556
Process finished with exit code 0
```

Looks like you're using NumPy  
Would you like to turn scientific mode on?  
[Use scientific mode](#) [Keep current layout...](#)

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: