Assignment 1

Secition A:

Ans- 1

a) To Prove that linear regression the sq, fit line will pass through $(\bar{X}, \bar{Y})$

In linear Regression

$$Y_i = W^T X_i$$

By MLE estimate we know the loss function will be

$$L = \frac{1}{m} \sum_{i=1}^{m} (W^T x_i - \hat{y}_i)^2$$

$x_i \Rightarrow$ dataset

$y_i \Rightarrow$ given value

Since we need to minimize this function

To minimize it $\frac{dL}{dw} = 0$

$$\frac{d}{dw} = \frac{1}{m} \cdot \frac{2}{m} \sum (w^T x_i - y_i) x_i = 0$$

$$\frac{2}{m} \sum_{i=1}^{m} (w^T x_i - y_i) = 0$$

$$w^T \frac{\sum x_i}{m} - \frac{\sum y_i}{m} = 0$$

$\frac{\sum x_i}{m} = \bar{X}$ AM of independent variable

$\frac{\sum y_i}{m} = \bar{Y}$ AM of dependent variable

$$w^T \bar{X} - \bar{Y} = 0$$

Hence $(\bar{X}, \bar{Y})$ will always pass through best fit line

b) Given 3 variables X, Y, Z given that X is in high correlation with Y also given that X is in high correlation Z and we need to find whether Y, Z are also in high correlation or not

It is not true that X and Z will also highly correlated.

Example : Lets Say X represents ethinicity
Y represents Height of person
Z represents Color of the individual

Lets Say Increasing the height belong to particular ethinicity hence forming a high correlation among X and Y.

Also we can say that People belonging to one typical ethnicity will have high chances to be having one type of Color. But it will not establish that people who are tall will belong to one type of color or vice versa hence Correlation cannot be transitive.

c) Weak laws of large numbers suggests that

$$P[[\bar{X} - \mu] > \varepsilon] \leq \frac{Var(\bar{X}_n)}{\varepsilon^2}$$

where $\varepsilon > 0$ μ = mean

$\bar{X}$ = Population mean

and when $n \to \infty$

$$P[(\bar{X} - \mu) - \varepsilon] = 0$$

$$\bar{X}_n = \frac{1}{N} \sum X_i$$

$$Var(\bar{X}_n) = \frac{1}{N^2} \sum Var(X_i)$$

$$\text{Assu } Var(X_i) = \sigma^2$$

$$\sum Var(X_i) = N\sigma^2$$

$$Var(\bar{X}_n) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$$

$$P[(\bar{X}_n - \mu) > \varepsilon] \leq \frac{\sigma^2}{N\varepsilon^2}$$

$$N \to \infty \quad \frac{\sigma^2}{N\varepsilon^2} = 0$$

Since $P() \geq 0$ $\quad P[(X_n - \mu) > \varepsilon] \leq 0$

$$P[(\overline{X}_n - \mu) > \varepsilon] = 0$$

$$X_n \rightarrow \mu$$

Suppose a bernoulli distribution of X

with $p = 1/2$

$$X = \begin{cases} 1/2 & x = 0 \\ 1/2 & x = 1 \end{cases}$$

$$\mu = 1/2 \times 1 + 1/2 \times 0 = 1/2$$

Pseudo code

```
n # large number
sum = 0
for i (0, n)
    sum += random (0, 1)
pop_mean = sum / n
epsilon = 0.001
if ( absolute ( pop_mean - 1/2 ) < epsilon )
    print("Hence Proved")
```

d) MAP solution for linear regression with gaussian prior

With MAP estimate

$$P(w/D) = \frac{P(D/w) \cdot P(w)}{P(D)}$$

We need to find estimate of weights and since $P(D)$ is not dependent on it we can remove $P(D)$ from it.

For linear regression we have $P(D/w)$

$$= P(y_i / w, x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{w^T x - u}{\sigma}\right)^2}$$

(u, σ^2) parameters of gaussian noise we

$$P(w) = \frac{1}{\sqrt{2\pi\sigma_w^2}} e^{-\frac{1}{2}\left(\frac{w}{\sigma_w}\right)^2}$$

$$P(w/D) = \prod \left(\frac{1}{\sqrt{2\pi\sigma_w^2}} e^{-\frac{1}{2}\left(\frac{w}{\sigma_w}\right)^2} \right) \cdot \prod_{j=}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}$$

Taking log and removing $e^{-\frac{1}{2}\left(\frac{w^T x - y_i}{\sigma}\right)^2}$ constants

$$-\sum \left(\frac{w}{\sigma_w}\right)^2 + \sum_{j=1}^{n} \left(\frac{w^T x_i - y_i}{\sigma}\right)^2$$

$=$

$$L = -\sum (w^T x_i - y_i)^2 - \sum (w)^2$$

$$w_{i+1} = w_j - \alpha \left(\frac{\partial L}{\partial w}\right)$$ → rate of change of loss

$$\frac{\partial L}{\partial w} = -2\Sigma(w^T x_i - y_i) x_i - 2\Sigma(w)$$

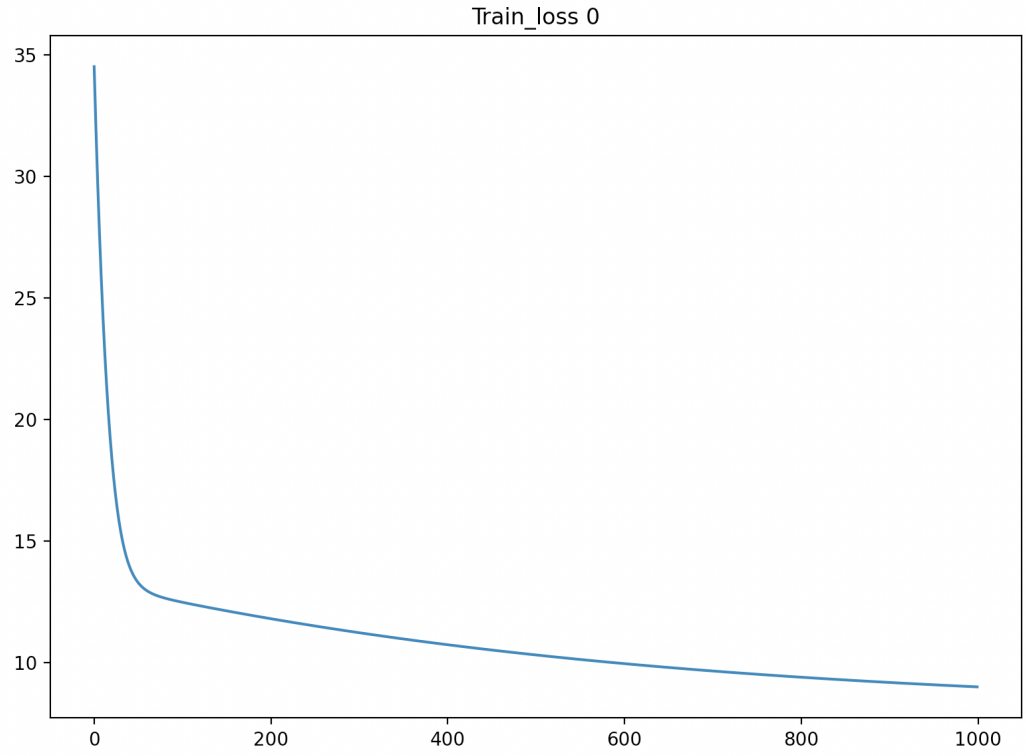
Which will be the new update rule for the weights

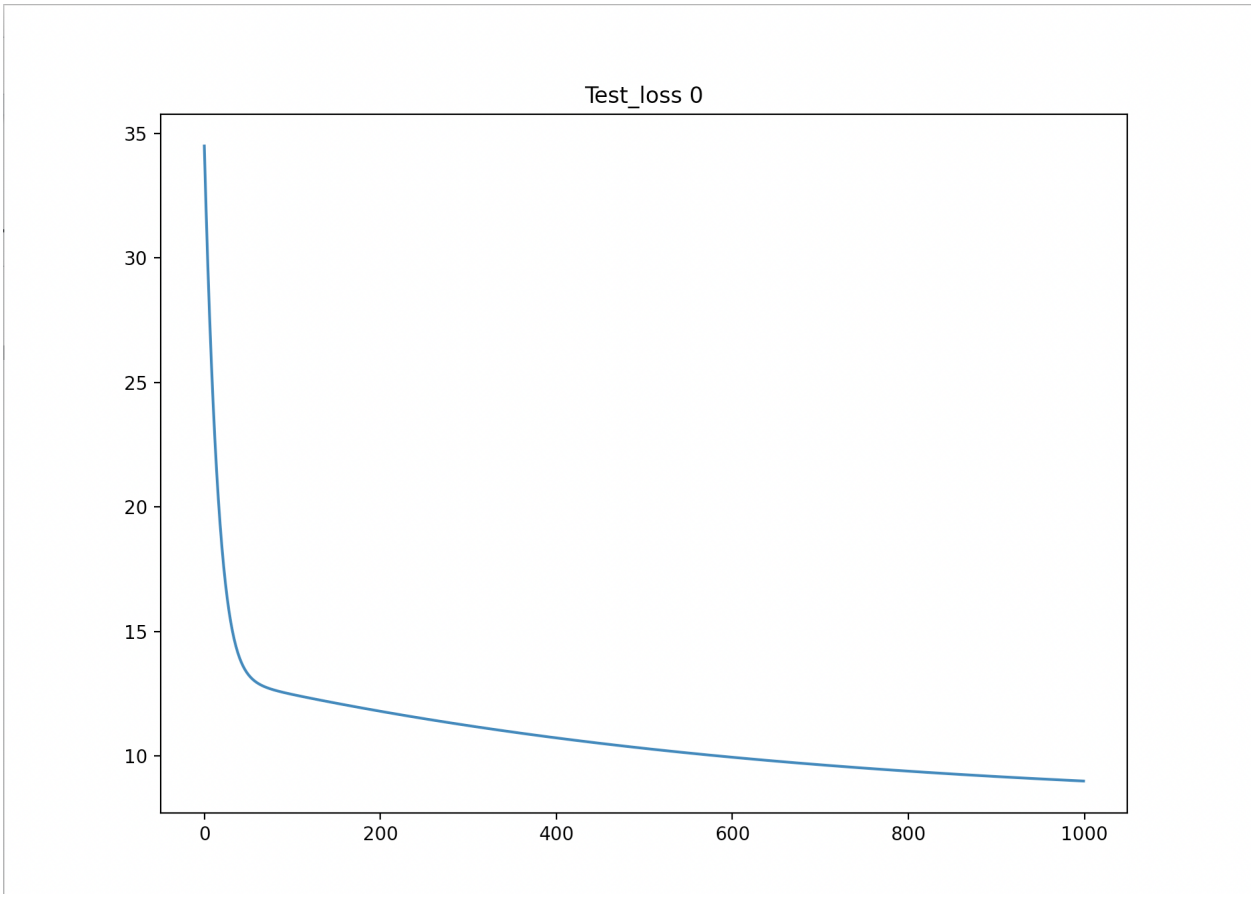
Section B:

1) For the given dataset, we apply Linear Regression analysis and try to find the optimal value of weights to find the best possible curve. In the code, we first initialize them with 0 weights and update the parameters as per mse loss function. For k-fold validation, we divide the dataset into k divisions and train it against the k-1 division and test it against one of the divisions and change the value of k from 2-5 and find the best optimal loss. For this, we get the minimum error in the case of k=5

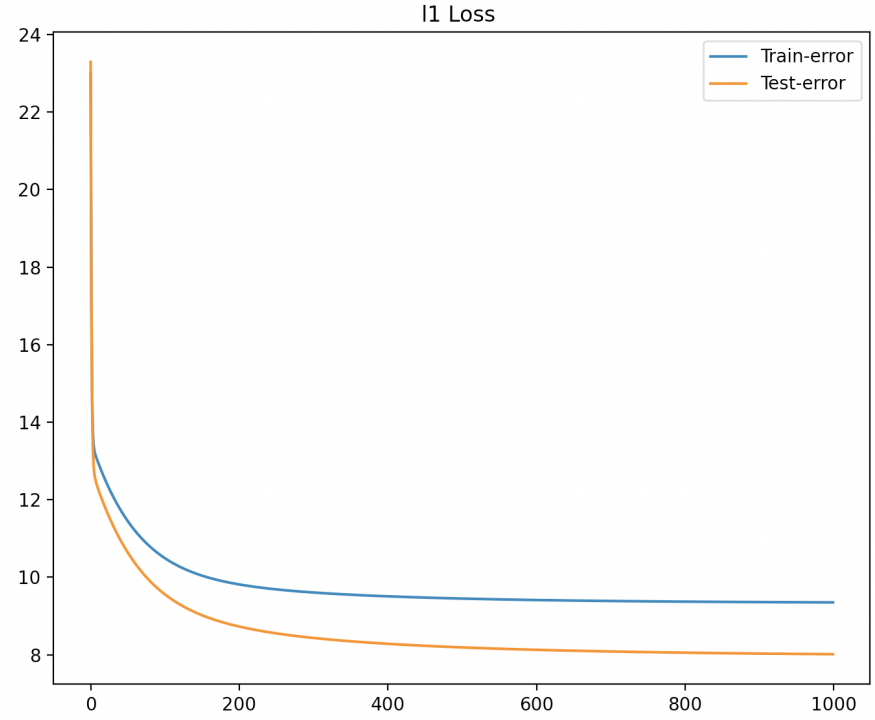
K-fold value	Loss value
2	9.158462245597544
3	9.114029363414012
4	9.07093568895518
5	8.99815711192188

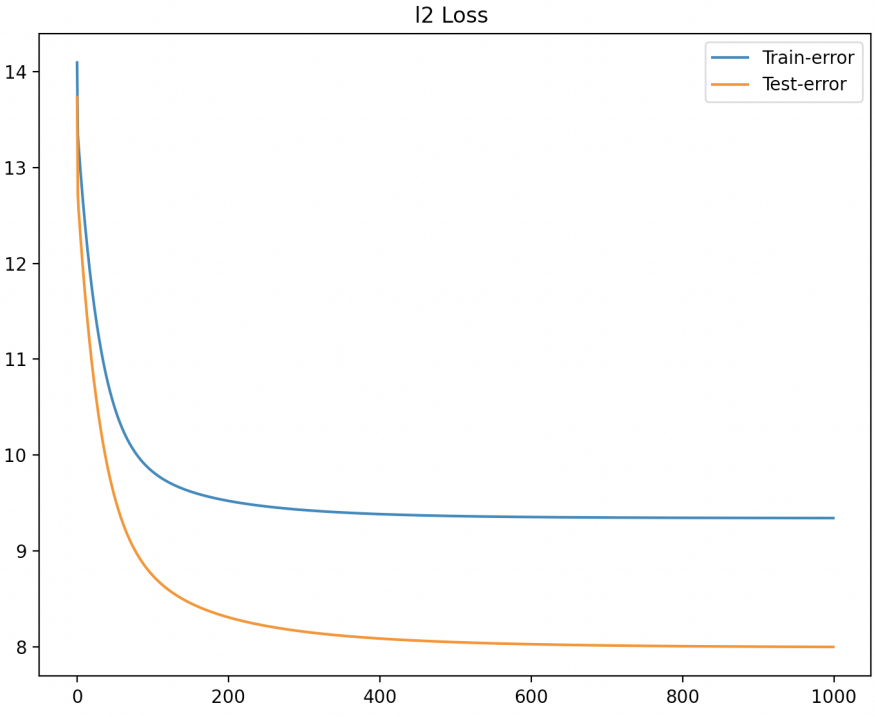
2) We have the best result when k=5, and for that, we find the graph of rmse training loss vs. iterations along with testing loss vs. iterations and plot the curve, which is as follows for such as that-


Train_loss 0
35
30
25
20
15
10
0
200
400
600
800
1000



3) In this part, we have modified the linear regression by adding l1 regularization for l1 regularization, adding another parameter to the loss function of the modulus of weights as a penalty, and in the case of l2, adding the penalty of the square of weights as the same. Changing the update rule as per the given regularization and plotting the training set loss and the validation set loss for the same.


l1 Loss
Train-error
Test-error
24
22
20
18
16
14
12
10
8
0
200
400
600
800
1000

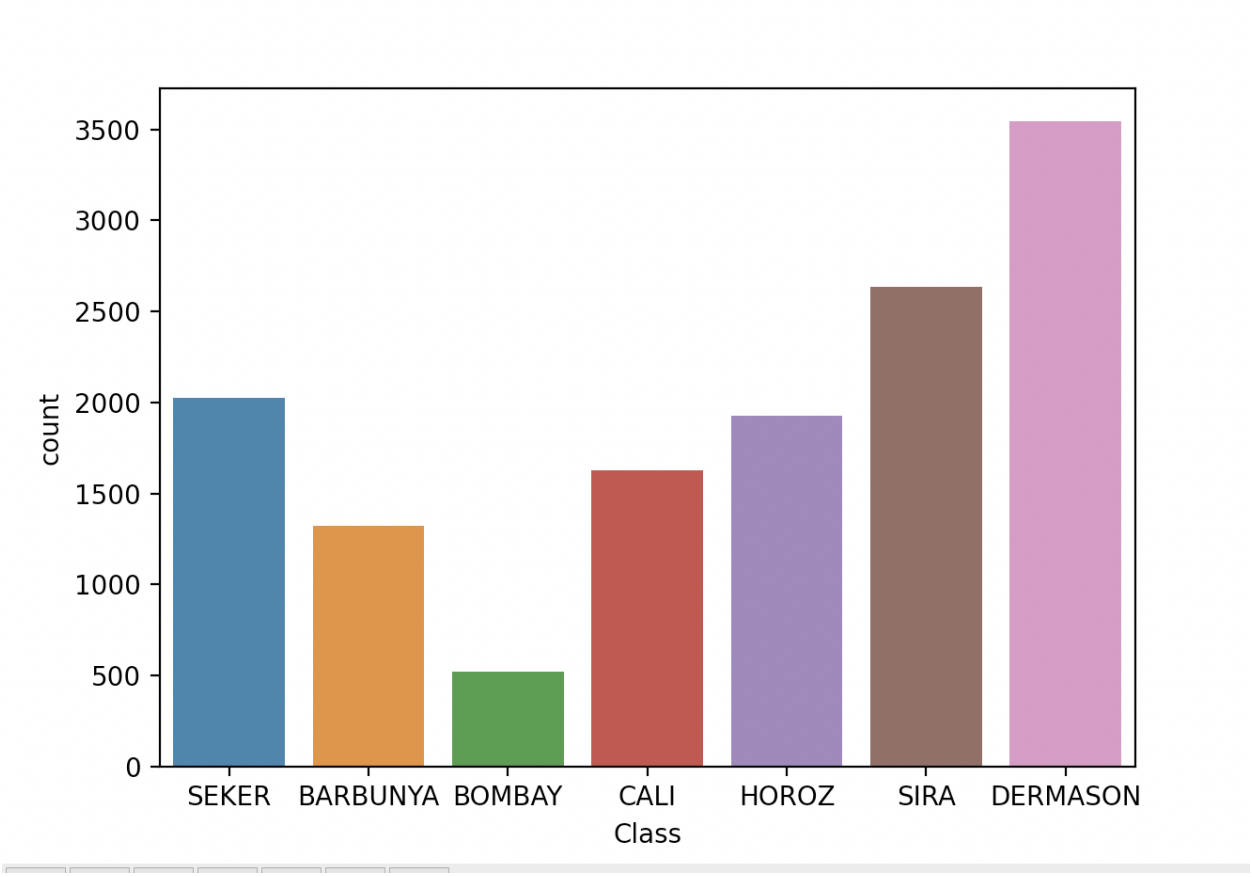

l2 Loss
Train-error
Test-error
14
13
12
11
10
9
8
0
200
400
600
800
1000

4) The normalization rule suggests instead of finding the minima to the loss function directly, finding the optimal value of weights as per mse loss function and obtain the value of weights as $(X.X^T)^{-1}.(X.Y)$. After that appling it on k-fold cross-validations on appropriate values of k. Then divide the dataset into k subparts where k-1 parts other part as testing set and finding the loss against each division.

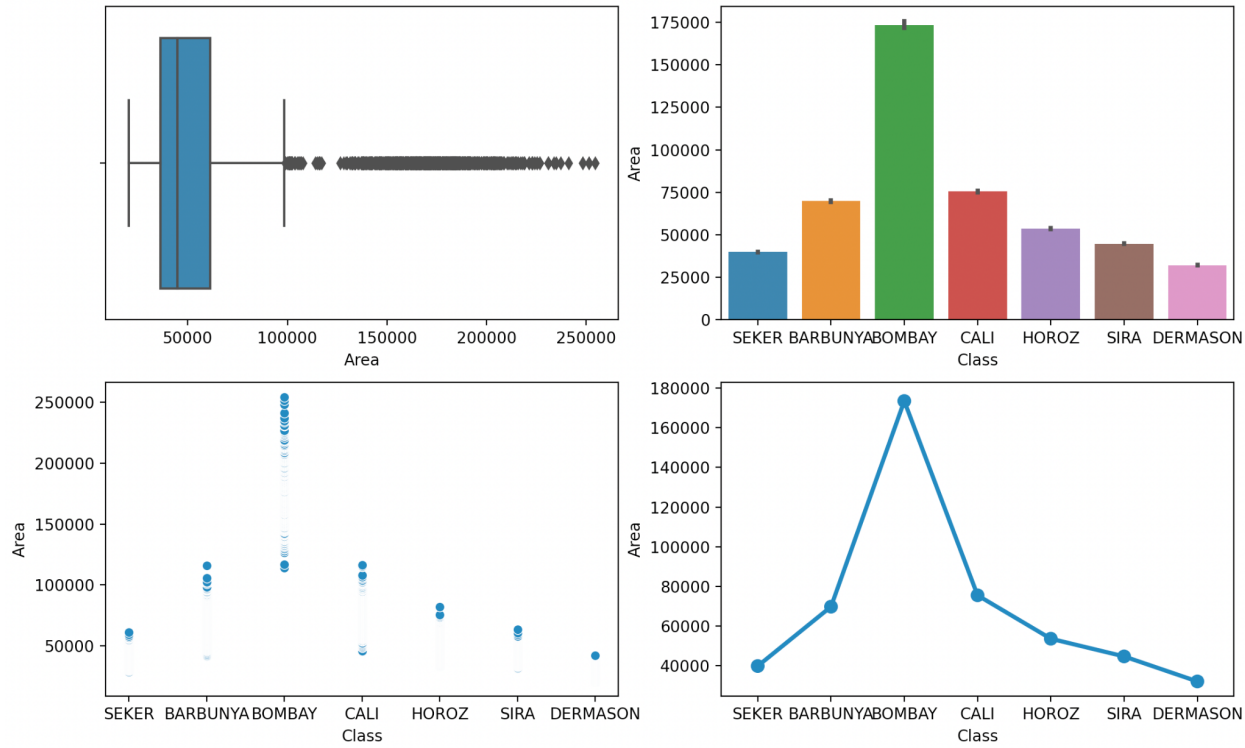
Validation set no.	RMSE loss on set
1	6.76
2	9.52
3	6.8
4	11.48
5	7.53

Section C (Bonus Component):

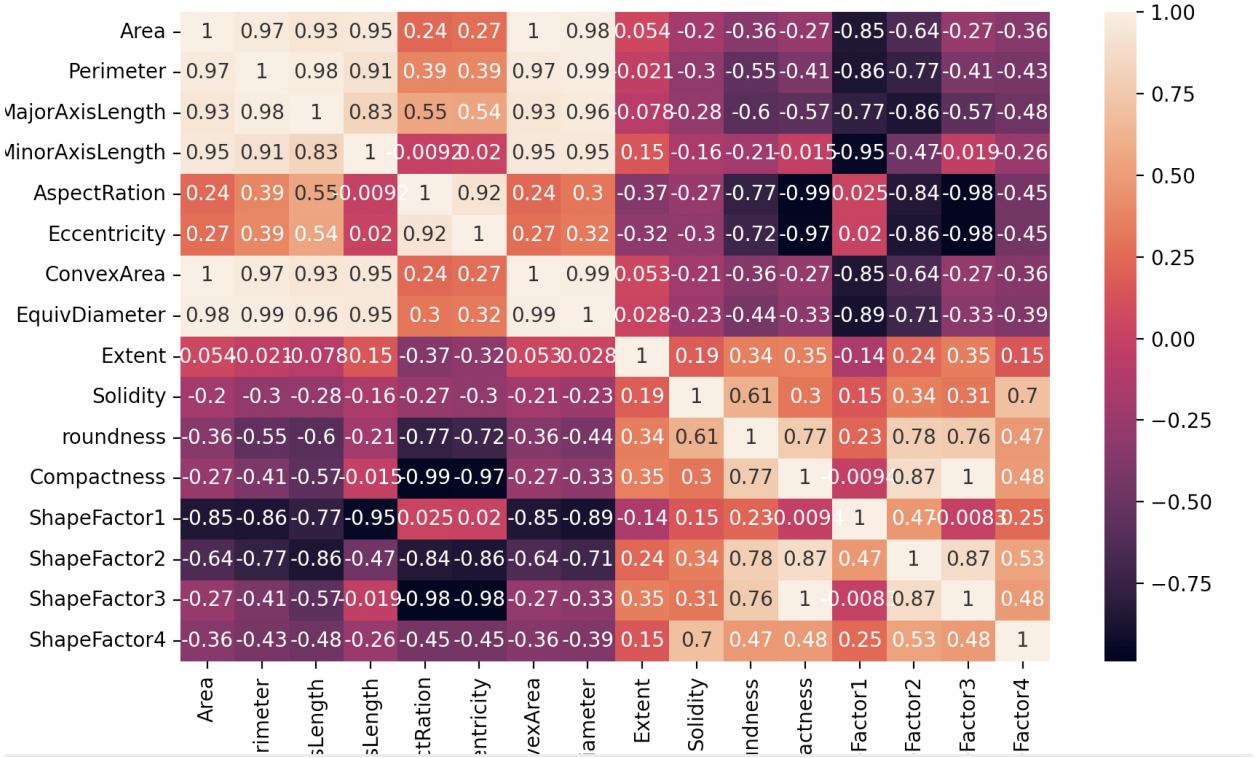
1) For the given dataset we find the find the no of samples in each class using count plot over the dataset using seaborn.



2) Appling Explainatory Data Analysis over the dataset
Boxplot: Helps us to understand the distribution of the data for an attribute.
Barplot: Helps us to understand the value corresponding to that class for a particular attribute.
ScatterPlot: Helps us to understand between what values of the attribute is the class found.
PointPlot:Helps us to understand the central tendency of the data.

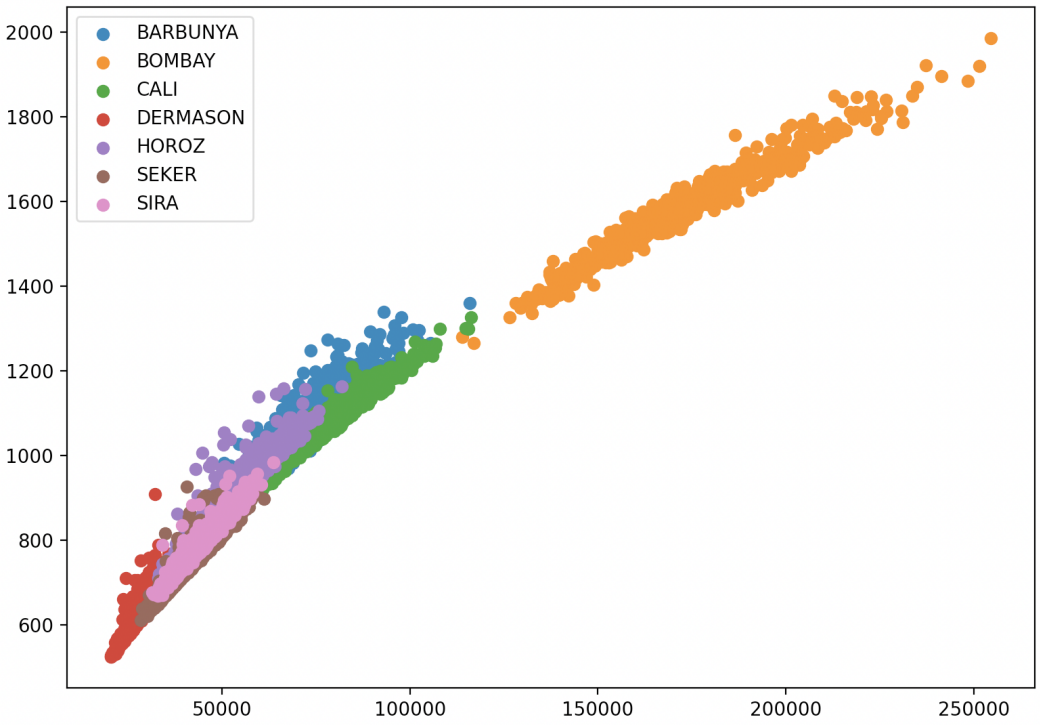


HeatMap: Helps to better visualize the covariance matrix.



The total number of missing values in the dataset are 0 for each column.

3) Tnse indicates that BOMBAY class can be easily separated from rest of the dataset rest class have some or the other form of intermingling present as well as some overlaps are also there.



4) For this part using multinomial naive bayes and Guassian naive bayes distribution. Basic internal working of both the implementation revolves around the basic assumption of naive bayes finding the maximum amoung all classes the posterior probability and then given on the conditions tring to find the best possible class. In gaussian it has another assumption that each class follows gaussian distribution. Whereas in case of multinomial naive bayes it assumes multinomial distribution for the classes (Metrics tables are in the last along with logistic regression metrics).

5) Using sklearn's pca library for dimensionality reduction and then training the model on logistic regression model finding accuracy metrics and finding which gives the best results
For n=4

Performance metrics	Rate
Accuracy	0.72
Precision	0.76
Recall	0.78
F1	0.75

For n=6

Performance metrics	Rate
Accuracy	0.73
Precision	0.77
Recall	0.78
F1	0.76

For n=8

Performance metrics	Rate
Accuracy	0.67
Precision	0.71
Recall	0.73
F1	0.68

For n=10

Performance metrics	Rate
Accuracy	0.73
Precision	0.76

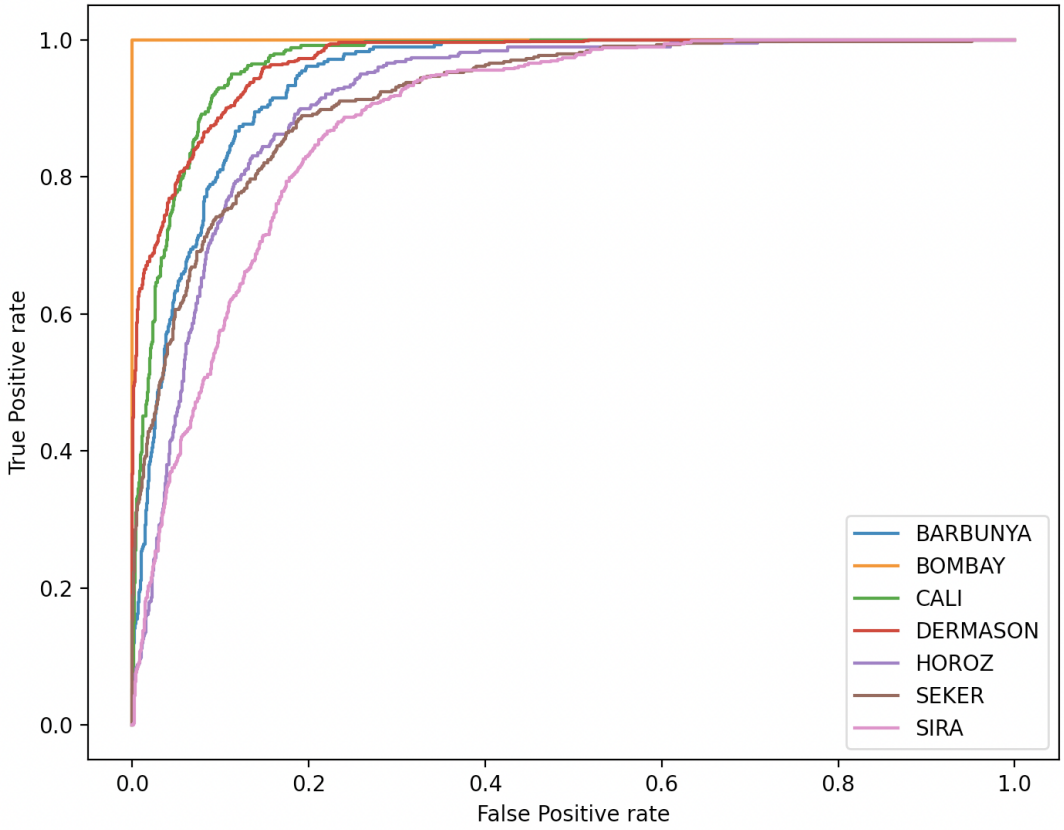
Recall	0.78
F1	0.74

For n=12

Performance metrics	Rate
Accuracy	0.69
Precision	0.72
Recall	0.77
F1	0.72

From the above tables we can see that the best result is given by n=6 components

6) ROC-AUC curve helps us to analyze which class has been analysed with the best probability by seeing the area under it's curve using false positive rate and true positive rate higher the area under the curve higher is the performance of the model to classify that class.



7) Using sklearn's logistic regression model to train the data based on multiclass classification and iterated only 100 times. The metrics were as followed

For MultinomialNB:

Performance Metrics	Rate
Accuracy	0.79
Precision	0.78
Recall	0.79

For GuassianNB:

Performance Metrics	Rate
Accuracy	0.76
Precision	0.76
Recall	0.76

For Logistic Regresssion

Performance Metrics	Rate
Accuracy	0.72
Precision	0.71
Recall	0.72