Data Science Project Report

!) Information About the dataset. :

The dataset is of bank notes which has two column as V1 and V2 . this column contains values which are necessary for the detection of quality of notes.

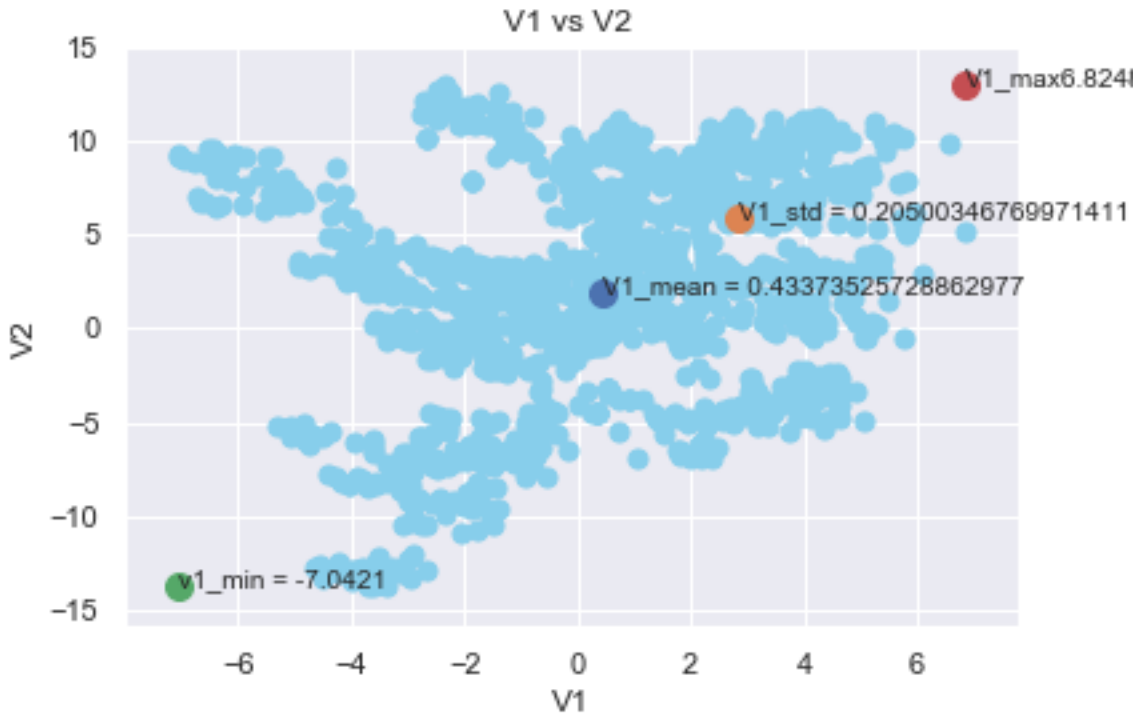
The resulting dataset has 1372 observation. i.e. we have information from 1372 images of both genuine and forged banknotes. while data from more images would be welcome to improve the accuracy of any models we make, the 1372 we have is large enough to give us an idea of whether this task is possible or not.

Here we can see statistical maeasures of data sets are as follows :

```
x_min = -7.0421
y_min = -13.7731
x_max = 6.8248
y_max = 12.9516

x_mean =  0.43373525728862977
X_std = 2.8427625862451658
y_mean = 1.9223531209912539
y_std = 5.869046743580378
25%      x = -1.773000    y = -1.708200
50%      x =  0.496180    y =  2.319650
75%      x =  2.821475    y = 6.814625
```

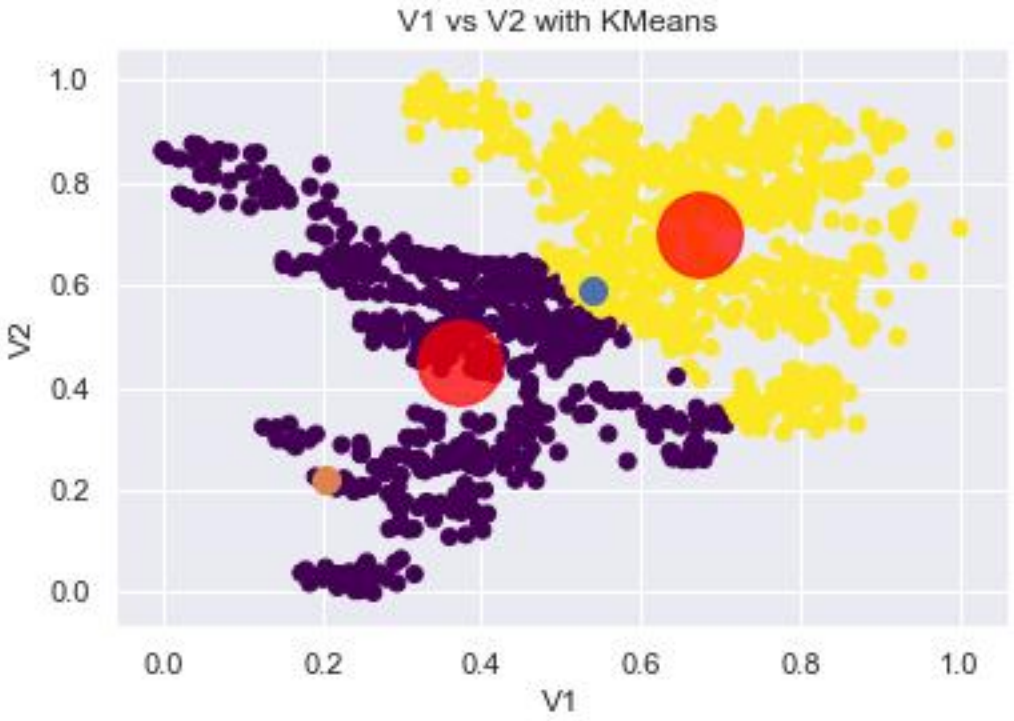
2) visualization of Data Points. :



From the given plot we can see that it is difficult to find any kind of relation between the points , so it is the perfect dataset for he k-means clustering project. The dataset also do not contain any kind of null value in it , so it is ready for k-means project after normalization.
From the graph we can not see any correlation in the data points.

3) Choosing the kMeans clustering:
Kmeans is unsupervised learning algorithm means this algorithm is used for data which do not contain any labels. This dataset also do not contain any labels. Therefore we choose kmeans for clustering.
Here k is no. of clusters here basically two qualities of notes.first is good quality and another one is great quality.

4)Applying Kmeans clustering :



From the graph we can easily two clusters of different quality of bank notes. Purple color is good quality and yellow is great quality. Red dots are the center of this cluster.
Here green dot is mean of the dataset. And the orange dot is standard deviation . This graph is shown after the process of normalization.

5) Summery of Kmeans result :

After re – running the code several times I did not see any kind of change in the graph . so I conclude that the Kmeans is stable.

If we increase the number of clusters , then we increasing the categories of the notes . for example if we shows three clusters then then categories will be good, intermediate and advance .

But we only shows Two categories good and great. Because we have only two features of data . by doing this we are increasing stability of algorithm.

6) recommendation for client :
This is good model for the for the authentication , But I still recommend that more data features are collected from the images of the banknotes and these

incorporated into the clustering model to possible improve the accuracy of the model. I would suggest that completely automating the detection of forged banknotes would not return accurate results for every banknote. Perhaps a less risky solution would be to use models such as these to identify banknotes suspected of forgery and further manually examine these. However even with this, you would run the risk of not identifying some forged notes should they be grouped as genuine.