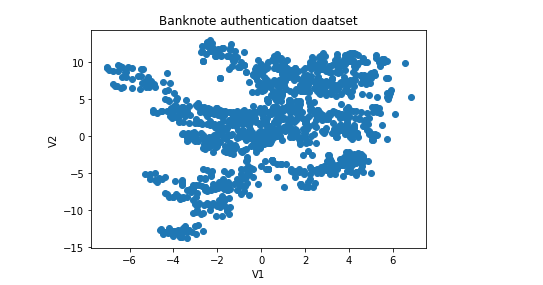
**1.Purpose of the Data Science Project:**

To write a report for a bank that is considering automating the detection of forged banknotes. As the data values of given CSV file is big and we cannot see the relation between variables. So we use Data Science to analyse, understand, evaluate and run KMeans to check the given Banknote is genuine or not.

**2.Description of Data:**

In the Banknote Dataset authentication file, they have given two features called V1 and V2. Variance method values are V1 and Skewness method values are V2. These methods are used for finding the given Banknote is Genuine or not. We have size of 1372 values and no missing values include. The data values are in numerical format with negative and positive values.

We can plot the Values of V1 and V2 in the graph as below:



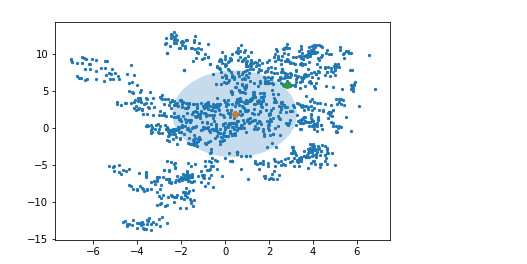
Mean of Variance method V1 is 0.43373

Mean of Skewness method V2 is 1.922353

Standard Deviation of V1 is 5.86690

Standard Deviation of V2 is 5.86990

We can plot the Values of V1 and V2 with Mean and Standard Deviation as Ellipse patch as below. This graph is used to identify Outliers.



**3. methods we used: How data were analysed:**

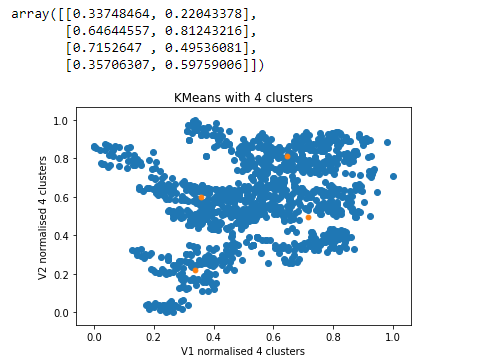
**3.1 Normalisation:**

Now we can see, the points are spread widely outside of Ellipse so there is high chance of outliers. We cannot find any relation between V1 and V2 values as it spread widely. So we normalised data and we plot normalised data values.

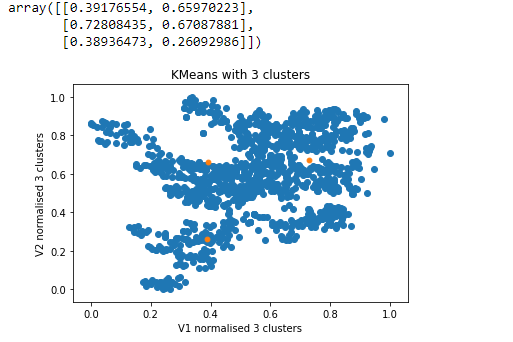
**3.2. Check for K\_Means:**

First we have to take number of clusters randomly and rerun until it stabilised the Centroids. So I choose number of clusters as 4 ,3,2, and run the normalised data and graphs are below

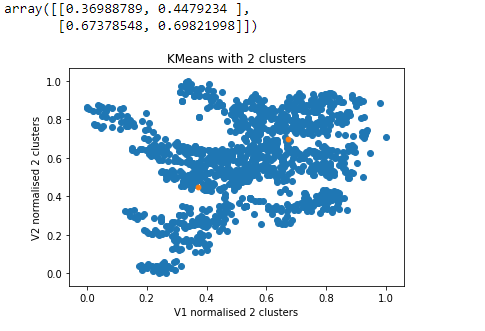
Plot of data with 4 clusters with centroid points are below:



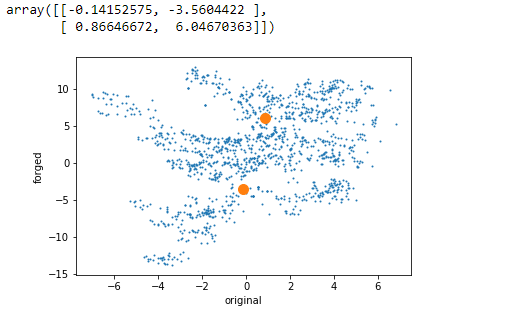
Plot of data with 3 clusters with centroid points are below:



Plot of data with 2 clusters with centroid points are below:



Plot of data with 2 clusters with original data are below:



**3.3 Finalising clusters:**

As we noticed from above graphs, Centroid points of K=4, K=3, k=2, we can see stability from K=2 , k=3 and between k=3 ,k=4 centroid points are varying more. So we have to choose k=2 as current number of clusters to run KMeans on the given data.

**4.Summary of results:**

For the given Banknote dataset values of Variance and Skewness method to find the note is genuine or not, as the values are varying widely we apply normalisation to the data and run the Kmeans on the data with 2 clusters. We got 2 clusters with 2 centroids and we can take 1 cluster as genuine and 1 cluster as fake notes. But we cannot find with one is genuine and which one is fake as we have enough data.

**5. Reccomendation for Client:**

As looking at values we cannot see the relation of V1 and V2 , we plot the values in the graph to see the relation of data. By looking at points in graph we can say as it spread widely so there is a chance of outliers . So we normalised data and run KMeans with 2 clusters and we got 2 clusters one for Genuine and one for Fake. So I recommend to use this Data Science method to understand, analyse, visualise, eyeball the data to find the genuine notes or not.