**Report on KMeans clustering to identify forge banknote**

**Introduction:**

In this exercise, a data set collected from bank notes some are forge and some are original. KMeans clustering is used to train and identify the both cases of bank note. The accuracy of the analysis is limited to the accuracy of the analysis and the number of samples.

Suitability of the data set:

In order to evaluate the data set the following statistical evaluation was done:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Property | Variable 1 | Variable 2 |  |  |
| Mean | 0.43373525728862977 | 1.9223531209912539 |  |  |
| median | 0.49618 | 2.31965 |  |  |
| Min | -7.0421 | -13.7731 |  |  |
| Max | 6.8248 | 12.9516 |  |  |
| Stdev | 2.841726405206097 | 5.866907488271993 |  |  |

Stdev shows that the data are probably suitable for the analysis with less extend the variable 2.

Before analysis the data has been normalized so that all dataset is between 0 and 1.Without normalization the algorithm will fail.

The kmeans algorithm was run with 2 centroids on the given dataset. It was repeated several times in order to confirm the stability of the analysis. This is another factors that tells us that most likely the analysis is with good confidence.

The x and y coordinates of the centroids are below:

X y

Cluster 1 centroid 0.67364922 0.69798136]

Cluster 2 centroid 0.36955897 0.44781178

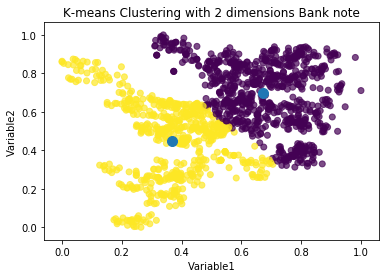


Figure calssification based on kmeans algorithm

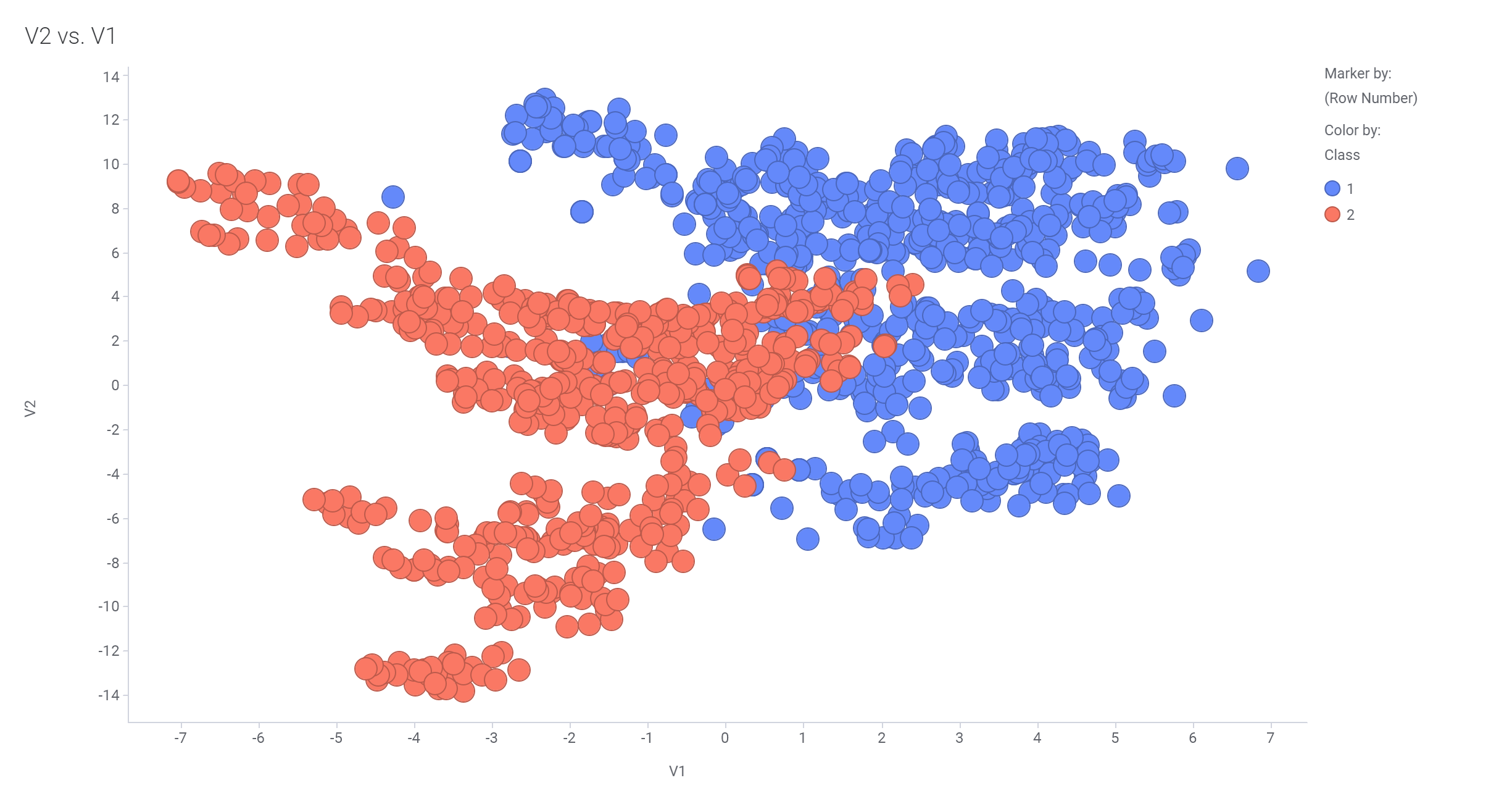


Figure classification of dataset from actual observation

Comparing two figures we can see high accurate classification of the kmeans clustering but not perfect. Adding the remaining variables (3 and 4) would help make the analysis more robust.

**Conclusion**: Kmeans clustering can be used to identify forge notes with high confidence. The more data we have the better the analysis.