Report on Forged Banknotes

The purpose of the project

The project was designed with the purpose of understanding better the given data on forged banknotes, exploring it, and bringing into light whether Data Science can provide us with help on understanding how to separate forged banknotes from authentic ones or not. In order to do that, the analysis consisted mainly in finding the statistical measures from the specific dataset and the attempt of using the technique of clustering to see whether is can be of use for this purpose or not.

The structure of the data analysed

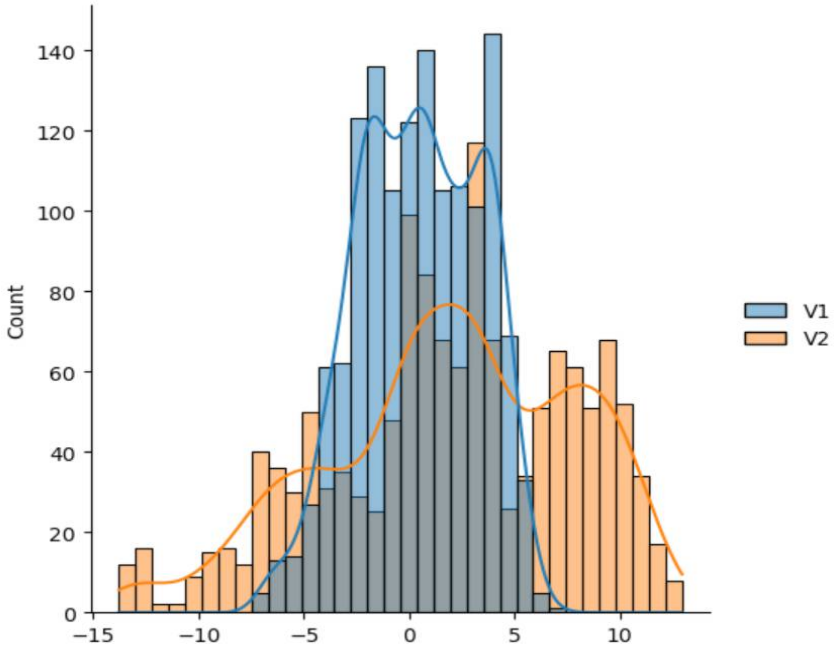
The data in context consists of two variables, or columns, namely: V1 or variance of Wavelet Transformed image (continuous), and
V2 or skewness of Wavelet Transformed image (continuous). Moreover, the data consists of 1372 observations.

	V1	V2
0	3.62160	8.66610
1	4.54590	8.16740
2	3.86600	-2.63830
3	3.45660	9.52280
4	0.32924	-4.45520
...	...	...
1367	0.40614	1.34920
1368	-1.38870	-4.87730
1369	-3.75030	-13.45860
1370	-3.56370	-8.38270
1371	-2.54190	-0.65804

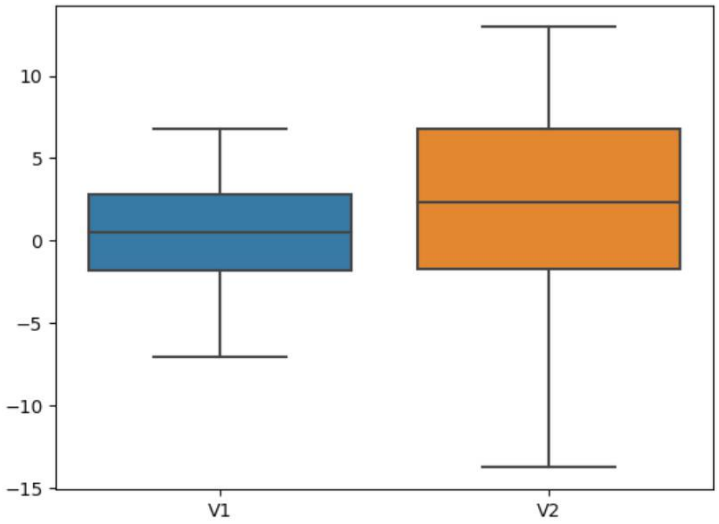
1372 rows × 2 columns

	V1	V2
count	1372.000000	1372.000000
mean	0.433735	1.922353
std	2.842763	5.869047
min	-7.042100	-13.773100
25%	-1.773000	-1.708200
50%	0.496180	2.319650
75%	2.821475	6.814625
max	6.824800	12.951600

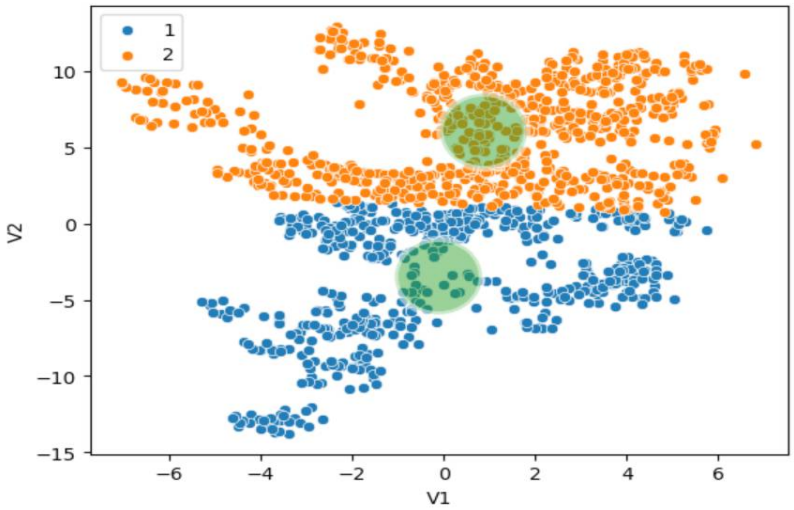
While the first image on the left offers a preview of how the data actually looks like, the image on the right provides a description of statistical measures of the same data. Important observations from the description are the differences between the mean and the median (expressed here as 50%), which suggest that both columns result in a slightly left skewed plot. Meanwhile the standard deviation, or std, proposes a somewhat higher variability for V2, which suggests that the observations are more spread out. This can also be seen while looking at the interquartile, that is the difference between the values of 75% minus 25%, as well as the distance between the maximum and minimum values. In order to look better at this description, two ways of visualizing were initially chosen: a box plot and a distribution plot.



The distribution plot above compares the distribution between V1 and V2. It's easy to see how much more centralized V1 is while V2 is far more spread out.



Meanwhile the box plot above also shows with clarity the difference in spread of data between both. The interquantile is also more spread for V2. The box plot helps us see that there are no outliers in neither variables. The lack of outliers allows us to study better with K-Means Clustering, without being much affected. The clustering will attempt to group the data into two centers representing values of authentic and forged banknotes in order to see if in fact data clustering can help identify the notes more easily.



Finally a scatter plot is designed with two clusters present following the values.

The legend proposes that the observations in cluster n.1 can be identified with that which would be the genuine banknotes, while cluster n.2 presents the a cluster of likely fake notes. The scattering allows us to see distinctions in the distributions and found values between genuine and forged notes.

Recommendation or Conclusion

As it can be seen, the use of K-Means clustering can indeed help identify banknotes that are forged apart from the genuine ones. That is because they present differences in values that explain the variability in the data. There are, however, strong limitations in the use of this technique. Although clustering identifies trends, all it does is calculate the distance from an observation and the center of a cluster. This can help increase the probability that an observation is forged, but criterion is still mathematical. All we see are patterns the rise red flags and can do it more quickly, but Data Science cannot really tell whether it is in fact forged or not. The data is helpful, but complementary.