Data Exploration

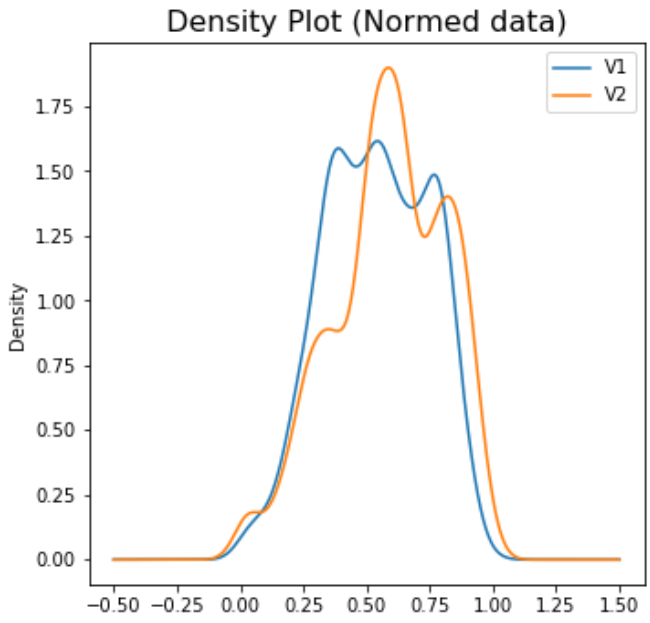
Comparing the Distributions of the Wavelets (V1 and V2)

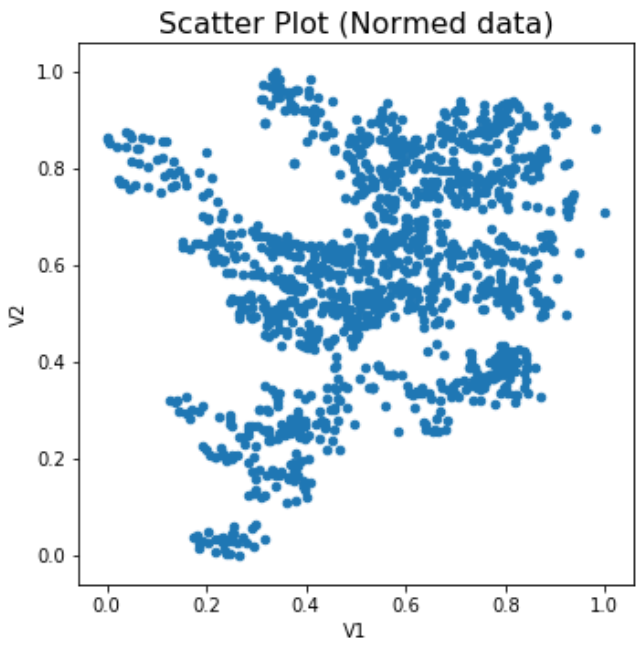
Summary of Raw Data

	V1	V2
count	1372.0	1372.0
mean	0.434	1.922
std	2.843	5.869
min	-7.042	-13.773
25%	-1.773	-1.708
50%	0.496	2.32
75%	2.821	6.815
max	6.825	12.952

Summary of Normed Data

	V1	V2
count	1372.0	1372.0
mean	0.539	0.587
std	0.205	0.22
min	0.0	0.0
25%	0.38	0.451
50%	0.544	0.602
75%	0.711	0.77
max	1.0	1.0





The dataset provided includes two observations (named V1 and V2) made on each of 1372 banknotes, some genuine and others counterfeit. The observations (called "wavelets") summarise two features extracted from the scan of each banknote. The dataset consists of a comma-separated file with one row for each banknote scanned. The dataset was loaded into a Jupyter notebook and explored using Python.

Key findings of the exploration are summarised in the figure above, which is composed of four facets: two tables and two graphs. A brief description of each facet follows, as a basis for judging whether this dataset is suitable for analysis by a clustering algorithm:

1. The first table (top-left) shows summary measures for the raw data. It can be seen that V2 (with a minimum of -13.77 and a maximum of 12.95) has a much broader range than V1 (which ranges from -7.0

to 6.8). The mean for V2 is more than four times higher than V1 and the standard deviation is twice as high. This suggests that any attempt to perform a cluster analysis on the raw data would produce a model that is much more influenced by V2 than by V1.

2. In an attempt to correct for this effect, both V1 and V2 where each scaled ("normalized") to fit within a range between 0 and 1. The effects of this scaling are shown in the second table. It is clear that the mean (0.54 and 0.59 for V1 and V2 respectively) and the standard deviation (0.21 and 0.22 for V1 and V2 respectively) are now much more comparable. Furthermore, when the two data columns are sorted and divided into four equal groups -- the 25, 50, and 75 percentiles (quartiles) -- the proportions in each group are comparable.

3. These effects are also shown in the density plot, the graph on the bottom left of the figure. This shows the shape of the distribution as a continuous curve between 0 and 1. It further confirms that the two distributions are centred on the same approximate mean, and the variance (indicated by the spread of the body of the curve) is roughly the same. Finally, neither of the curves skews to the left or right, but roughly symmetrical, although the peaks along the way in each curve do show that there are important differences between the two variables.

4. Finally, the scatter diagram allows a visual assessment of the correlation and clustering of V1 and V2. While there is a large U-shaped concavity on the top left of the scatter plot, the overall shape is that of a convex cloud (overall round, if irregular outline), with an internal granular structure. It is immediately apparent that the cloud has a rough axis of symmetry going diagonally from top-left to bottom-right.

In conclusion: the dataset provided by the client, after normalization, results in features comparable in location (mean) and spread (variance, or standard deviation). The two features are scattered in a spatial form that suggests the K-Means algorithm would be suited for investigation of the internal clustering, and for an effort to develop a model than can distinguish genuine from counterfeit banknotes.