**K-Means Bank Note Report**

**Purpose**

The context for this report includes a comparison of two bank notes. One set is true and one set is false. K-means was chosen as the method of analysis to determine whether or not the authentic bank notes could be differentiated from the false bank notes.

**K-Means Overview**

The K-Means approach uses clusters to organize data as part of patter recognition. The “K” simply denotes the number of clusters. The number of clusters can be customized according to the scenario and researchers. Exploratory data analysis may involve the use of more clusters. For this scenario, there were two main groups of interest. The first group was the authentic bank notes, and the second group was the fake bank notes. Given there were only two main groups of interest, the “K” for this analysis was two.

**Python & Jupyter**

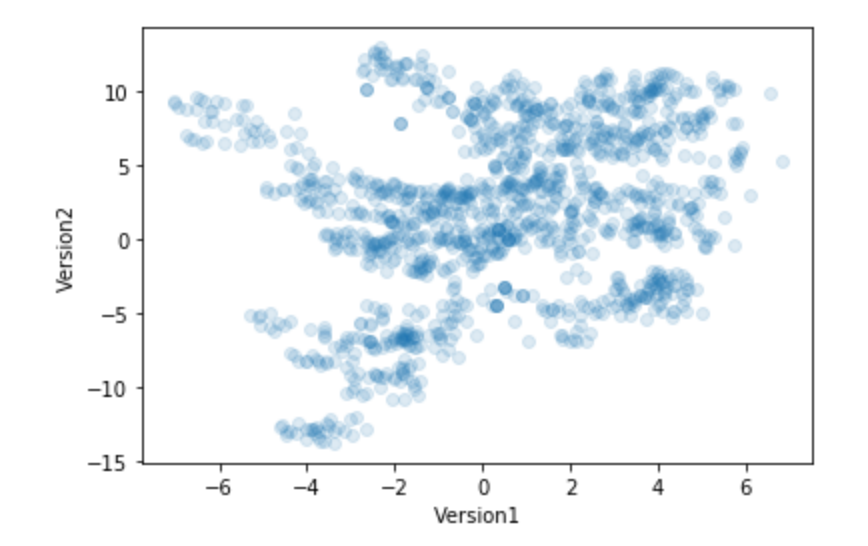
Analysis conducted included Python using Jupyter. Python is an open-source programming software which is heavily across the globe in data science. In Jupyter is a rapidly growing GUI, or graphical user interface used with Python. Currently Jupyter is being adopted to additional programming languages, such as R. R is a data science programming language used in academic and research settings, with an emphasis on statistics and data visualization.   
 Python has several data visualization software packages and tools. These include Dash as well as Seaborn, among others. The package used for this project was simply matplotlib. This package worked well with Python given the scatterplots and clustering processes used.  
  
**Data Analysis**

The initial data provided included two columns of data. There were two versions of bank note values, Version1 and Version2. Overall there were 1,372 rows of data. The following table summarizes the descriptive statistics for Version1 and Version2:

|  |  |  |
| --- | --- | --- |
|  | Version1 | Version2 |
| Total Rows | 1,372 | 1,372 |
| Maximum | 6.8248 | 12.9516 |
| Minimum | -7.0421 | -13.7731 |
| Standard Deviation | 2.8417 | 5.8669 |
| Mean | 0.433 | 1.92235 |
| Sum | 595.084773 | 2637.46848 |

The first step was to do a simple scatterplot of the data. This was completed using Version1 and Version2. Figure 1 highlights the spread of the data. Comparing the max and min values across Version1 and Version2, one difference that quickly becomes apparent is the range difference between the versions. Version1 has a range roughly half that of Version2.

Figure 1. Exploratory Scatterplot of Version1 and Version2 Bank Notes

  
   
 In order to be able to compare the values within Version1 and Version2, then placing the values on the same scale is needed. This normalizing of values between Version1 and Version2 includes the ways in which each data point can be standardized onto a scale between 0-100. The lowest value for Version1 was standardized to 0, and the highest value to 100. The same process was repeated for Version 2.   
 Once the normalization process was completed the scatterplot was revisited. The normalized scales were presented with the normalized values for both Version1 and Version2. Additionally, K-means were used. As mentioned earlier, K-Means were used to classify the categories. For these K-Means analysis, there were two main groups presented, each for Version1 and Version2.   
 Even though the bank note data was normalized, an important aspect from the raw scatterplot indicates one of the groups had a disproportionate amount of negative bank notes. This could be concerning given the extent to which this would not reconcile with the internal bank note data. The bank would have less money on the books than anticipated! Below is a scatterplot highlighting the respective clusters for normalized Version1 and Version2. The gold circles represent the center of the clusters for Version1 (upper right) and Version2 (bottom left). 