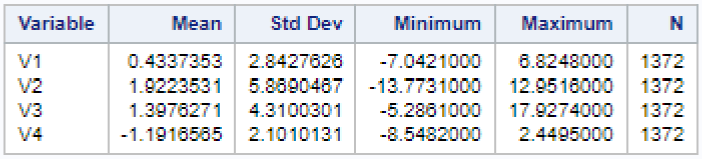
Introduction

Counterfeiting banknotes has been a problem since the introduction of color photocopiers and computer image scanners. The bank industry has suffered from counterfeits due to inflation and reduction in the value of real money. Here we have been given a dataset that contains different entries for different bills that have been analyzed. There are four different variables (variance of wavelet, skewness of wavelet, kurtosis of wavelet, and entropy) within the dataset, and a fifth Boolean variab le stating whether or not the banknote is real of counterfeit. With this dataset we have been tasked with creating a model to predict counterfeit banknotes using SAS we can perform clustering, and potentially create a binary regression model that will predict future banknotes authenticity.

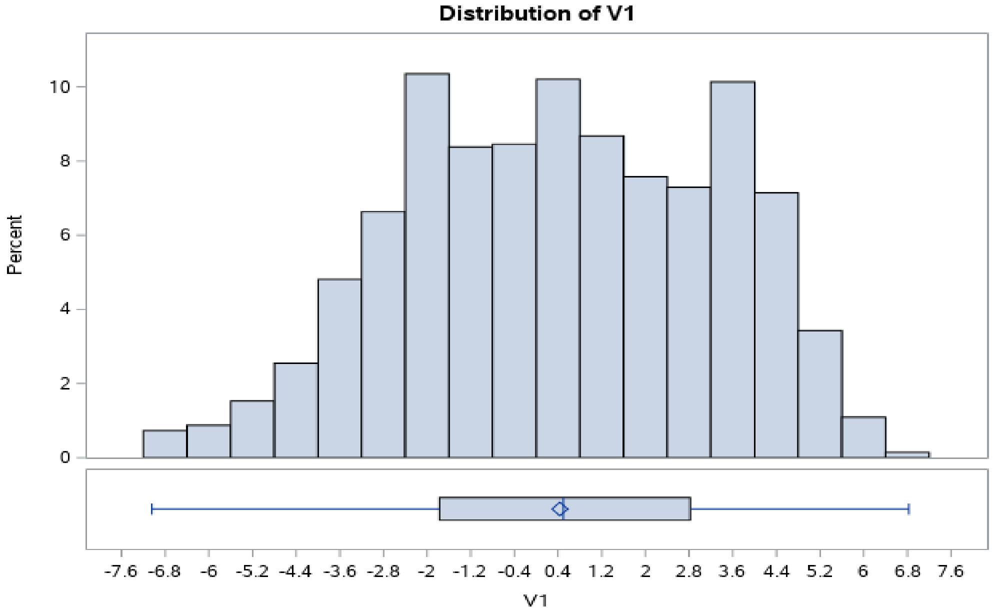
Part 1 - Summary Statistics

Here are the summary statistics for all the variables within the Banknote CSV file:



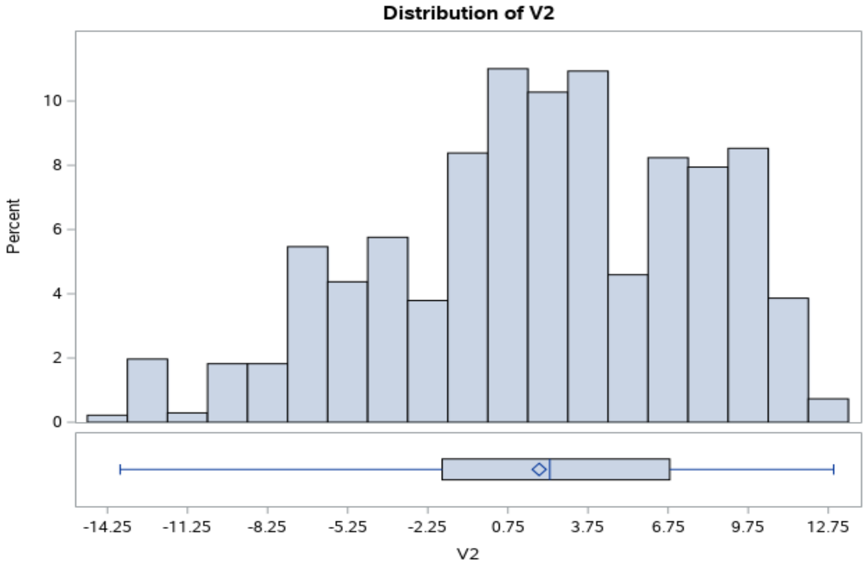
Variable	Mean	Std Dev	Minimum	Maximum	N
V1	0.4337353	2.8427626	-7.0421000	6.8248000	1372
V2	1.9223531	5.8690467	-13.7731000	12.9516000	1372
V3	1.3976271	4.3100301	-5.2861000	17.9274000	1372
V4	-1.1916565	2.1010131	-8.5482000	2.4495000	1372

Here is the graphical summary of the variable V1:



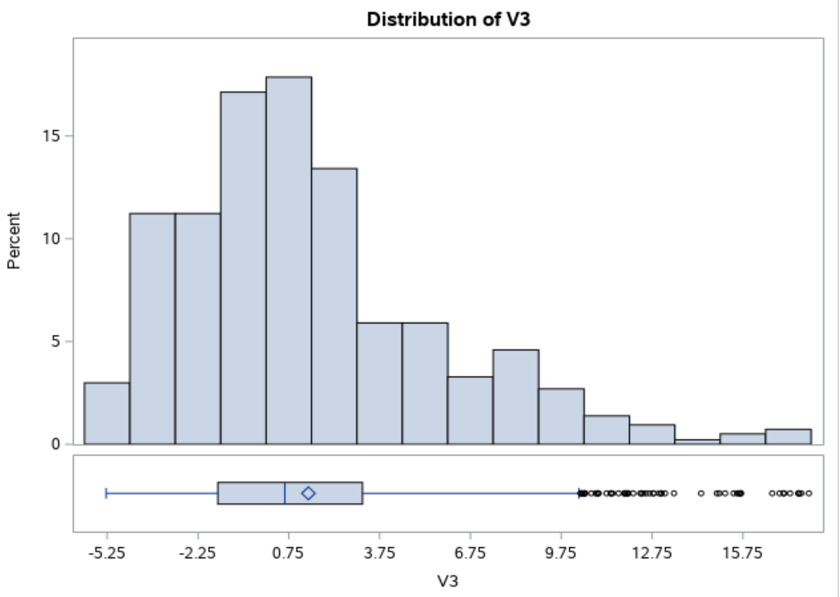
Here we can see V1 has a symmetric Gaussian distribution, with no skew in either direction.

Here is the graphical summary of the variable V2:



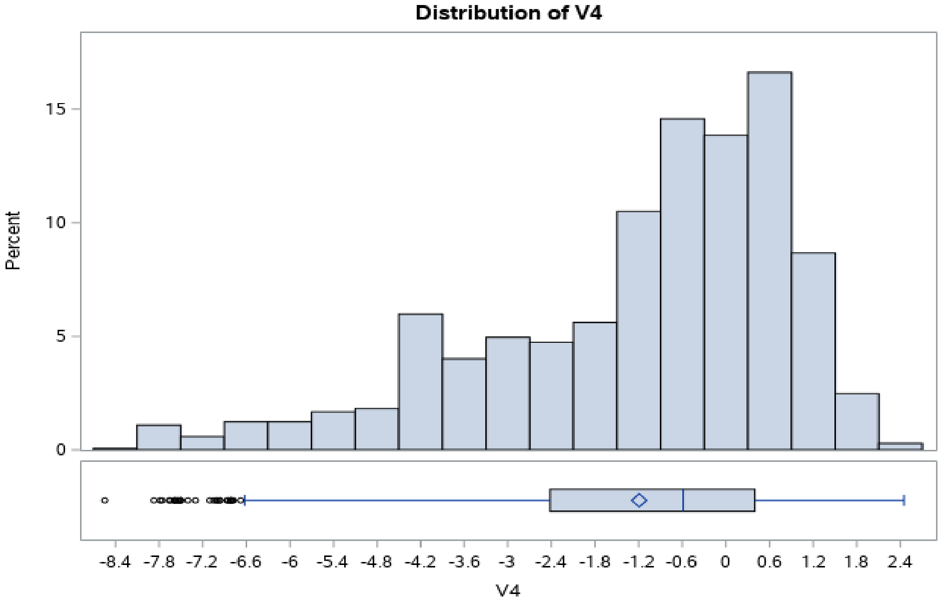
Here we can again see V2 has a normal symmetrical distribution, with a slight skew to the left.

Here is the graphical summary of the variable V3:



Here we can see that V3 has a normal distribution, but is quite skewed to the right.

Here is the graphical summary of the variable V4:



Here we can see V4 has again a normal distribution, with a skew to the left.

First, we can look at the different variables are all titled V1-4, this is for the sake of giving the variables a shorter name, an arbitrary value that has been assigned to each. So that there are zero questions, here are all of the continuous variables with their new names (for the rest of the paper they will be referred to V1, V2, V3, and V4):

- Variance of Wavelet Transformed image = V1
- Skewness of Wavelet Transformed image = V2
- Kurtosis of Wavelet Transformed image = V3
- Entropy of image = V4

There is also another binary variable that has not yet been shown here, that shows whether or not a banknote was forged. This variable is called Class, and will be shown later on in this report.

Next, we can look at the initial table that is filled with the summary statistics to start diving into what exactly this dataset contains. First, we can see the total number of observations in each of the columns, since there are equal amounts for N in each of these rows, this is a pretty good indicator that there are no missing values within the dataset. Then we can start looking at the other statistics found from the variables and the first thing that really jumps out at me is all of the means for the variables are within the range of –2 to 2. The second thing that is noticeable from this information is that V2 has the standard deviation, and therefore the highest variance out of all the variables. This is also apparent in when looking at the minimum and maximum columns as they contain values that are further away from the mean on either side, than any other variable. This doesn't mean much to us yet, but the point here is to look at the statistics and try to get familiar with the data.

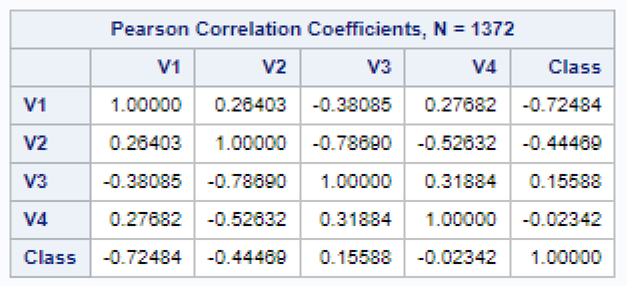
Then we can start analyzing the histograms and boxplots for each of the variables. When we look at these, we can see that V1 and V2 both have normal distributions that appear to taper off, fairly evenly on both sides. V2 does have a slight skew to the left, but not far enough that there seems to be any outliers. V3 and V4 on the other hand, both skew in opposing directions and do seem to contain outliers that can be seen with the use of the boxplot. According to the Engineering Statistics Handbook (2012), This skew is generally caused by the possibility of a bound being on the opposing side of the skew (NIST, 2010). This doesn't appear to be the case in this situation as there doesn't seem to be any apparent bounds in any of these graphs.

Part 2 - Potential Anomalies

According to McLeod (2019), one of the best ways to analyze the data for any possible outliers or anomalies, is to simply look at histograms and boxplots. When originally looking at histograms, this may be a little difficult to see potential anomalies, but when the boxplot is present, this becomes very apparent. The middle box on the graph shows the inner quartile range, which show s the inner 50% of the data points. The whiskers then extend from this inner range and extend to the outer ranges of the data. These whiskers do exclude data points from the whiskers that are seen as outliers. As seen in part 1, there are already some outliers that are anomalies that have been noticed due to the points that lay outside the whiskers of the box plots. It is important to note that these outliers are found on the boxplots from the variables that have a more serious skew to their histograms.

Part 3 - Associations

To find the associations within SAS, it is easiest to look directly at the table of associations. This table gives you the Pearson Correlation Coefficient, that tells how well variables are correlated with other variables. Here is the table that Sas produced with the banknote dataset:



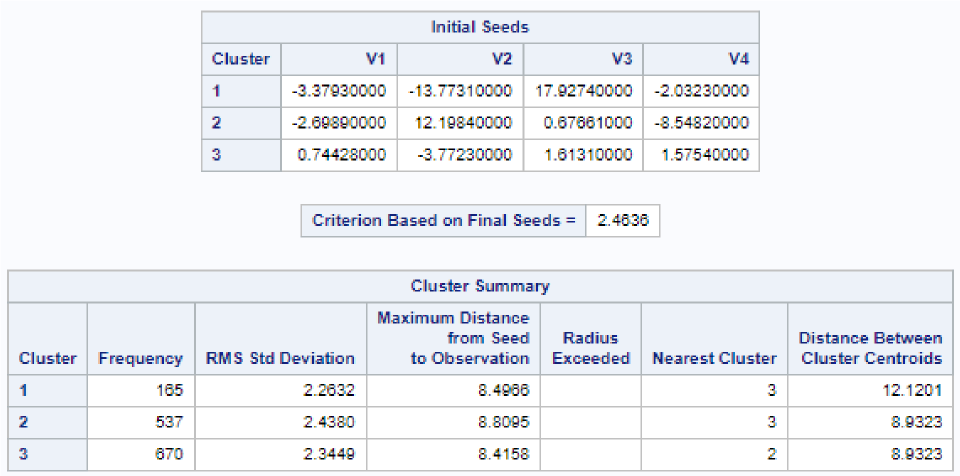
Pearson Correlation Coefficients, N = 1372

	V1	V2	V3	V4	Class
V1	1.00000	0.26403	-0.38085	0.27682	-0.72484
V2	0.26403	1.00000	-0.78690	-0.52632	-0.44469
V3	-0.38085	-0.78690	1.00000	0.31884	0.15588
V4	0.27682	-0.52632	0.31884	1.00000	-0.02342
Class	-0.72484	-0.44469	0.15588	-0.02342	1.00000

Here we are looking for values that are close to 1 or –1, signifying that we have variables that do show signs of potential correlation. The Pearson coefficient is basically putting the two variables on graph, and drawing a straight line through the graph and calculating the distance the points are from that graph. If there is either a 1 or –1 there is a perfect linear correlation, while if there is 0 there is no correlation (Laerd Statistics, 2020). Since we are trying to predict forgery, the variable Class is showing if an item was found to be forged. This is the variable that we are making a model to predict, therefor this is the variable we will be trying to find correlations with. For this we can look at the bottom row or the right-hand column and take the absolute value of all the values in either the row or column and find the values that are close to 1. We can see that V1 and V2 have the highest correlation with the Class variable, while it appears that V4 has the smallest correlation. This does not mean that they will not be used in the model creation, just shows that they are less promising to be significant factors for predicting forgery.

Part 4 - Clustering

To make different clusters within the data, we have used FASTCLUS. FASTCLUS uses the k-means model, or the Euclidean distance, which bases the cluster centers on least squares estimation (SAS Documentation, 2011). These cluster seeds will then grow into the clusters which are seen below. Here are the results provided by SAS from running the FASTCLUS Procedure as well as the graphical representation:



Initial Seeds

Cluster	V1	V2	V3	V4
1	-3.37930000	-13.77310000	17.92740000	-2.03230000
2	-2.69890000	12.19840000	0.67661000	-8.54820000
3	0.74428000	-3.77230000	1.61310000	1.57540000

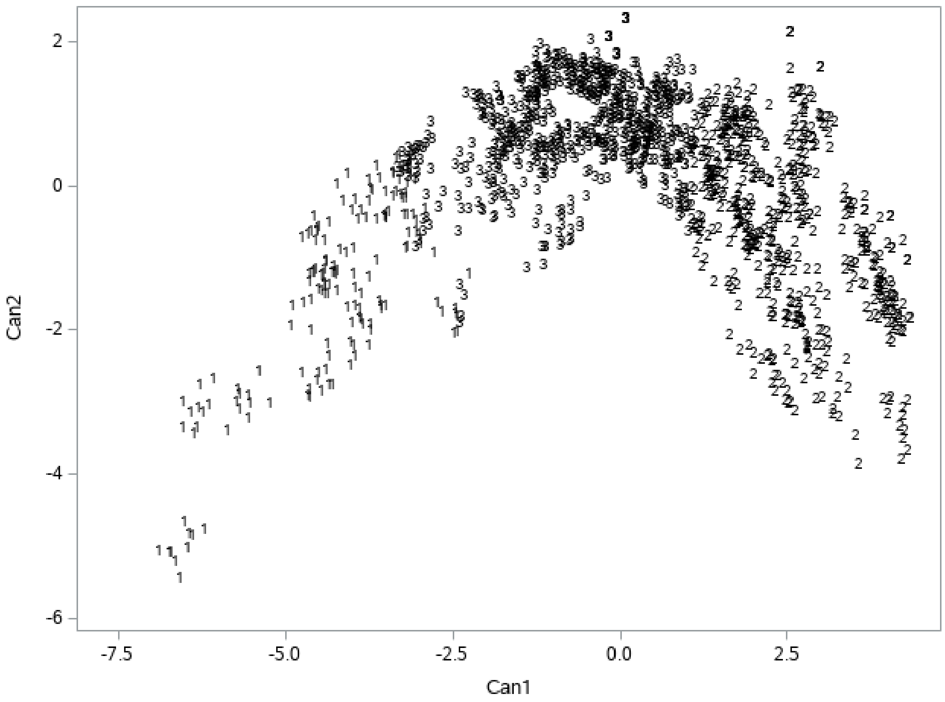
Criterion Based on Final Seeds =	2.4636

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	165	2.2632	8.4966		3	12.1201
2	537	2.4380	8.8095		3	8.9323
3	670	2.3449	8.4158		2	8.9323

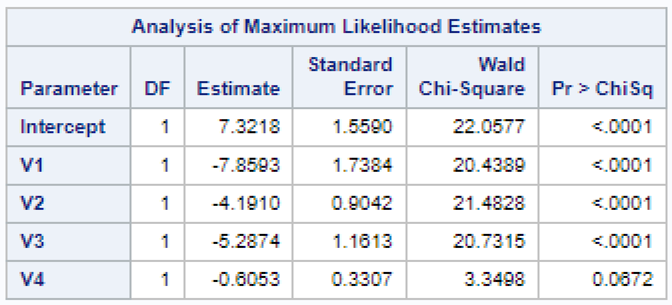
The first table shows the initial seeds for the clusters, and these are showing the exact points for each of the variables that the cluster seed contained. The second table shows the valuable information from the clustering procedure. As we can see here, there are three different clusters that have been made, with the third cluster containing the most data points, and the first containing the least. Then we can also see the standard deviation or spread for each of the clusters, with the third cluster containing the least amount of spread, and the first and second containing more and more spread respectively.

Here is the graphical representation of the clustering procedure, and the different clusters can be seen by the different numbers that have been assigned to each of the data points:



Part 5 - Classification

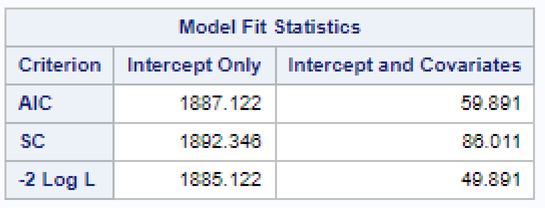
Here we are using a binary regression model to predict whether or not a banknote has been forged. Here are the results from the model that was built in SAS:



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	7.3218	1.5590	22.0577	<.0001
V1	1	-7.8593	1.7384	20.4389	<.0001
V2	1	-4.1910	0.9042	21.4828	<.0001
V3	1	-5.2874	1.1613	20.7315	<.0001
V4	1	-0.6053	0.3307	3.3498	0.0672

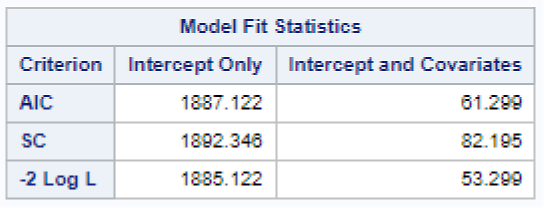
This is the point where we can see which variables are significant factors of predicting forgery. Since our confidence level is not specifically set, we can assume that a 0.05 level is set, and compare the values that are presented in the Pr > ChiSq column to the confidence level of 0.05 (Tan et al., 2020). All of the points that are below this value, are significant factors in predicting forgery, meaning that V1, V2, and V3 are significant. V4 on the other hand is not a significant predictor.

Now that we have a model that has been created it is important to validate the model and look for improvements. The easiest way to validate the model is to check the model fitment table. Here is the table provided from SAS:



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1887.122	59.891
SC	1892.346	86.011
-2 Log L	1885.122	49.891

Here we can see the AIC level for this model. According to Maydeu-Olivares, C. García-Forero (2010), the AIC model is not good for hypothesis testing, but rather a good check for fitment. Thus, meaning the lower the AIC level is, the better the fitment of the model is. Since we saw previously that V4 was not a significant predictor of forgery, we can remake the model excluding this variable to see if we can get a better score here. Here is the new model fit statistics table, with V4 excluded:



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1887.122	61.299
SC	1892.346	82.195
-2 Log L	1885.122	53.299

Since the AIC value has risen, it is safe to say that the first model did a better job at predicting forgery. Another possibility to make this model more effective would be to find other variables that have data that can be used to predict forgery. The more variables that are used in the model to predict an outcome, that also have significant predicting abilities will always make the model more accurate. Since we don't have any other variables to add in this case, it is safe to say that we have made the best binary logistic regression that is possible with the given data.

References

Laerd Statistics. (2020). Pearson Product-Moment Correlation. Retrieved September 8, 2021, from https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php.

Maydeu-Olivares, A. García-Forero, C. (2010). Goodness of Fit Testing. In C. García-Forero (Ed.), International Encyclopedia of Education (Vol. 3rd, pp. 190–196). University of Barcelona.

McLeod, S. A. (2019, July 19). What does a box plot tell you? Simply psychology: https://www.simplypsychology.org/boxplots.html

NIST, US Department of Commerce (2012). Histogram Interpretation: Skewed (Non-Normal) Right. Engineering Statistics Handbook. Retrieved September 10, 2021, from https://www.itl.nist.gov/div898/handbook/eda/section3/eda33e6.htm.

SAS Documentation. (2011, July 15). The FASTCLUS Procedure. SAS/STAT(R) 9.3 User's Guide. Retrieved September 11, 2021, from https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_fastclus_sect001.htm.

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2020). Introduction to data mining. Pearson.