|  |
| --- |
| Team 4: Aparna Unni Caitlyn May Kathryn Walker Partha Ganrai Pooja Singh  10-24-2016 |

|  |
| --- |
| University of Connecticut |
| **Loan Statistics Data Preparation** |

Contents

[Executive Summary 2](#_Toc465089473)

[Strategy 3](#_Toc465089474)

[Data Dictionary 3](#_Toc465089475)

[Sample 4](#_Toc465089476)

[Exploration & Modification 5](#_Toc465089477)

[Change Data Types 5](#_Toc465089478)

[Column Headers 5](#_Toc465089479)

[Correlation 6](#_Toc465089480)

[Funded Amount and Loan Amount 6](#_Toc465089481)

[Total Principal from Investor 7](#_Toc465089482)

[Interest Rate(%) and Grade 8](#_Toc465089483)

[Column Removal 8](#_Toc465089484)

[Data Modification 9](#_Toc465089485)

[Binning Annual Income 10](#_Toc465089486)

[Data Visualization 11](#_Toc465089487)

[Missing Value Treatment 12](#_Toc465089488)

[Outlier Analysis 15](#_Toc465089489)

[Univariate 15](#_Toc465089490)

[Bivariate 17](#_Toc465089491)

[Multivariate 18](#_Toc465089492)

[Scaled Variable 19](#_Toc465089493)

[Principle Component Analysis 20](#_Toc465089494)

[Conclusion & Recommended Next Steps 22](#_Toc465089495)

[Appendix A: Data Dictionary 23](#_Toc465089496)

[Appendix B: Data Modification 24](#_Toc465089497)

[Appendix C: Data Visualizations 28](#_Toc465089498)

[Appendix D: Outlier Analysis 30](#_Toc465089499)

**“The work presented in this report is our team’s work and our team’s work alone”**

# Executive Summary

To prepare data for a predictive modeling exercise, the team has compiled a sample size of 30,000 records that is ready for modeling. Data and the data dictionary were received from a reliable source, and data preparation was carefully done as per basic pre-processing standards.

This report walks through the various processes the team carried out in SAS JMP Pro to explore the data, treat the missing values and to analyze the outliers. The report will also mention steps the team took to reduce dimensionality to decrease model complexity.

Two datasets have been provided to the modeler, or the person that utilizes the model ready data. Our primary dataset, which is recommended for use to create a predictive model, includes normalized variables and principal component values. The secondary reference dataset includes non-scaled variables which can be used as a reference if modeling requires original data.

# Strategy

After extensive analysis on the data dictionary and provided data, the team decided that Loan Status would be our target variable. Target variable is the one which receives properties of other variables. Hence, the status of the loan will be predicted based on other variables. The team categorized Loan Status into three categories: Current, Fully Paid & Default. The team believes that predicting the loan status will have great business value of predicting whether a Current loan will turn into a Fully Paid loan or a Default loan. In addition, in the future before a loan is issued, the loan advisor could have the knowledge of how this loan will benefit their business or could possibly be bad for business.

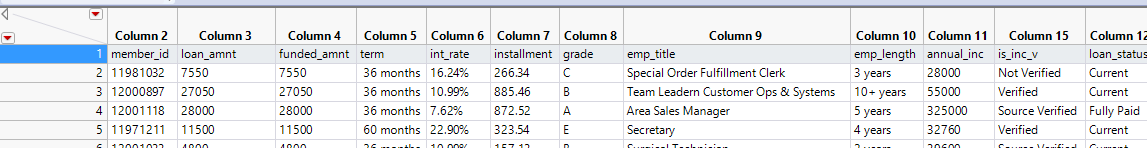
The team followed the SEMMA approach in data mining to pre-process the given dataset. First, the team selected a random sample from the given data to make inferences about the whole population. Second, the team iteratively explored and modified data to create a model ready data set. Recommended modeling next steps were discussed during the exploring and modification stages of the analysis and are documented in the conclusion section of this report.

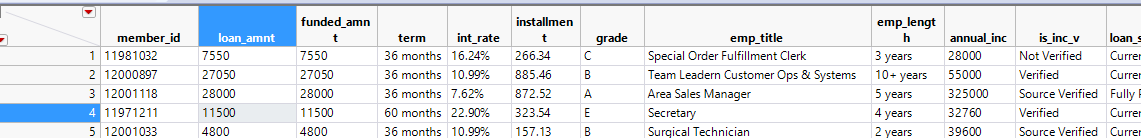
# Data Dictionary

A finalized Data Dictionary is included as a part of the model ready data. The user is recommended to read the data dictionary for clean direction of what each variable is before taking further modeling action. The data dictionary is attached with this report, and can also be seen in [Appendix A](#_Appendix_A:_Data).

# Sample

Initially, the variable names or column headers of the population were displayed as row 1 of the data. This data was imported again from MS Excel and in the import wizard, it was indicated that row 1 will be the column headers. Now, the dataset has appropriate and informational headers.



Exhibit 1: Data set before and after appropriate column headers

After importing the data, there were numerous completely blank rows at the end of the dataset. These rows were deleted since there was no value. Exhibit 2 shows the process in JMP to select randomly. After indicating the sample needed 30,000 random values, the selection was inverted and those rows were deleted. The remaining rows are the sample of 30,000 randomly selected rows for data cleansing and preparation for modeling.

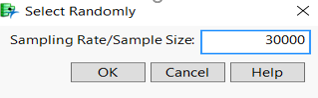
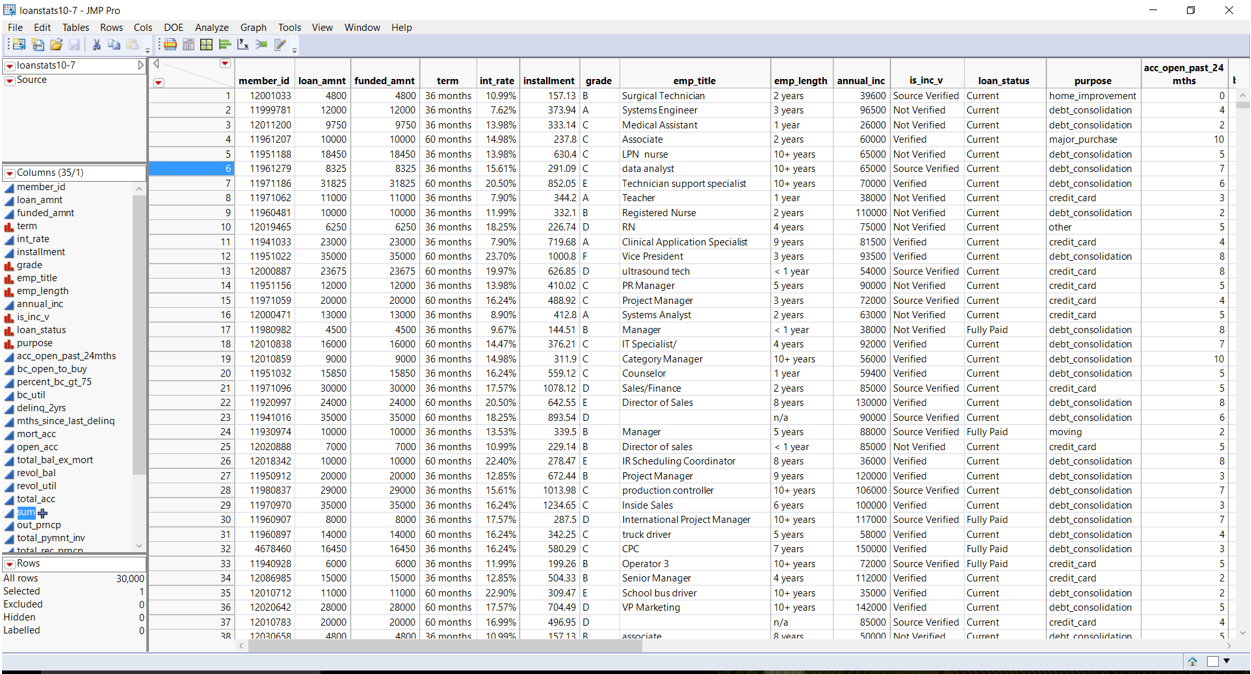


Exhibit 2: Process of selecting a random sample in JMP

Exhibit 3 shows the initial part of the selected sample of 30,000 observations.

Exhibit 3: Initial part of the data observations

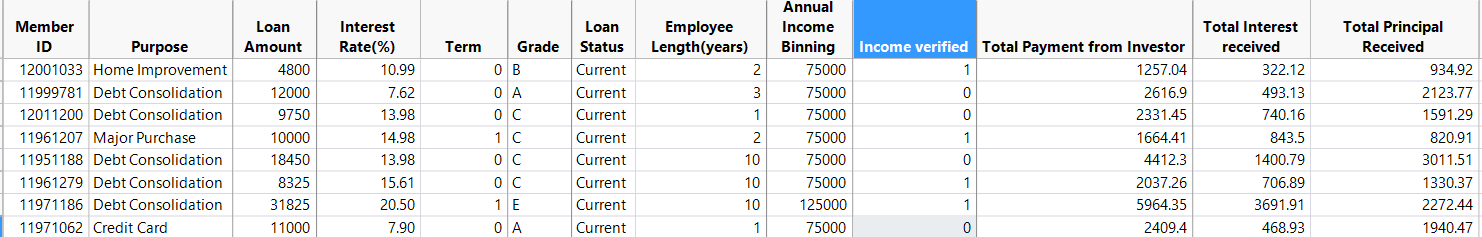
# Exploration & Modification

After choosing a sample size for data cleaning, the next step of the process is to Explore and Modify. While exploring, the team found relationships, noticed trends and gained an understanding of the data. Iteratively while the data was explored, the data was connected to the business value it could have. During exploration, the team started to make modifications to the data. By exploring and modifying in interactions, the team learned more about the data with each modification which led to a model ready data set. The below processes define the team’s exploration and modification of the data set to compile the primary modeling data set.

## Column Headers

The given data sample had column headers with abbreviations and underscores, most likely from the original database source of the data. Since the data will be used for modeling and visualization, the team decided to change the column names to be understandable and visually appealing. The team removed underscores and made the title proper case.

See Exhibit 4 for a sample of what the model ready column headers look like. The value is any consumer of the data will understand what each column is most likely without referring to the data dictionary.

Exhibit 4: Data sample with appropriate column headers

## ****Change Data Types****

The data imported into JMP was realized as nominal, or categorical data. The team explored each column’s data by either scrolling through, reading the data dictionary, or by viewing the recode option. Each variable was changed to a continuous, nominal or ordinal variable depending on what was suitable for the data type. See Exhibit 5 for the before and after view of the data type change modification. The value of having an accurate data type is when visualizing, modeling and interpreting the data, the correct data type will ensure the data has the accurate statistic meaning in calculations.

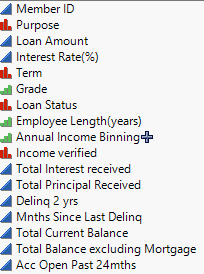
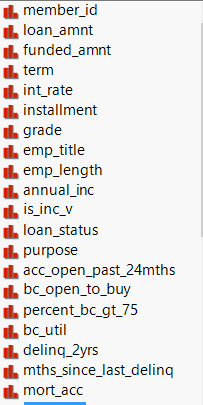


Exhibit 5: Before and after changing data types

## Correlation

Correlation between columns in a data set is an explorative method to standardize the measure of dependence. If two columns have a correlation close to 1 or -1 , these columns also have a high dependence on each other. The team took appropriate action on the columns depending on the correlation of each column.

### *Funded Amount and Loan Amount*

The team plotted a scatterplot matrix between *Funded Amount* and *Loan Amount.* The scatter plot informed the team the correlation was close to one, inferring high correlation. To remove redundant information from the dataset, the team decided only one variable is needed and a model would only need one of these data attributes. See Exhibit 6 where the correlation is indicated as .9999 between the two attributes. Funded amount was kept in the dataset since it would be used for further calculations of columns such as installments and total principal received. As per the data dictionary, funded amount is the final loan amount granted by the loan company, therefore it can be used for all calculations.

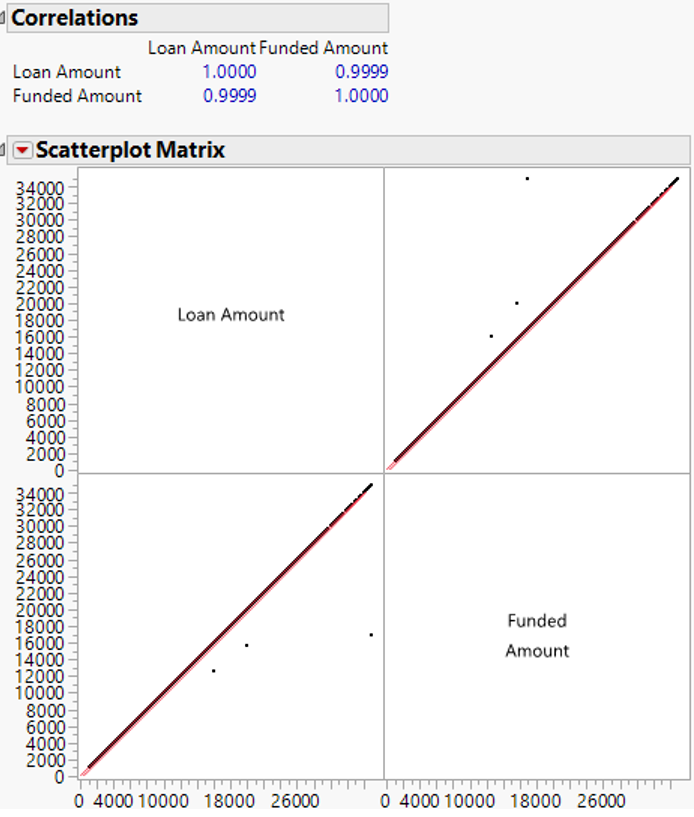


Exhibit 6: Correlation between Funded amount and Loan amount

### *Total Principal from Investor*

A similar analysis was completed for *Total Principal from Investor* as was done for *Loan Amount* and *Funded Amount*. As seen in Exhibit 7, the correlation between the field, and the sum of Principal Received and Interest Received is .9994. With this high of a correlation, the team assumed that Total Payment from Investor could be derived from *Principal Received* and *Interest Received*. With this, Total Principal from Investor can be removed from the dataset since there is no need to store redundant data.

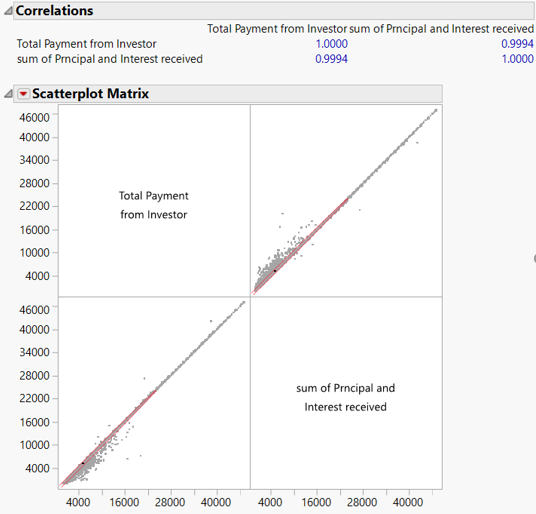


Exhibit 7: Correlation between Total payment from investor and sum of principal and interest received.

### *Interest Rate(%) and Grade*

By plotting a Oneway analysis of *Interest Rate (%)* by *Grade*, the team saw that *Grade* was determined by grouping a range of interest groups together. Exhibit 8 shows that Grade A has a range of 6%-10%, Grade B has a range of 10%-14%, and so on for the other Grades. The team kept this column in our data set since we found it could be a useful way to group interest rates. *Interest Rate* was also kept in the data set since it provides information about the loan and could be used to calculate other valuable information. This example shows that though the two variables are related, they have valuable business information for modeling and can be used in different ways.

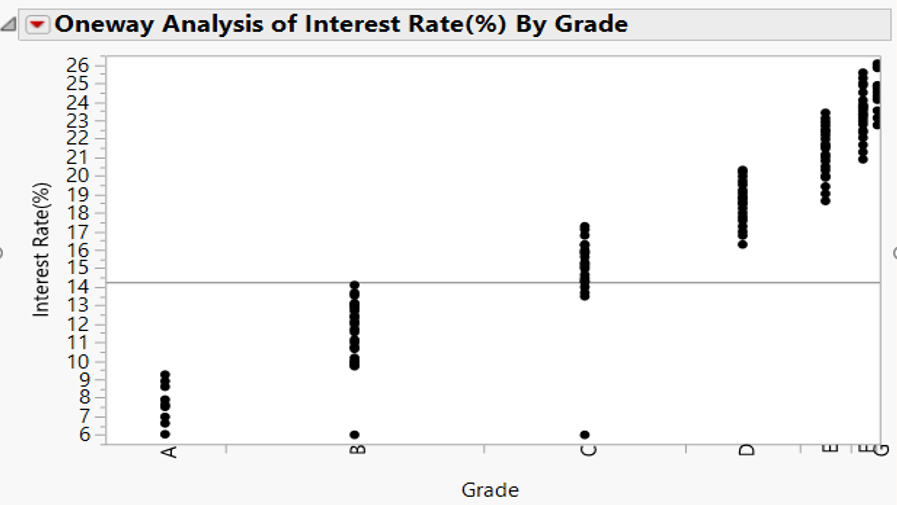


Exhibit 8: Oneway Analysis of Interest rate and Grade

## Column Removal

While exploring the data, the team saw little value from a few columns that lead to the decision to remove them from the data set. The reason for each column that was removed is listed in Table 1 below. The team focused on removing columns with little business value because having too many columns in a data set can make the data overwhelming and unnecessarily large. There were also some columns that could be derived if they were needed in the model.

The team focused on creating a concise informative data set without providing data overload to the end user (curse of dimensionality), therefore removed columns as needed. If the team could not find a valid reason for removing a column, the column was kept in the data set. If no correlation was found and there was valid data available, the column was kept in the dataset.

|  |  |
| --- | --- |
| **Column** | **Removal Reason** |
| Loan Amount | Funded Amount and Loan Amount were correlated. |
| Installment | The installment can be calculated using term, interest rate and funded amount. |
| Employee Title | This column was removed because all employment titles range drastically throughout different industries and companies. Having this data with such inconsistencies does not allow a clear analysis of the data especially for modelling purposes. |
| Months Since Recent BC Delinquencies | This column was removed because 80% of the values were missing and the data did not provide adequate information to keep and analyze. |
| Outstanding Principal | This column was removed because the value can be determined by the difference of Funded Amount and Total Principal Received. |
| Total Payment Investor | This column was removed because the value can be determined by the sum of Total Received Principal and Total Received Interest. |
| Last Payment Amount | This column was removed because the value of how much the customer owes per month can be determined by the term, funded amount and interest rate. Also, the column was assumed to have limited insights in a predictive model as it only specifies the most recent payment. |
| Policy Code | This column was removed because there was no variance in the data. 29,999 rows of Policy Code had the same value of 1. (See [Appendix B.6](#_Appendix_B:_Data) for Policy Code recode) |

Table 1: Reasons for removal of columns

## Data Modification

After removing columns that were not necessary for the dataset and deciding our target variable, the team explored the remaining columns to develop ways to make the data more consistent. For each column remaining, the team made appropriate decisions to modify the data and simplify the dataset. In Table 2 below, see the decision for each column. The team used the recoding process to implement each modification decision. Screenshots and examples of how the team modified the data can be seen in [Appendix B](#_Appendix_B:_Data). The value of going through this exercise is to resolve inconsistencies and create a simple and clean dataset.

|  |  |
| --- | --- |
| **Column** | **Data Modification Decision** |
| *Term* | For simpler and clearer understanding of the categories, the team replaced 36 months with the value 0 (zero) and 60 months with the value 1 (one). |
| *Interest Rate* | The percentage symbol (%) was removed from rows and added to the column title for more clarity. |
| *Employee Length* | As per the definition given in the initial data dictionary, employment length <1 year was replaced with the value 0 (zero) and employment length >10 years with the value 10 (ten). The word ‘years’ was removed from the rows and added to the column title. |
| *Is Income Verified* | Verified and source verified incomes were added to one grouping of the value 1 (one) and not verified with the value 0 (zero).  This was done for a simpler and clearer understanding of the categories. |
| *Loan Status* | Loan status has been streamlined into three categories. These categories are Current, Fully Paid and Default in the data. The breakdown of these columns is that current denotation includes current accounts and those accounts in grace period, Fully Paid denotation includes fully paid accounts, and Default indicates bad repayment behavior like are late, defaulted, and charged off. |
| *Purpose* | Underscores were removed and proper title case was used for the data. |

Table 2: Data modification decisions ([Appendix B](#_Appendix_B:_Data))

## Binning Annual Income

Binning refers to dividing continuous variables into groups. This is done to discover a set of patterns in continuous variables, which is otherwise difficult to analyze. After binning is complete, the column is easier to analyze. In the column titled *Annual Income* the practice of binning was completed to better analyze and interpret the data. The team chose to do the practice of binning because it provides a clear division of continuous data into specific groups (or bins) through the data set.

As seen in Exhibit 9, the team has binned the annual income values from $0 to $250,000 into five bins each increasing by $50,000 per bin. Within each of the bins, the team has assigned the mid-value to represent the corresponding bin for the ease of data analysis purposes. To not lose information, the team has created smaller bins to produce better results in a future predictive model.

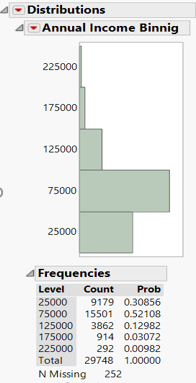
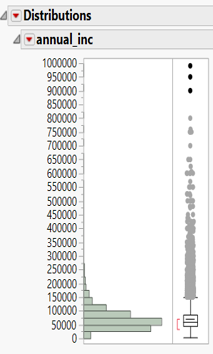


Exhibit 9: Before and after annual income binning

## Data Visualization

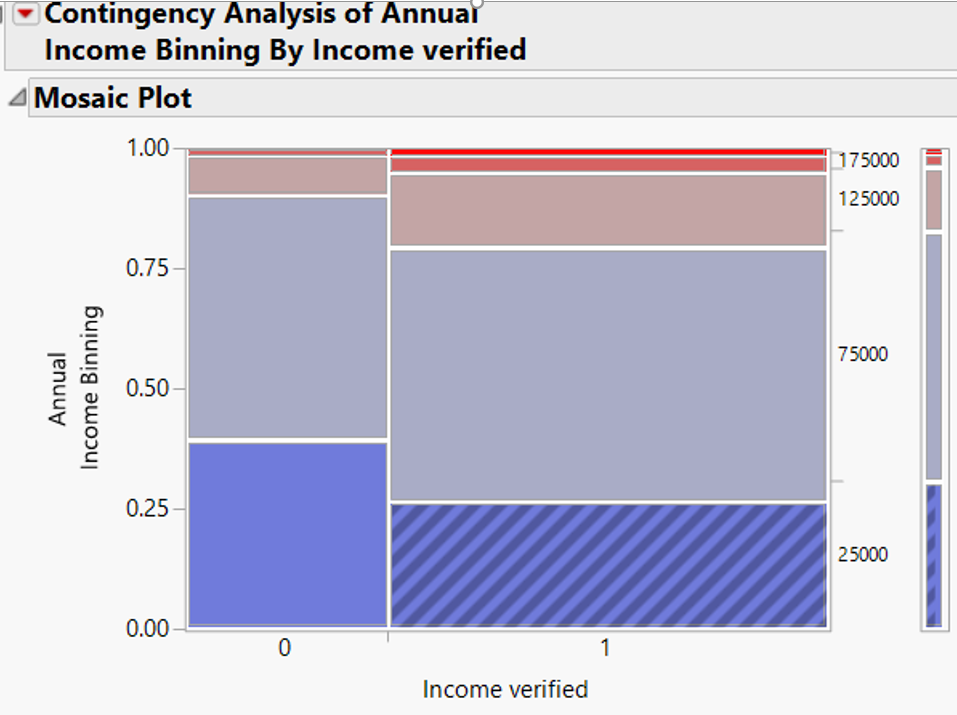
With the remaining columns that have been modified, the team explored the spread of data in the sample with visualizations. An example is in Exhibit 10 where the team explored the relation between *Annual Income Binning* and *Income Verified*. Through some of the exploratory methods, the team decided next steps of dealing with missing information, outlier analysis and other methods to reduce the number of variables in the dataset. The value of taking the step to visualize the data was to notice trends, easily see patterns and understand other pre-processing steps that need to occur. Other data visualization examples that were explored can be found in [Appendix C](#_Appendix_C:_Data).

Exhibit 10: Mosaic Plot of binned annual income and income verified.

## Missing Value Treatment

The given data sample contained several columns with missing values which needed to be reviewed for the data to truly be a reliable source for future modelling. Within this data sample, there were several examples of modifications that were made with caution to handle the missing values. The specific techniques used for handling the missing values included: replacing values with mode, replacing value with median, replacing values with prominent outcome and imputing missing values by least squares imputation. Exhibit 11 displays the missing value pattern for all the variables in the sample.

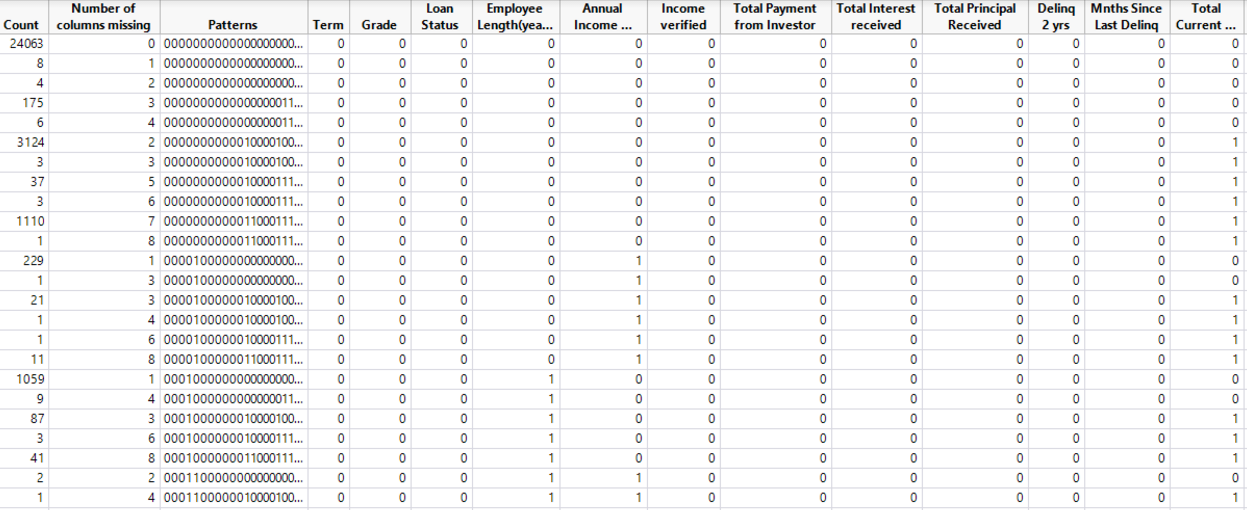


Exhibit 11: Missing values pattern for the sample

In certain variables, there are very few numbers of missing variables and hence we have used mode or median for imputation of the missing values as doing this was not affecting the standard deviation of the variable. In others, where there are large number of missing values, we have used multivariate normal imputation method to impute them. Below explanations of each specific treatment can be seen to gain a greater understanding of why the steps were taken.

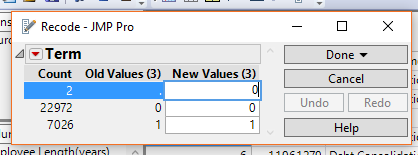


Exhibit 12: Missing value treatment for *Term*

In this first example, as shown in Exhibit 12, the team chose to replace the values ‘. ‘with the mode of the term length. The term length is distinguished by a zero (for 36-months) or a one (for 60-months). For the two values that were missing, 0 was added in place of ‘.’ as 77% of the values in term length are 0. The missing values were replaced with mode as the standard deviation, mean and median of the data will not be affected with this replacement.

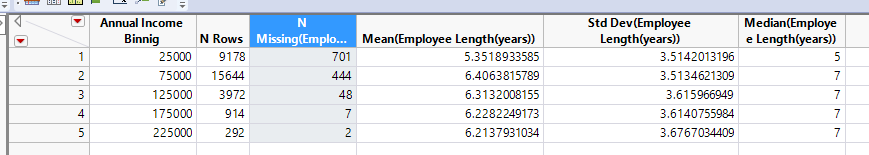


Exhibit13: Before missing value treatment for *Employee Length*

In the second example, as shown in Exhibit 13, the team chose to replace the missing values in *Employment Length* with the median of the *Employee Length* for the existing data. A summary table was derived which included the mean and median of *Employee Length* for each of the *Annual Income* bins.

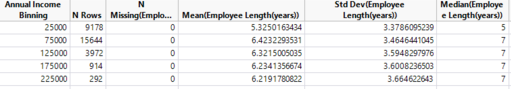


Exhibit 14: After missing value treatment for *Employee Length*

When replaced with median, extremes will not be altered and the mean will be preserved. This can be seen in Exhibit 14.

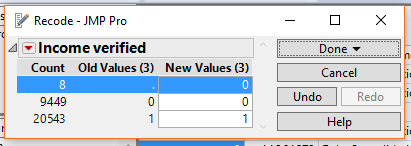


Exhibit 15: Missing value treatment for *Income Verified*

In the third example, as shown by Exhibit 15, the team chose to replace the missing values in *Income Verified*, with the value of zero (or not verified annual income amount). The decision-making process behind this step was due to the higher cost of error that could be associated with replacing the missing values with Verified or 1 because it would indicate a verified source of annual income amount.

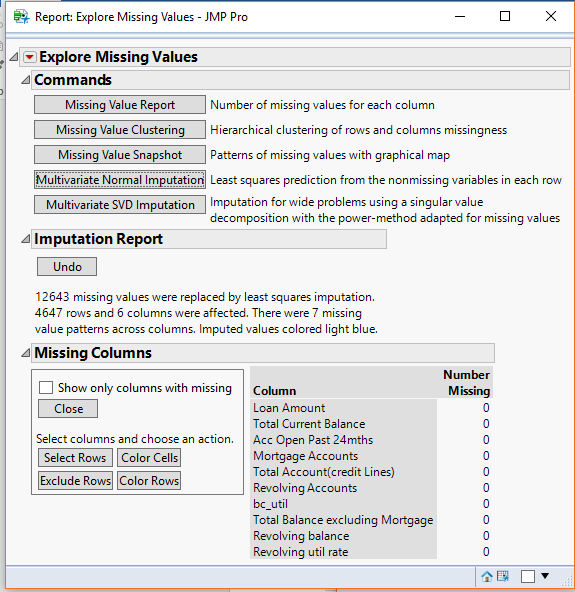
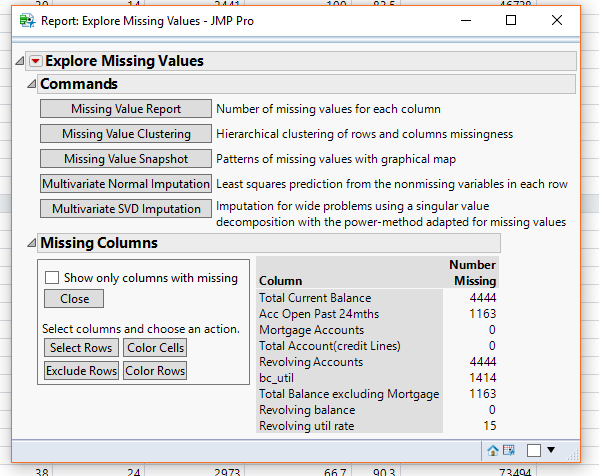


Exhibit 16: Missing value treatment for *Total Current Balance, Acc Open past 24 months, Revolving Accounts, BC Util, Total Balance excluding mortgage, Revolving Util Rate*

In the final example, as shown by Exhibit 16, the team chose to replace the missing values in columns titled *Total Current Balance, Acc open last 24 months, Revolving Accounts, BC Util, Total Balance Excluding Mortgage* and *Revolving Util Rate* using the Multivariate Normal Imputation utility.

This method uses least squares imputation algorithm which imputes the missing value of a variable by forming a linear relationship between the variable and the other similar variables. We have used this method because it is more robust than the mean-median-mode imputation method and because it does not affect the standard deviation of a variable.

## Outlier Analysis

Outlier analysis within the data was a large portion of the exploration and modification process of the team's data set. To truly understand the outliers within the data set several outlier detection methods were used to visually see the values that did not conform to the expected pattern of the column. As the team, does not know the entirety of the business scenario, the decision was made that all the outlier variables should remain included due to their potential importance in the data set.

### *Univariate*

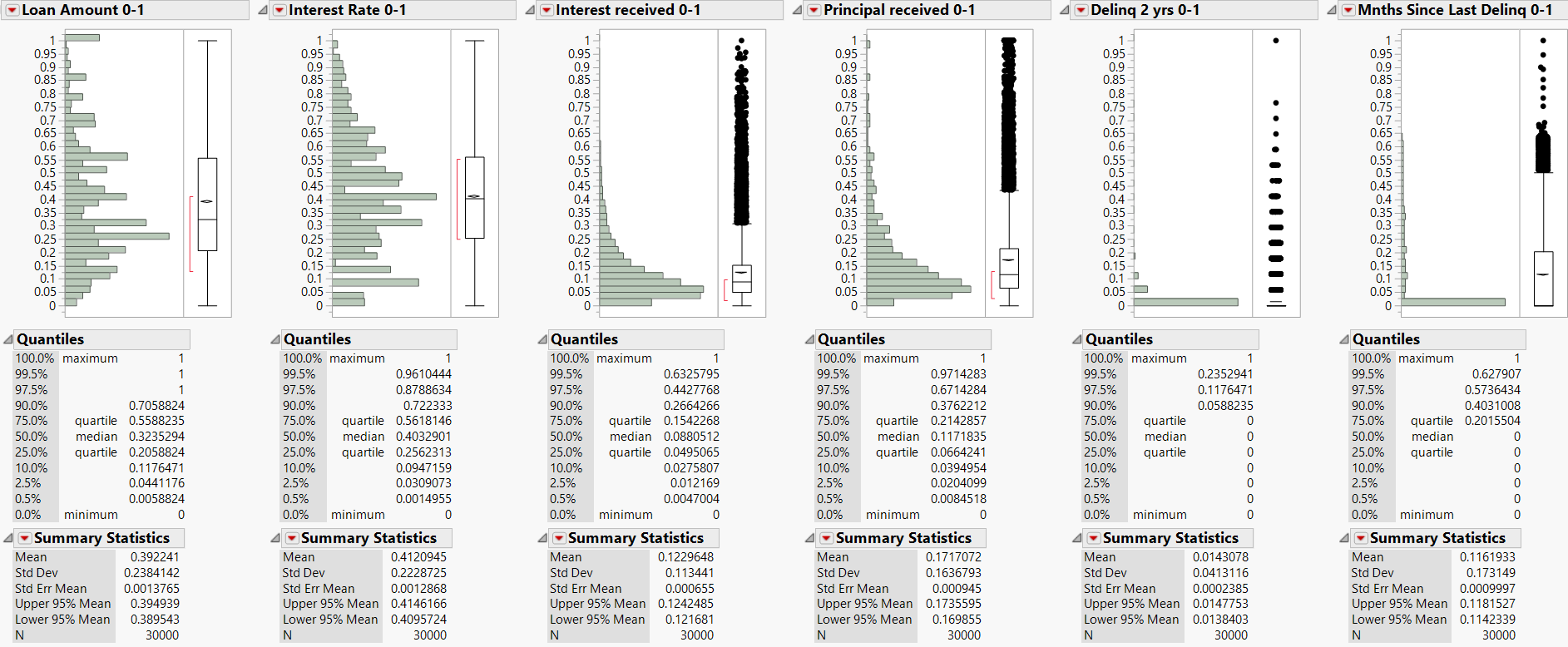
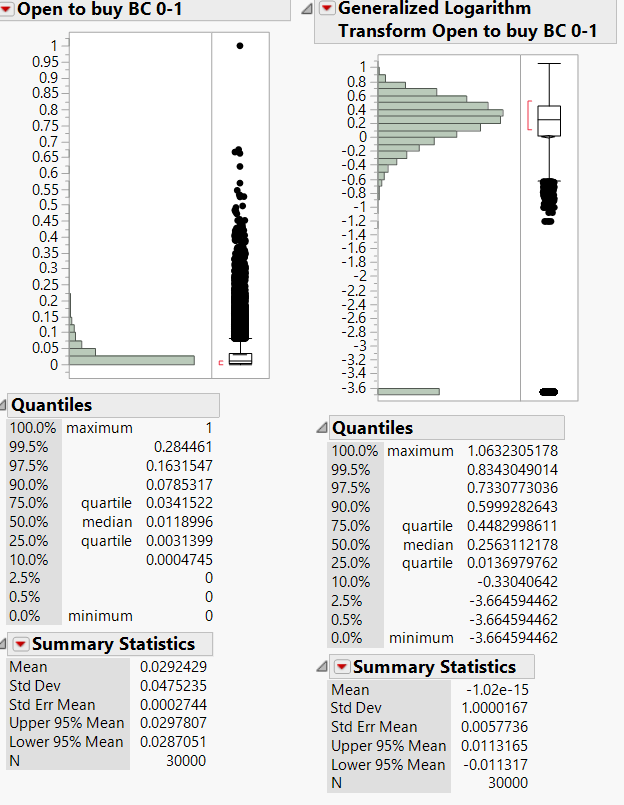
The team first completed its univariate outlier detection by using distribution graphs; this gave insight into the data and how the variables placed for each columns distribution. Exhibit 17 provides an overview of the distribution graphs for a sampling of the variables which shows the significant amount of values that fell outside the whiskers of the box plot. The entirety of the distribution graphs for all variables can be seen in the [Appendix D](#_Appendix_D:_Outlier).

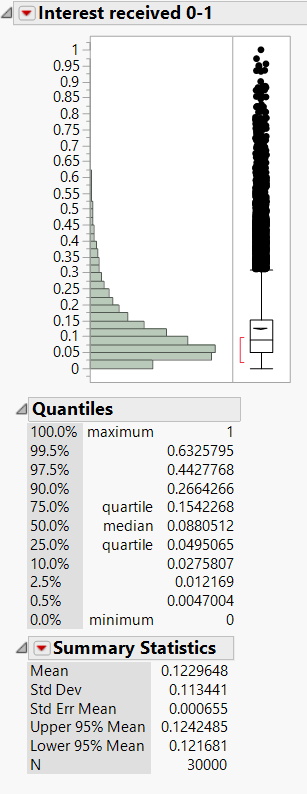
Exhibit 17: Distribution graphs for univariate outlier analysis.

As seen below in Exhibit 18, there were columns that contained extremely skewed variables such as *Delinquency 2years*, *Open to Buy BC*, and *Months since last delinq*. The team chose to transform these columns for ease of analysis, reducing skewness or for equal spread. The team chose to apply a continuous fit to the extremely skewed variables and the best transformation was plotted (as seen in Appendix D) and as seen in Exhibit 18, it was noted that only for *Open to buy BC*, the transformed data spread out.

Exhibit 18: Extremely skewed *Open to Buy BC* before and after transformation

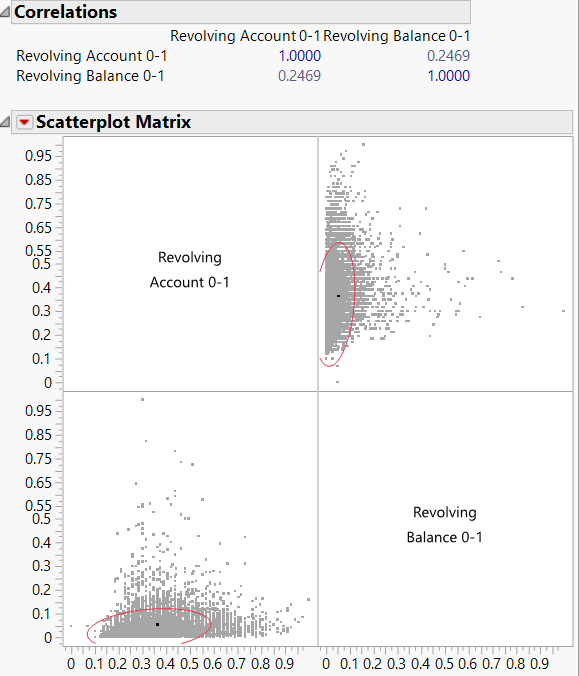
In addition, the team chose to take a specific evaluation of the Inter Quartile Range (IQR) which measures the dispersion or variation and can be depicted with the equation IQR= Quartile 3- Quartile 1. For normal variables, if a value is higher than the 1.5\*IQR above the upper quartile Q3, the value will be considered an outlier. Similarly, if a value is lower than 1.5\*IQR below the lower quartile, the value will be considered an outlier. But, if the distribution is skewed, the 1.5\*IQR might lead to a significant number of values to be outliers. Therefore, here the team has used a method of capping or also referred to as flooring of three. So, the lower limit for acceptable range of values become Q1-3\*(Q3-Q1) and upper limit for acceptable range is Q3+3\*(Q3-Q1).

For example, in the case of *Interest Received* as seen in Exhibit 19 below, as the data is significantly skewed and a large proportion of the values lie outside 1.5\*IQR. These are not outliers. So, with capping, values greater than Q3+3\*(Q3-Q1) and lesser than Q1-3\*(Q3-Q1). Here, Q1 is 0.0495 and Q3 is 0.154. Lower limit for acceptance range= 0.0495-3\*(0.154-0.0495) =-0.264. As the value is negative, and the interest received start from 0, there are no lower outliers. Upper limit for acceptance range= 0.154+3\*(0.154+0.0495) =0.7645. All values above 0.7645 will be outliers for *Interest Received*. Similarly, outliers can be found out for all continuous variables, normal or skewed.

Exhibit 19: Skewed distribution graph for *Interest Received*

### *Bivariate*

To conduct a bivariate outlier analysis, the team discovered correlation between *Revolving Account* and *Revolving Balance*. In JMP by default, a 95% bivariate normal density plot is shown in the scatterplot. If the pair of variables have a bivariate normal distribution, the ellipse encloses about 95% of the points. The narrowness of the ellipse will determine the correlation between two variables. The values lying outside the ellipse is usually considered outliers. In the below example, as shown in Exhibit 20, we can see that the variables are not correlated and a lot of values will be considered as outliers. Similarly, a scatterplot matrix was plotted for all continuous variables and a full overview including the correlation review of the matrix can be found in [Appendix D](#_Appendix_D:_Outlier).

Exhibit 20: Scatterplot matrix of Revolving Accounts and Revolving Balance

### *Multivariate*

Lastly, we use the robust estimation method for outlier detection called Mahalanobis Distance which can be seen in Exhibit 21. This is an alternative to outlier rejection where robust procedures will down weigh the outliers which reduces their effect on the estimation without the heavy-handed step of rejecting them. By capping the dataset, the team determined the upper limit for acceptable range (for example, Q3+3\*(Q3+Q1) = 4.522+3\*(4.522+3) = 9.088). Accordingly, all values with Mahalanobis distance greater than 9.088 can be considered outliers for this specific data set.

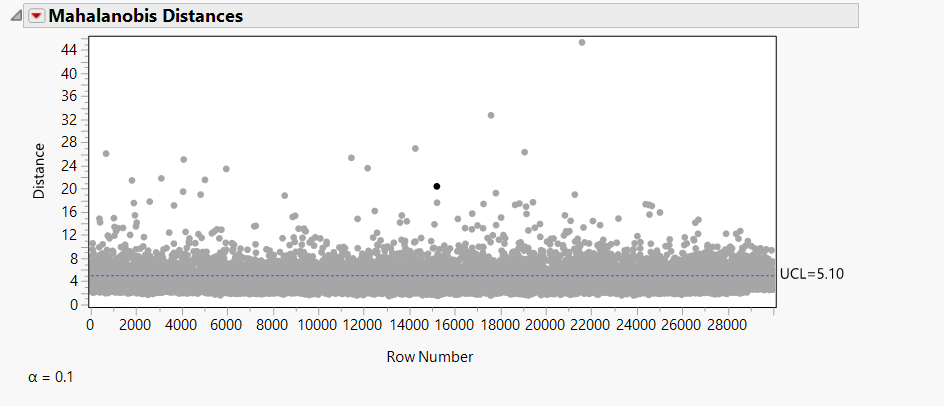


Exhibit 21: Mahalanobis distances plotted for the continuous variables with an Alpha level 0.1

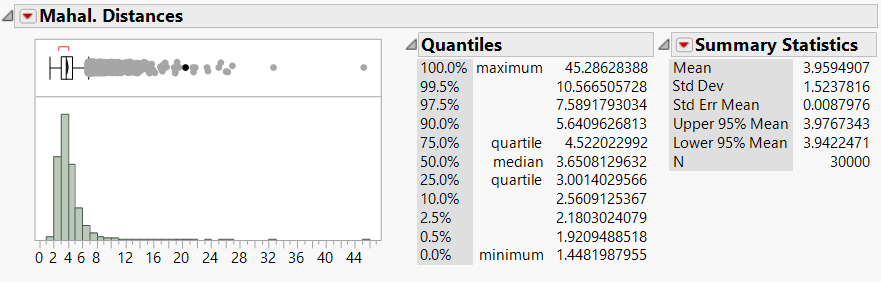


Exhibit 22: Distribution graph of saved Mahalanobis distance

As mentioned, the team have decided not to remove any values because an outlier is not always an error and sometimes it provides important information about the data. Valid data should not be removed just because it does not conform to our theory of reality. In addition, the team agreed that the reason why the outlier occurred should always be investigated and taken into consideration for modeling. As the business scenario is unknown, this investigation for these outliers is not possible. The modeler should make the decision on whether these outliers are to be removed or kept.

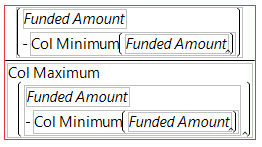
## Scaled Variable

In the final dataset, the user will find that all continuous variables have been scaled. These columns can be identified in the header as ‘*Column Name 0-1’.* The MinMax formula (Xi-Min(X)/(Max(X)-Min(X)) was applied to each continuous column to create a new scaled value for every record and it is scaled from zero to one. The value of scaling each continuous variable is when finding relationships between two variables, any variable in a larger scale will exert more influence on the relation when compared to other variables. Since there are no variables in the dataset that should be given more weight, all variables have been brought to the same scale.

Variables measured at different scales do not contribute equally to the analysis. For example, a variable that ranges between 0 and 100 will outweigh a variable that ranges between 0 and 1. Using these variables without standardization in effect gives the variable with the larger range a weight of 100 in the analysis. Transforming the data to comparable scales can prevent this problem.

See Exhibit 23 for an example of a new scaled column that was created from the MinMax formula. The scaled variables are recommended for modeling since they have the same variance of the original column, just shown between the value of zero to one. The scaled values have been provided in the primary data set. If the user would like to use the original data, we have kept unscaled data in the secondary reference data set.

Variables measured at different scales do not contribute equally to the analysis. For example, in boundary detection, a variable that ranges between 0 and 100 will outweigh a variable that ranges between 0 and 1. Using these variables without standardization in effect gives the variable with the larger range a weight of 100 in the analysis. Transforming the data to comparable scales can prevent this problem. Typical data standardization procedures equalize the range and/or data variability



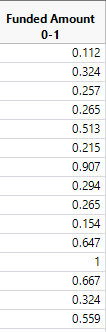
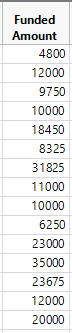
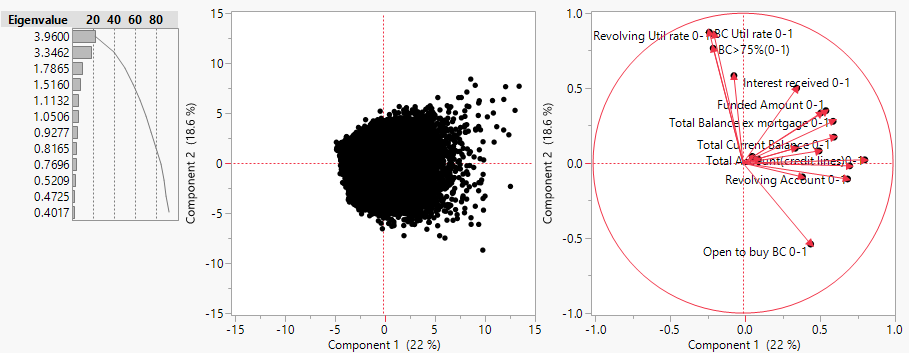


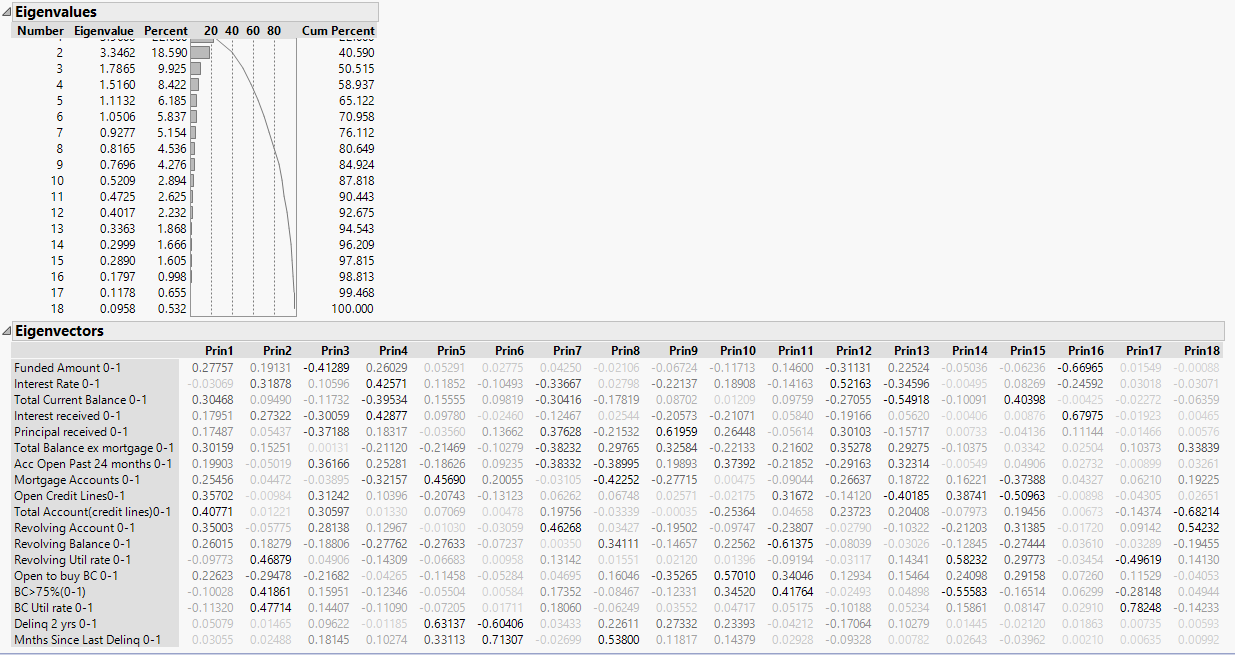
Exhibit 23: Funded Amount before and after scaling

## Principle Component Analysis

While exploring the data, the team noticed even after reducing the number of columns that were unnecessary there was still numerous continuous variables in the data set. To reduce the dimensionality, while keeping variance, the team performed a principal component analysis. If successful, the value of this exercise provides calculated columns that together still provide the information of the original data to the model, with less number of variables.

While performing a principal component analysis on the data set, it was found that to keep about 95% of the variance, 13 of the principal component variables would need to be used (See Exhibit 24). The team analyzed why so many variables would be needed and concluded that while removing columns that correlated or could be derived from our dataset, the principle component columns would be less useful because there was not much correlation left in the data set between variables. However, reducing the number of continuous variables from 18 to 13 could be useful for a modeler depending on the business scenario and how much variance is acceptable. Since the team is not able to make this decision for the modeler, the principal component variables have been added to the primary dataset.





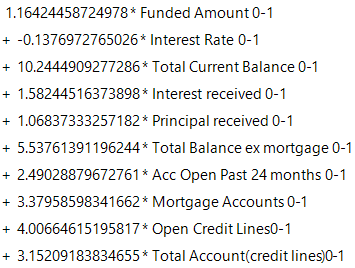


Exhibit 24: Principal Component Analysis on the sample,

## Conclusion & Recommended Next Steps

In conclusion, through the entirety of this process the team could gain tremendous insight to the data that was given through exploration and completed the necessary modifications, which led to a model ready data set. The team chose to submit two datasets for the modelling user to refer to as mentioned in the executive summary, which contained a primary and secondary dataset version.

For the future use of this data, it is recommended to use the column Loan Status as the target variable for modelling. The training dataset could include the Fully Paid and Default categories, and the testing dataset could be the Current loans. With this, the business could predict what loans will become fully paid, and which will default. Even though this report outlines the basic pre-processing that the team has completed, it is possible once the modeler starts to use the data, additional data processing will need to be completed depending on the business scenario.

Appendix

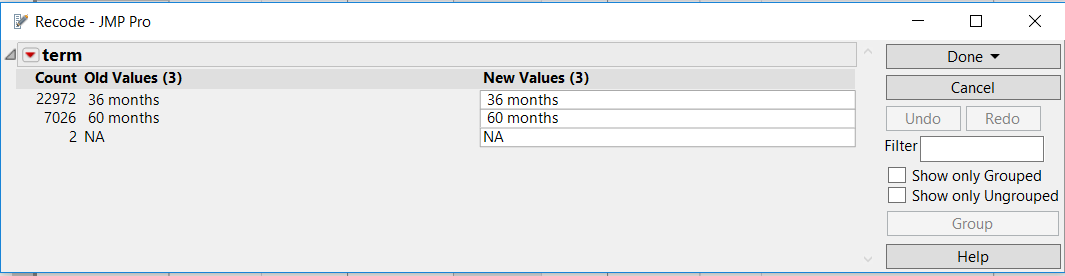
## Appendix A: Data Dictionary

|  |  |  |
| --- | --- | --- |
| **Title** | **Variable Type** | **Description** |
| Member ID | Continuous | A unique LC assigned Id for the borrower member. |
| Purpose | Nominal | A category provided by the borrower for the loan request. |
| Funded Amount | Continuous | The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value. |
| Interest Rate (%) | Continuous | Interest Rate on the loan |
| Term | Nominal | The number of payments on the loan. Values are in months and is 36 months with the value 0 (zero) and 60 months with the value 1 (one). |
| Grade | Ordinal | Loan grade assigned based on loan Interest rate |
| Loan Status | Nominal | Status of the loan |
| Employee Length (years) | Ordinal | Employment length in years. Possible values are between 0 and 10 where 0 means less than or equal to one year and 10 means ten or more years. |
| Annual Income Binning | Ordinal | The annual income provided by the borrower during registration, is placed into corresponding "bin" for ease of data analysis. |
| Income Verified | Nominal | Indicates if the income was verified or not. Income verified or source verified is grouped to value 1 (one) and not verified is grouped to 0 (zero). |
| Total Interest Received | Continuous | Total interest received to date |
| Total Principal Received | Continuous | Total principal received to date |
| Delinq 2 Yrs | Continuous | The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years |
| Months Since Last Delinq | Continuous | The number of months since the borrower's last delinquency. |
| Total Current Balance | Continuous | Total current balance of all accounts |
| Total Balance excluding Mortgage | Continuous | Total credit balance excluding mortgage |
| Accounts Open Past 24 mths | Continuous | Number of trades opened in past 24 months. |
| Mortgage Accounts | Continuous | Total number of mortgage accounts. |
| Open Credit Lines | Continuous | Total number of open credit lines in the borrower's credit file. |
| Total Account (credit lines) | Continuous | Total number of credit lines currently in the borrower's credit file |
| Revolving Accounts | Continuous | Total number of revolving accounts |
| Revolving Balance | Continuous | Total credit revolving balance |
| Revolving Utilization Rate | Continuous | Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. |
| Open to buy BC | Continuous | Total open to buy on revolving bankcards. |
| BC > 75 (%) | Continuous | Percentage of all bankcard accounts > 75% of limit. |
| BC Utilization Rate | Continuous | Ratio of total current balance to high credit/credit limit for all bankcard accounts. |

## Appendix B: Data Modification

Appendix B.1 *Term*

**Original:**



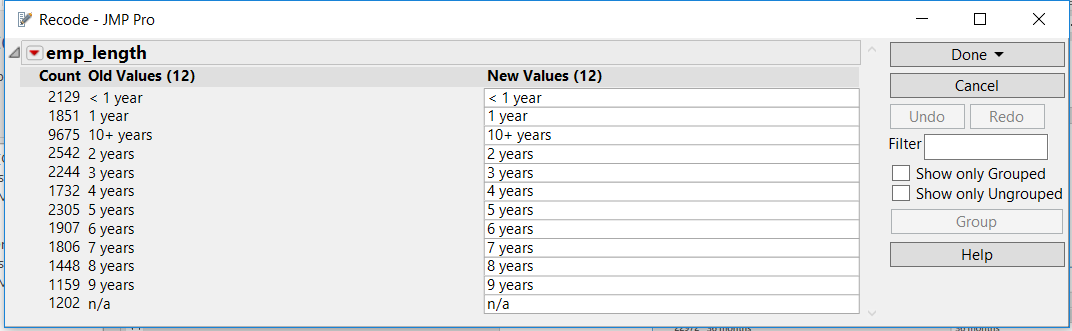
**Action**: Replaced 36 months with 0 and 60 months with 1.

**Result:**



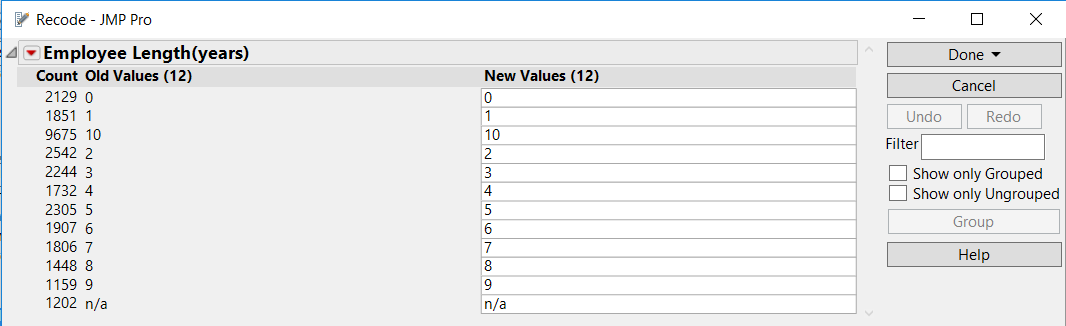
Appendix B.2 *Employee Length*

**Original:**



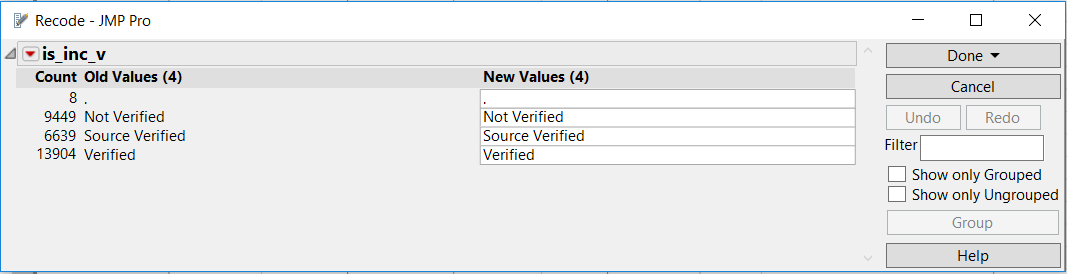
**Action**: As per the data dictionary, changed <1 year to 0 and 10+ years to 10

Result:

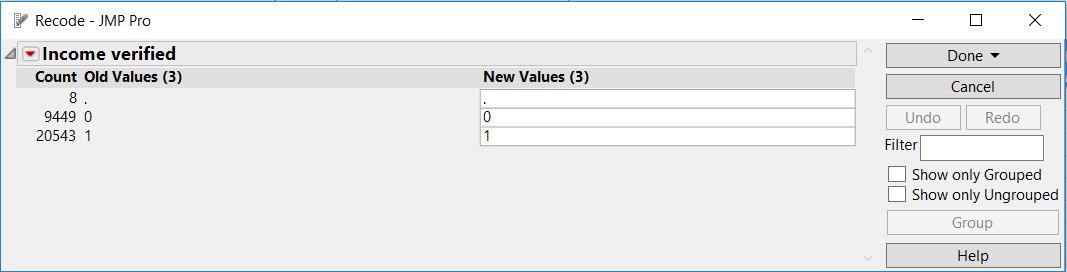


Appendix B.3 *Income Verified*

**Original:**

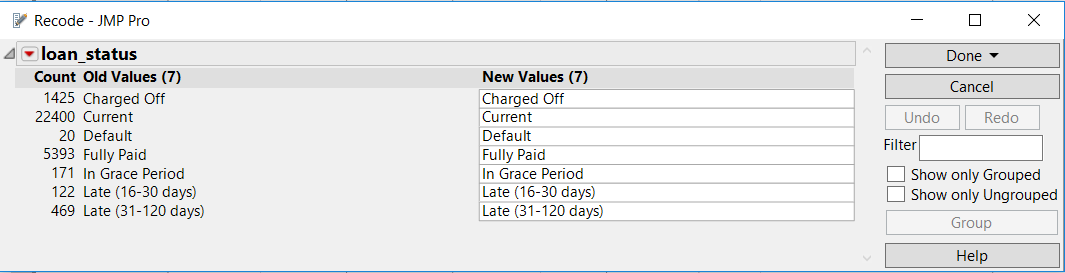


**Action**: Replaced Verified and Source verified with 1, not verified with 0



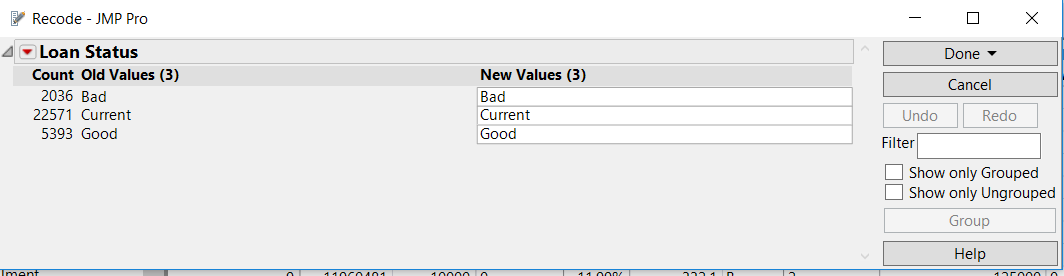
Appendix B.4 *Loan Status*

**Original:**



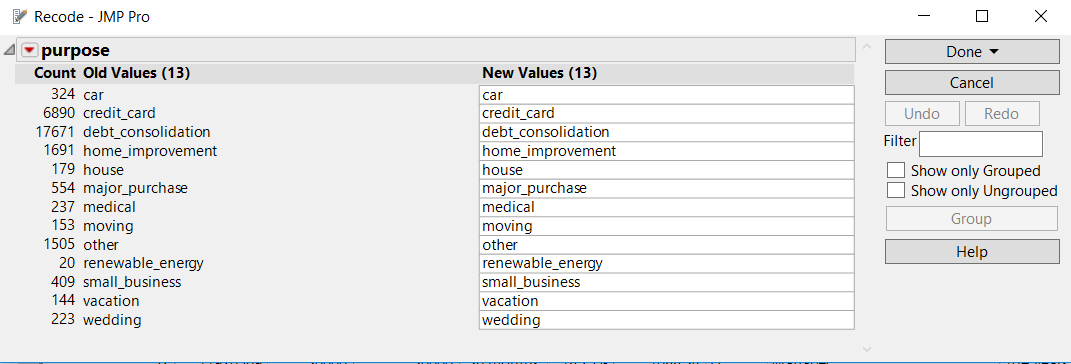
**Action**: Replaced Full paid with ‘good’, Current and grace period with ‘Current’ and the rest with ‘Bad’.

**Result**:



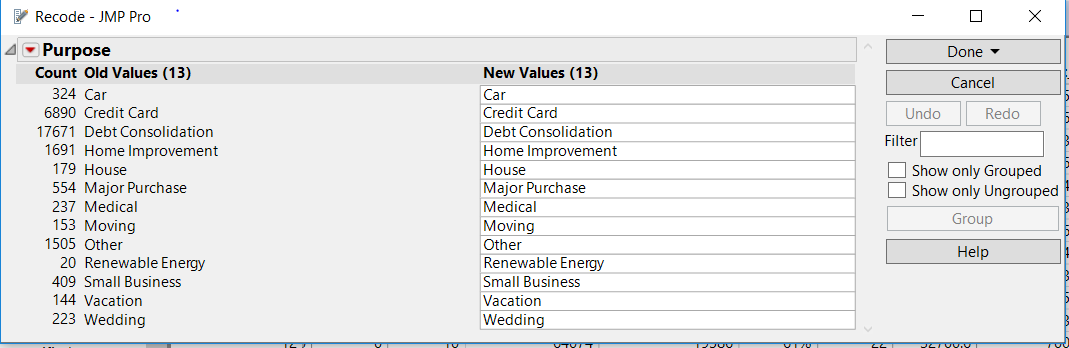
Appendix B.5 *Purpose*

**Original**:



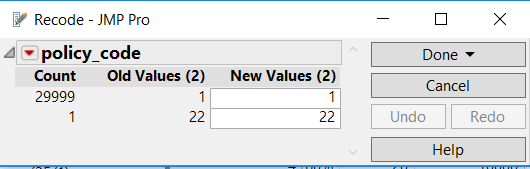
**Action**: Replaces underscore with space and used proper case.

**Result**:



Appendix B.6 *Policy Code*

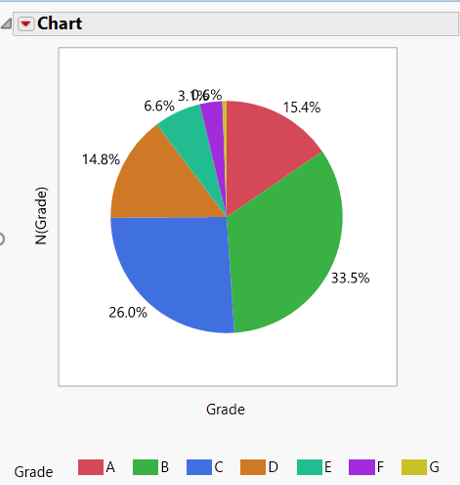
**Original**:



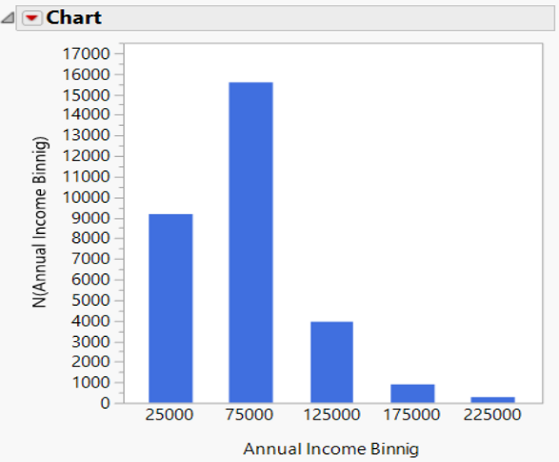
**Action:** Column was removed as there is no variance in the data.

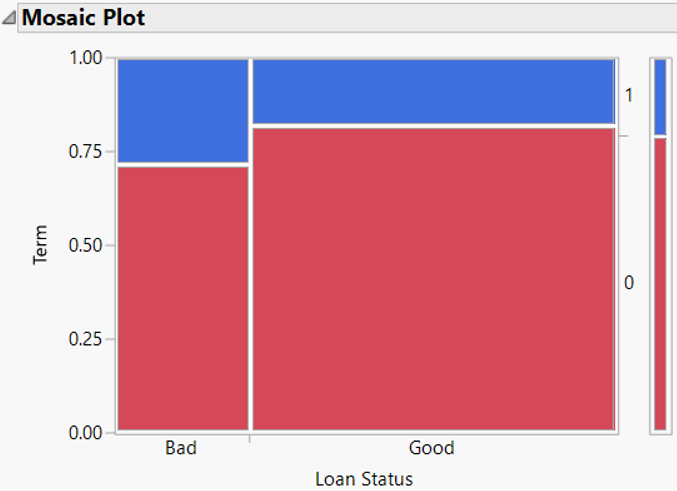
## Appendix C: Data Visualizations

Appendix C.1 *Grade*

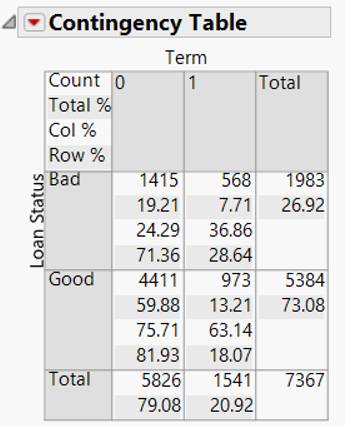


Appendix C.2 *Annual Income*

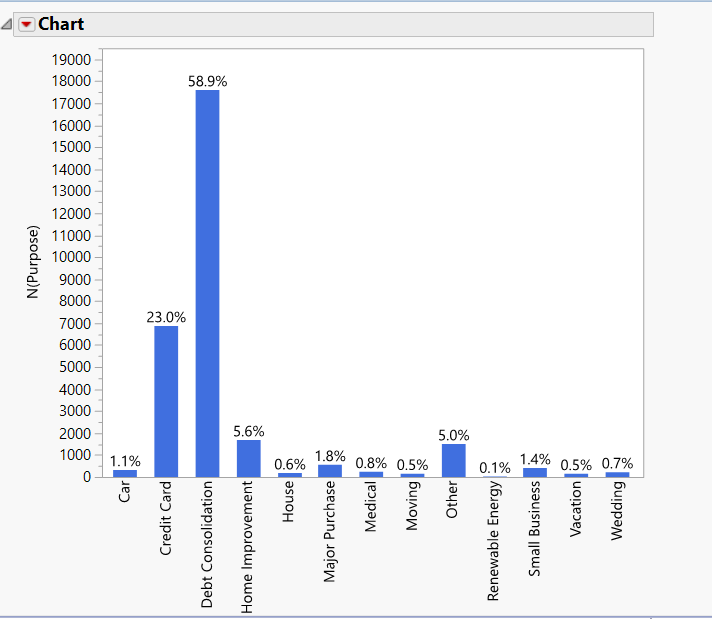


Appendix C.3 *Loan status* and *Term*

Appendix C.4 *Loan status* and *Term*

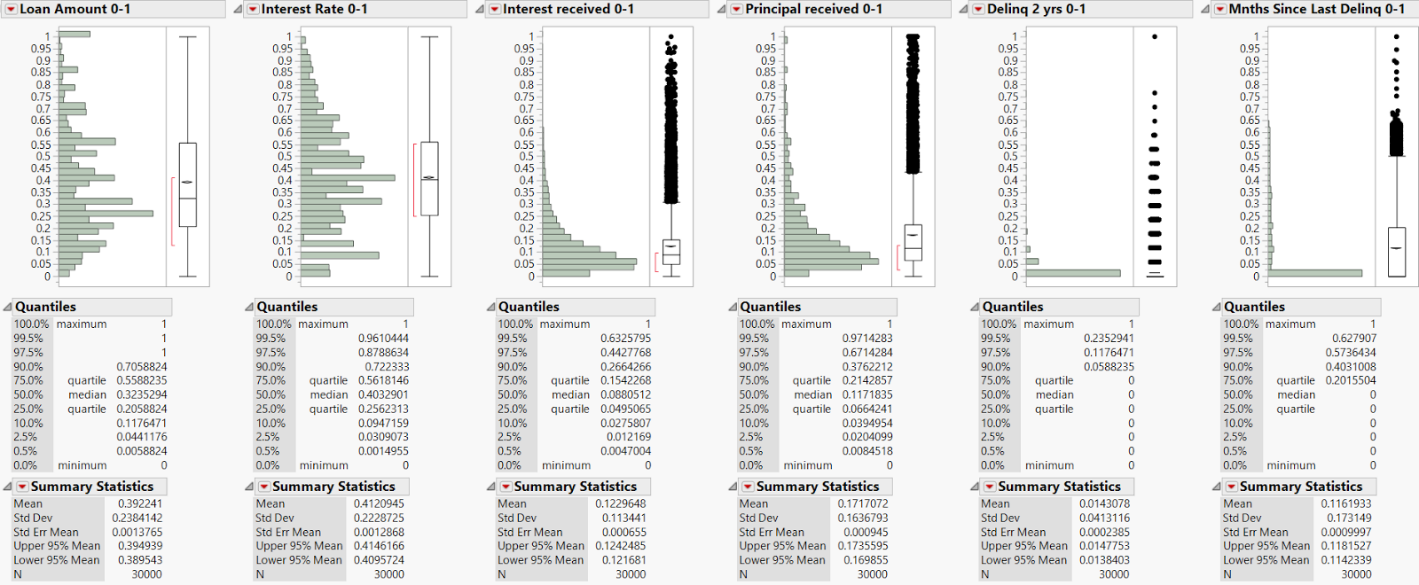


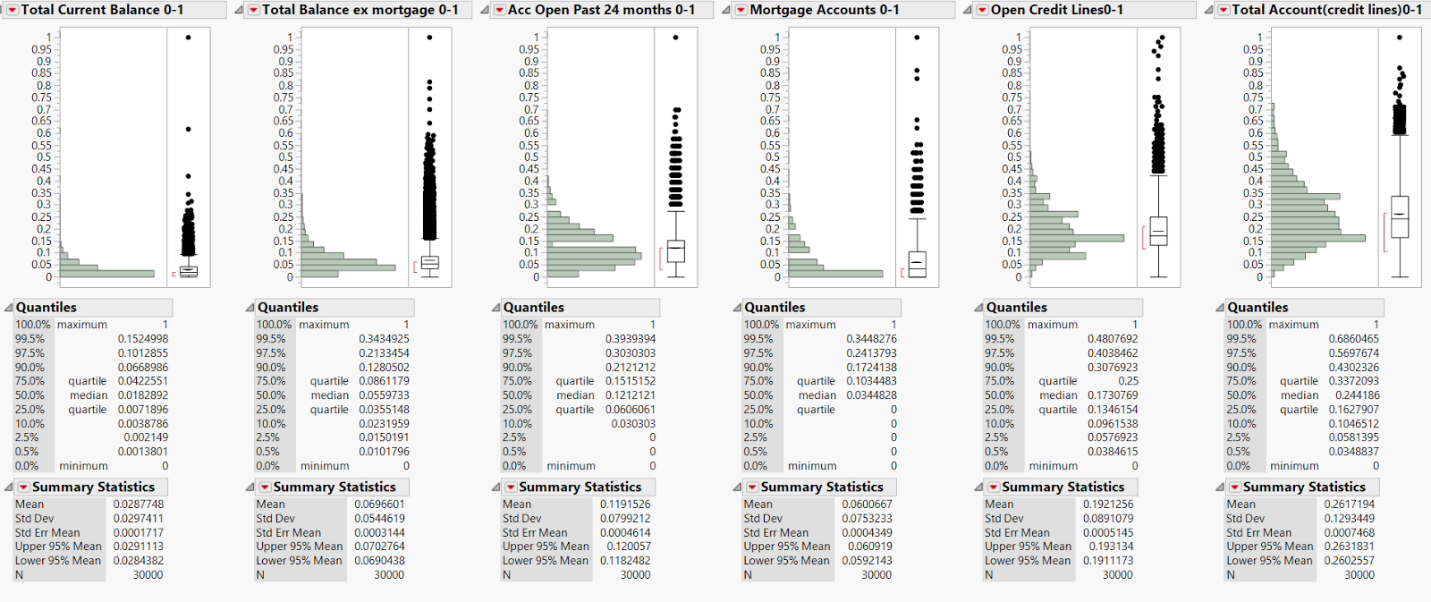
Appendix C.5 Bar graph of *Purpose*

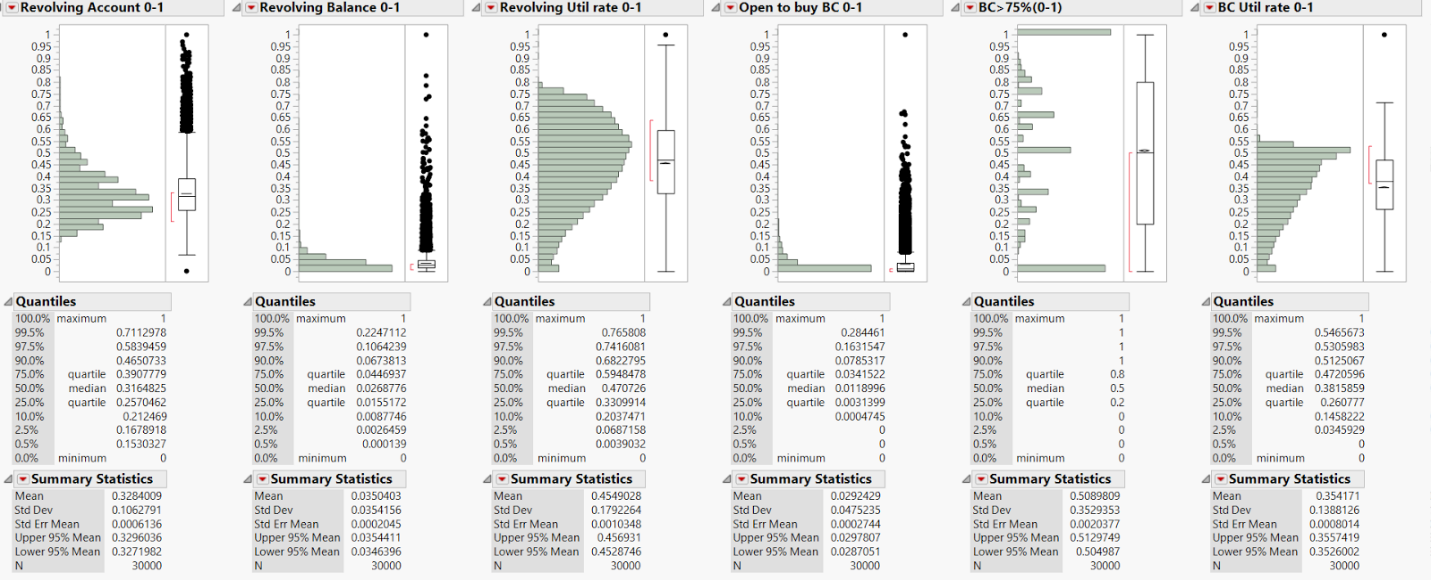


## Appendix D: Outlier Analysis

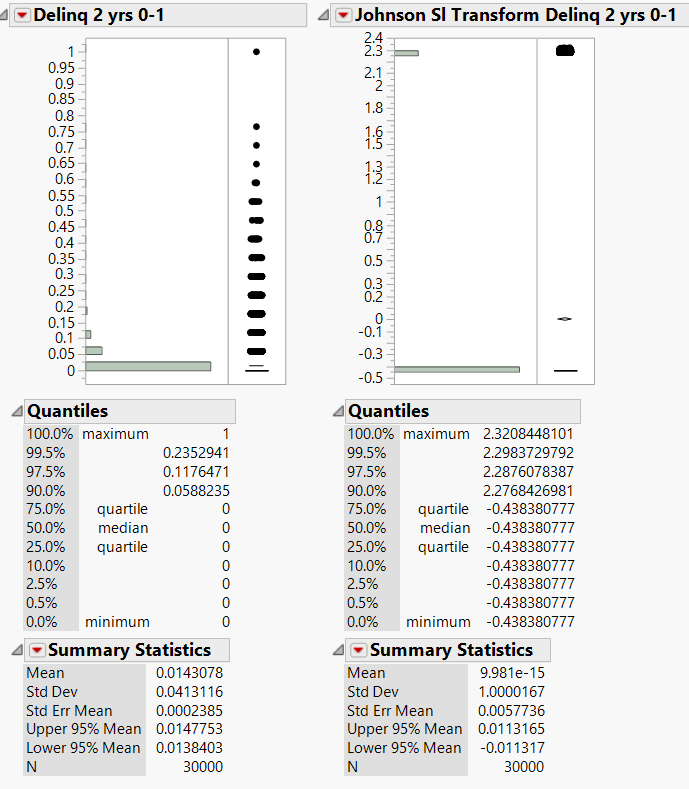
Appendix D.1 **Univariate Distribution**

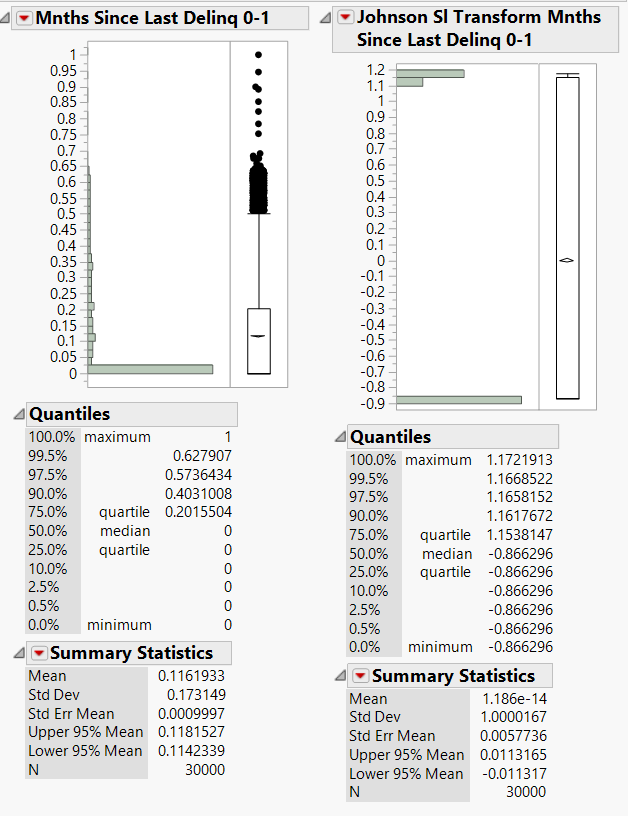






Appendix D.2 **Transformed Data**





Appendix D.3 **Bivariate Distribution**

