**Investigating Late Loan Payments**

Team 7

Abe Archer, Kennedy Maloney, Isabella McCoy, Grady Smith

University of Arkansas

DASC 1223H – Role of Data Science in Today’s World

Dr. Schubert

05/03/2024

**Introduction**

This project was concerned with analyzing the data surrounding loans that are late on their payments. In order the examine the information, this project was split into four different sections, each of which is distinguished by a different hypothesis. The first hypothesis was that there would be a prominent correlation between installments and the amount of interest a loan has. The second hypothesis was that there would be clear correlations between annual income and the number of delinquent accounts for a customer, as well as between annual income and remaining outstanding principle for total amount funded. Additionally, the second hypothesis speculated that income would be a major predictor of financial responsibility for a prospective customer. The third hypothesis was that there would be strong correlations between variables that had similar origins, such as the funded amount and the funded amount by investors. It was also hypothesized that there would be similar correlations between values in the late payments and in the full dataset. The fourth hypothesis was that there were subgroups the data could be broken into to help banks accurately anticipate when a customer might make a late payment. All of these hypotheses were individually explored and then examined together in order to find conclusions on the late-paid loans.

**Hypothesis 1**

This section of the analysis was concerned with exploring variables using ANOVA. Put most simply, ANOVA is a statistical analysis used to investigate the variance among variables, and although ANOVA can be seen as more of a simple method of statistical analysis, it is very powerful in helping see if there is or isn’t correlation. In this case, an ANOVA analysis was performed on interest and on other target variables that it was compared to. The residual mean squared value was calculated and compared for each of these. From these ANOVA analyses, histograms and ggPlot graphs were used to examine the area of variance between variables. As a basic rule of thumb for the histograms, a more typical-looking distribution indicates correlation, and for the qqPlot, a more positive linear relationship indicates correlation.

Before the ANOVA analysis was performed, some of the data involved was surveyed to find any additional information about the interest rates of late payment loans. This was done by developing a boxplot (Figure 1) and a density function of the interest rate (Figure 2). One interesting thing that was found about interest rate, before even going into the further statistical analysis, was that there was an almost normal distribution to the interest rates density besides a dip in the middle and some subtle hills as it tapers outside. It was also found that the mean of the interest rate was 15.95, the maximum value of the interest rate was 28.99, and the minimum value of the interest rate was 5.32.

A diagram of a box plot

Description automatically generated

**A graph of a function

Description automatically generatedFigure 1**

**Figure 2**

The hypothesis for the ANOVA analysis were first that there will be a prominent correlation between installments and the amount of interest a customer a customer’s loan contains. This is because installments will be a good indicator of whether a customer is having to pay a volatile amount that could be compromised. It was also hypothesized that the interest will hold the weakest correlation with annual income, due to the fact that there are some very extreme outliers for the annual income.

The residual mean squared value found for the installments was 18 (Figure 3, Figure 4). For the annual income, it was 18.8 (Figure 5, Figure 6), and for the loan amount, it was 18 (Figure 7, Figure 8). The outstanding principal amount had a mean squared value of 17 (Figure 9, Figure 10), the DTI’s was 19 (Figure 11, Figure 12), and the total payment’s was also 19 (Figure 13, Figure 14).

A graph of a graph of a rate

Description automatically generated with medium confidence

**Figure 3**

A graph with a line

Description automatically generated

**Figure 4**

A graph of a graph of income

Description automatically generated with medium confidence

**Figure 5**

A graph with a line

Description automatically generated

**Figure 6**

A graph of a loan amount

Description automatically generated

**Figure 7**

A graph with a line

Description automatically generated

**Figure 8**

A graph of a number of percent

Description automatically generated with medium confidence

**Figure 9**

A graph with a line

Description automatically generated

**Figure 10**

A graph of a graph of a person

Description automatically generated with medium confidence

**Figure 11**

A graph with a line drawn on it

Description automatically generated

**Figure 12**

A graph of a graph of a payment

Description automatically generated with medium confidence

**Figure 13**

A graph with a line drawn on it

Description automatically generated

**Figure 14**

It can be concluded that the variable with the highest correlation to interest was the outstanding principal amount with a residual mean squared error value of 17, which did not follow the hypothesis. After doing some further research, it was found found that this does indeed add up because banks very often take the outstanding principal amount into account very early on when constructing a customer’s interest rate. The variable that had the lowest correlation to interest was the total payment and DTI (Debt to Income ratio), which both had a residual mean squared value of 19, which also did not follow the hypothesis. This was surprising, as it was expected that the DTI would be one of the variables that held a stronger correlation with the interest rate.

**Hypothesis 2**

For Hypothesis #2, the main characteristic of customers that was being investigated was their annual income. Obviously, income is valuable to any organization that is extended a loan, however, this was done to see to see if annual income would be an overextending measure of reliability within the group of late customers. It was predicted that there would be clear correlations between annual income and delinquent accounts and between annual income and remaining outstanding principle for the amount funded. More specifically, it was anticipated that the higher annual income for a prospective customer, the less risky it is to give them a loan.

The definition of a delinquent account is an overdue account that has not received payment for over a certain period that is expected. Put simply, this means that the account owner has breached the terms of the loan by not paying it within the specified period required. In addition, another variable discussed is the remaining outstanding principle for total amount funded, which represents the amount of the original loan that still needs to be paid off. After providing these definitions, understanding the results became easier.

The first relationship that was investigated was between annual income and delinquency. After doing several filtering maneuvers through the “late” dataset (dataset containing late customers), it was concluded that there were 112 instances of delinquency. Of these 112 customers, 104 had 1 delinquent account, 7 had 2 delinquent accounts, and 1 had 4 delinquent accounts. The average income for 1 delinquent account was $80,137.54, the average income for 2 delinquent accounts was $75,928.57, and the customer that had 4 delinquent accounts had an income of $100,000. Overall, the customers in this group of delinquent accounts surprisingly have a higher average annual income than the pool of late customers(($80,051.82 vs $70,571.32). After seeing this information, it seemed as if there was not a correlation between annual income and delinquency, and if anything, there would be a positive relationship between these two variables.

The next step in finalizing the results was running a linear regression model. After computing within the “lm” function in R, a p-value of 0.03704 and an adjusted R-squared value of 0.0002401 were found. To interpret these values, the p-value suggests that there is a statistically significant relationship(p<0.05). However, the adjusted R-squared value tells us that this information is limited. Due to the fact that their R-squared value is an extremely low value (0.002401), only 0.002401 of the variance in delinquency is explained by annual income. The next step in this analysis was to look at the relationship between income and the proportion of a loan paid back. To do this analysis, a new variable was created, out\_prncp\_to\_fundedamnt, to represent the ratio between the outstanding principle remaining and the funded amount. Essentially, this would represent the proportion of a loan that had been paid back by a given customer. Like the above, a linear regression model was used to explain the relationship between this new column and annual income, and was visualized in a linear regression plot to compare owed to loan ratio and annual income (Figure 15).

A graph showing a number of black dots

Description automatically generated with medium confidence

**Figure 15**

Based upon this visualization, there does not seem to be an underlying relationship, as the points of data are scattered, and the regression line has a slope of near zero. To confirm this notion, the corresponding p-value and adjusted R-squared were computed. The resulting p-value was 0.00003094 and the adjusted R-squared was 0.001172. To interpret these results, the p-value of 0.00003094 suggests that there is a statistically significant relationship, but the power of this information is limited by the low R-squared value once again. Since the R-squared value is extremely low (0.001172), only 0.001172 of the variance in proportion paid back is accounted for by annual income.

In conclusion, although there are statistically significant relationships between both annual income and delinquency and annual income and proportion paid back, there is little evidence that would support the claim predicted by Hypothesis #2. Since the adjusted R-squared values are considerably low, it can be concluded that annual income is not the best predictor of either the possibility of delinquency or proportion paid back. Income is undoubtedly an important factor for banks and creditworthiness, or else the bank would not ask for that information. However, there are other factors that might need to be considered equally. For example, a businessman, who has an above average annual income, can qualify for a loan for a business venture because of his income, but this venture can come with risk and the return on investment could be low. Another exception could be psychological factors. Naturally, some customers might be more psychologically inclined to be more impulsive with financial decisions, leading to an unbeknownst risk for the bank. Based upon this hypothesis testing, income should not be an overextending factor for predicting a customer’s behavior.

**Hypothesis 3**

This section of the analysis was concerned with finding correlations between variables using correlation matrices, and exploring those further. The hypothesis developed for these correlations was that variables that are similar in nature, such as the funded amount of the loan (funded\_amnt) and the funded amount by investors (funded\_amnt\_inv), would have a strong correlation between them. It was also hypothesized that there would be similar correlations between variables for the late payments and for the full dataset.

Some brief EDA was performed before the analysis was done, intended to help find any correlations between large variables. The annual income was compared to the loan amount for the late payments (Figure 16) and full dataset (Figure 17), and was colored by grade. The R-value for the late payments graph was 0.461, and for the full dataset, an N/A value was returned. Next, the installment was compared to the loan amount for both late payments (Figure 18) and the full dataset (Figure 19), and again colored by grade. These graphs in particular show a stratifying of the grades, which seems to indicate that the grade is a strong indicator used when determining what installments borrowers will have to pay. The R-value for the late payments was 0.9448 and for the full dataset it was 0.945.

A graph showing a number of payments

Description automatically generated

**Figure 16**

A graph of data set up

Description automatically generated with medium confidence

**Figure 17**

A graph showing a line of multicolored dots

Description automatically generated

**Figure 18**

A graph showing a rainbow of dots

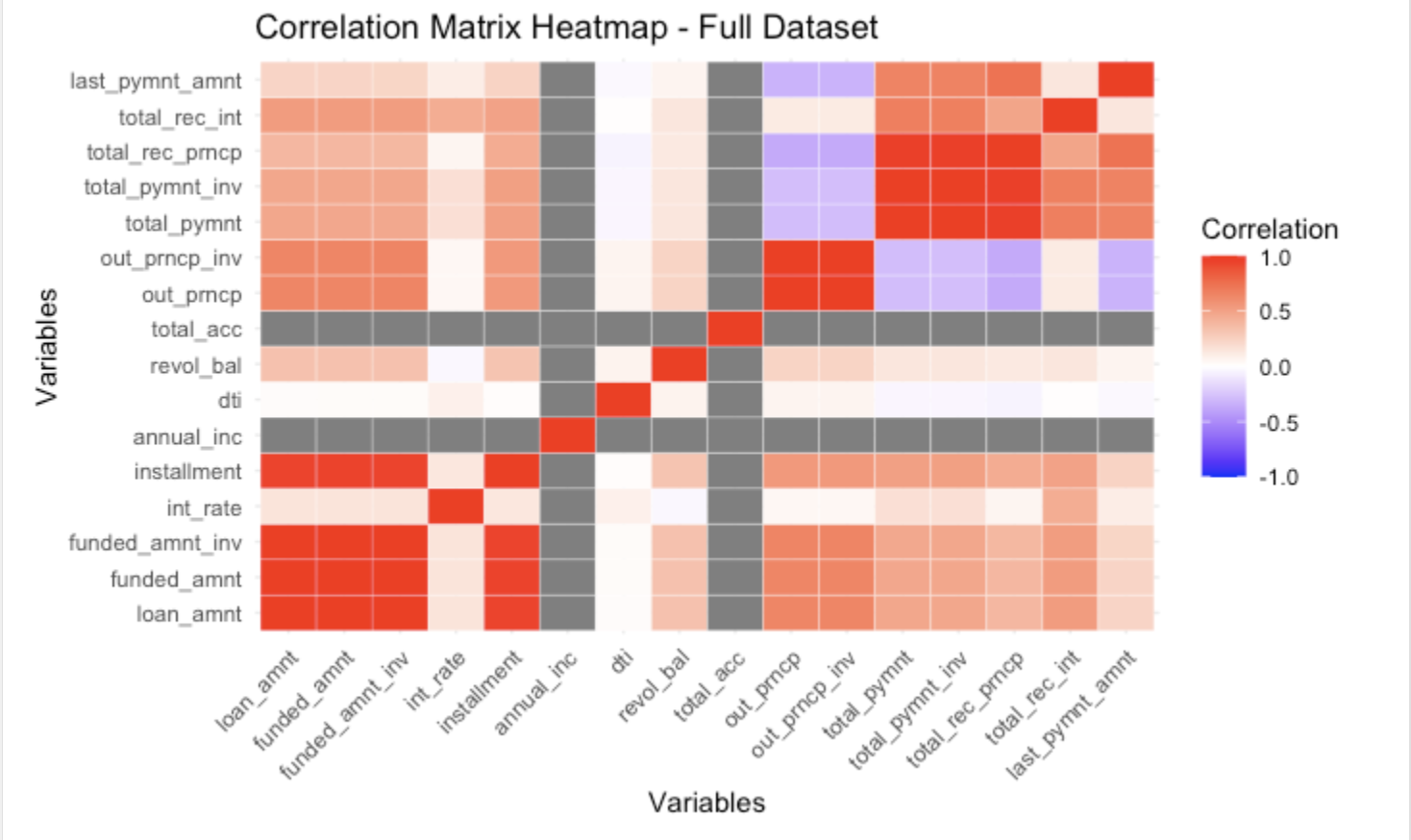
Description automatically generated with medium confidence

**Figure 19**

In R, it was difficult to develop a correlation matrix for every numeric variable in the set, as many did not have entries in every row. For these variables, the matrix only display a null variable. In order to avoid this, only columns that were numeric and had an entry in every row were included in the correlation matrix for the late payments (Figure 20). Next, a similar matrix was constructed for the full dataset, using the same variables as the late payments dataset. As some variables would have entries for every row in the late payments dataset but not necessarily for the full dataset, some of these correlations returned as null. These are still visible in the full dataset heatmap in order to more clearly compare it to the late payments dataset (Figure 21).A red and white graph

Description automatically generated

**Figure 20**



**Figure 21**

Based on this matrix, it is clear that there are correlations between the expected variables of funded\_amnt and funded\_amnt\_inv, as well as loan\_amnt when compared with both of these, and installment compared with funded\_amnt, funded\_amnt\_inv, and loan\_amnt. Similar correlations could be seen for total\_payment and total\_payment\_inv. Next, these correlations were explored further using scatterplots and comparing the R-values for the plots. While not all of the correlations are the same for both matrices, many are, particularly those we are concerned with.

The first scatterplots are for funded\_amnt vs. loan\_amnt. Here we can see the correlations between these variables for the late payments data (Figure 22), which has an R-value of 0.9998, and the full dataset (Figure 23), which has an R-value of 0.9992. Both have been colored by the status of the loan, where we can see the differences in how late different loans are, as well as the differences in the entire dataset. Clearly, both have a strong correlation, particularly the late payments specifically, which we can see by the coloring in the full dataset where the late payments are the data points that specifically follow a very straight line.

A graph with a line drawn on it

Description automatically generated

**Figure 22**

A graph with a line graph and a chart with text

Description automatically generated with medium confidence

**Figure 23**

The next scatterplots developed were for funded\_amnt\_inv vs. funded\_amnt. The structure and coloring of these graphs was similar to the funded\_amnt vs. loan\_amnt graphs, and there is a similar trend. The late payments follow a very specific trend (Figure 24), and while the full dataset also follows a strong trend (Figure 25), the non-late payment entries are more scattered. The R-value for the late payments was 0.9999, and the R-value for the full dataset was 0.998.

A graph with a line drawn on it

Description automatically generated

**Figure 24**

A graph with a line graph and text

Description automatically generated with medium confidence

**Figure 25**

The next scatterplots were for total\_pymnt\_inv vs. total\_pymnt. Again, we can see an extremely strong trendline in the late payments (Figure 26), although this time with slightly more outliers than the previous two late payment graphs (Figure 27). The full dataset also follows a trend, although less direct than the late payments. Unlike the previous two graphs, there are some data points that are not late payments that do follow the line set by the late payments. The R-value for the late payments was 0.5744, and for the full dataset was 0.5158.

**A graph with colorful dots

Description automatically generated**

**Figure 26**

A graph with a line

Description automatically generated

**Figure 27**

**Hypothesis 4**

A null hypothesis was established to analyze late payments being broken into subgroups. The null was that the number of clusters was greater than two with the alternate being that the number of clusters present was equal to 2. A histogram of loan status was made to understand how the data was distributed (Figure 28). The result was approximately 75% of the late payments fell between 31-120 days while the other 20% was comprised of the 16-30 day late categories.

A graph of a distribution of loan status

Description automatically generated

**Figure 28**

After establishing the null hypothesis, the data had to be put into a readable format for R. This began by filtering the dataset into just late payments. The data was then stored as df\_filtered and put through a series of pipes that removed the loan status column. This column was removed so the algorithm couldn’t “cheat” and use the group labels when clustering. To cleanse the data further, everything in df\_filtered was made into a numeric value so it could be put into a scaled matrix. The new dataset, named df\_scale, was then put through another series of pipes to take a random sample of 5,000 rows. A random sample was taken to ensure the process could run efficiently and reflect the main data set without using as much computing power.

To determine the optimal number of clusters for the k-means clustering analysis, the NbClust package in R was employed. This involved running the NbClust function on the scaled dataset, denoted as df\_scale, utilizing the Euclidean distance metric. The function was configured to explore the range of clusters from a minimum of 2 to a maximum of 10, employing the k-means clustering method. Subsequently, the results were visualized using the fviz\_nbclust function from the factoextra package (Figure 29). This visualization provided insights into the optimal number of clusters by presenting various clustering indices, such as silhouette width or gap statistics, across different cluster counts. This process aided in making an informed decision regarding the appropriate number of clusters for subsequent clustering analyses.

A graph of a graph showing a number of different colored dots

Description automatically generated with medium confidence

**Figure 29**

It was determined that two was the best number of clusters for this data. The optimal number of clusters was then used in a K-means analysis.The k-means clustering analysis was conducted using the kmeans function, which clustered the data into two distinct groups (centers = 2). The procedure was set to iterate a maximum of 10 times (iter.max = 10) with a single initial configuration (nstart = 1). This clustering step facilitated the generation of kmeans\_result, which was subsequently utilized for visualization. To visualize the results of the k-means clustering analysis, the fviz\_cluster function was employed, utilizing the k-means result object generated from the k-means clustering procedure. The function was configured to display the clustering results as points on a plot, with the df\_scale dataset used as the data source. Additionally, the geom parameter was specified as "point" to represent each data point as a point on the plot. The clustering results were not standardized (stand = FALSE) to preserve the original scale of the data.

**Conclusion**

Each conclusion was either accepted or rejected. The first conclusion was rejected, due to the variable with the highest correlation to interest being the outstanding principal amount with a residual mean squared error value of 17. The variables with the lowest correlation to interest were the total payment and the DTI, both of which had a residual mean squared value of 19. This did not follow the hypothesis either. The second hypothesis was also rejected, determining that income was not as reliable of an indicator as it had been hoped. While there were meaningful relationships between income and delinquent accounts, as well as the proportion of the loan paid, there was not enough evidence to suggest that income alone is a significant predictor of borrower responsibility. The third hypothesis was accepted, with the values having similar sources having strong correlations to each other. There were also strong similarities between total correlations between variables when comparing the late payments to the full dataset. However, this could be partially due to the late payments having unusually strong correlations for certain variables, which could be pulling up the correlation in the full dataset. The fourth hypothesis was rejected, with the clusters that were developed only showing what was already seen from the data.

Altogether, three of the four hypotheses were rejected. However, this does not indicate that there are no ways of determining what loans are likely to eventually have late payments. There could be other personal factors not involved in the dataset that would be strong indicators of late payments that more situational to the borrower than something like income. Even still, there could be other factors that are in the dataset that were simply not analyzed in this research. There could still be many indicators of late payments, and in order to find these, the issue needs to be studied further.