FIN 434 Freddie Mac Data Project

Team 3: Gordon Miller, Chandler Dessenberger, Sean Irons, Chris Shin, Zach Wieder

Contents

**[Introduction](#_Toc89554562)** [1](#_Toc89554562)

**[Variable Selection:](#_Toc89554563)** [1](#_Toc89554563)

**[Univariate Analysis:](#_Toc89554564)** [3](#_Toc89554564)

**[Bivariate Analysis:](#_Toc89554565)** [9](#_Toc89554565)

**[multivariate Analysis:](#_Toc89554566)** [14](#_Toc89554566)

**[Conclusion](#_Toc89554567)** [16](#_Toc89554567)

# **Introduction**

In this report we are providing the results of the data analysis of the 2017 Freddie Mac Mortgage data (historical\_data\_2017Q1.txt) which can be found at [Freddie Mac Standard Loan-Level Dataset (embs.com)](https://freddiemac.embs.com/FLoan/Data/downloadA.php). This report is divided into four parts: Variable selection ,univariate analysis, bivariate analysis, and multivariate analysis. Additional information about the dataset used for this analysis can be found at [Microsoft Word - SF LLD User Guide Release 29.docx (freddiemac.com)](http://www.freddiemac.com/fmac-resources/research/pdf/user_guide.pdf).

# **Variable Selection:**

When first choosing variables, we read carefully over the user guide for the data referenced in the introduction. Following reading over the user guide we decided that our direction for our analysis would be to find which variables that we think would be best to maximize the proportion of explained variation in the Interest Rate. After reading over the user guide, we used our finance knowledge to decide that our variables should be credit score, debt to income, loan to value, first time homebuyer flag, loan term, property state and loan purpose. We planned on using first time homebuyer flag and loan purpose as dummy variables in a regression, and wanted to take the state date to make a map of average interest rates for each state.

We decided to create a correlation matrix for numeric variables to aid our decision. From the correlation matrix we realized that loan term has the highest correlation with interest rate, so we would like to explore that further in our analysis (*Figure 1*). We also found that loan to value seems to have a weak positive correlation with interest rate and wanted to explore this more but there was an issue. The problem is there is a statistically significant correlation of 0.297 between loan term and loan to value. To avoid possible multicollinearity in our multiple regression model we decided to drop one of the two variables. Because loan term has a higher correlation with interest rate than loan to value, we decided to drop loan to value. Credit scores surprisingly have a correlation close to zero with most of the numeric variables.

For categorical variables we used the user guide to infer which variables would be best for our regression and most interesting to plot and explore. We ended up choosing Property State so that we could create a heat map of the average interest rate for each state. Additionally, we chose Loan Purpose and First Time Homebuyer Flag to use as dummy variables in our regression. This left us with eight variables for our analysis: credit score, first time homebuyer flag, original debt to income, original loan to value, original interest rate, property state, loan purpose, and loan term. Before carrying on with our analysis we omitted any of the NA values for each variable that were specified in the user guide.

Figure 1: Correlation Matrix

Chart, treemap chart

Description automatically generated

## **Univariate Analysis:**

For this section of our analysis, we will be analyzing the distribution and descriptive statistics for each variable in our analysis. Firstly, we will be looking into the distribution for each variable by creating histograms for numeric variables and bar charts and relative frequency tables for categorical variables.

Table 1: Descriptive Stats

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **vars** | **n** | **mean** | **sd** | **median** |
| Credit Score | 1332826 | 746.107 | 45.574 | 753 |
| Original Debt To Income | 1332826 | 35.135 | 9.308 | 36 |
| Original Interest Rate | 1332826 | 4.190 | 0.439 | 4.18 |
| Original Loan Term | 1332826 | 326.400 | 68.784 | 360 |

Figure 2: Credit Scores Histogram

Chart, histogram

Description automatically generated

In figure 2 we can see the distribution of credit scores in our dataset. It appears that credit scores are strongly skewed left in our data, with most of the data falling between credit scores of 750 and 800. Another interesting observation is that there are some credit scores below 600 but very few. This tells us that Freddie Mac usually buys most of their mortgages with credit scores over 700, and they do purchase mortgages with credit scores under 600 but very rarely.

Figure 3: Debt to Income Histogram

Chart, histogram

Description automatically generated

In figure 3 we can observe the distribution of debt to income which seems skewed left. Most of our data falls around the 40% mark for debt to income. This seems a bit odd considering that most of our credit score data fell around values that are considered very good credit scores, but our debt-to-income distribution is also skewed left with most values in the higher range. However, as mentioned earlier credit score doesn’t have a strong correlation with any variables in our data, and there is actually a slight negative correlation between the two variables.

Figure 4: Interest Rate Histogram

Chart, histogram

Description automatically generated

When observing figure 4 it seems like interest rates may be normally distributed. To test this we calculated the kurtosis and found that it was 3.50 and calculated the skewness which was -0.072. This seems to suggest that interest rates distribution is leptokurtic and is very slightly skewed left. To find out if these numbers were significant, we did a Jaque-Bera test for normality. The test outputted a p-value of 0.00000000000000022 so we will reject the null hypothesis that interest rates is normally distributed and conclude that the distribution is in fact leptokurtic.

Figure 5: Loan Term Histogram

Chart

Description automatically generated

This graph shows that the overwhelming majority of loans are 30-year loans, with 180-month (15-year) loans being a viable second option. This makes sense considering the fact that the generic mortgage term is 30 years. Although we can’t see it in figure 5, the maximum loan term value is 504 months. This is a very unreasonable value so it may be possible that this is a typo in the data.

|  |  |
| --- | --- |
| No | Yes |
|  |  |
| 77.47% | 22.53% |

Table 2: First Time Homebuyer Flag Relative Frequency Table

It appears that a large portion of the first-time homebuyer data consists of homebuyers who have bought homes before (table 2). Regardless, this variable could be used to show a significant difference in interest rates between first time buyers and non-first-time buyers when used as a dummy variable.

Figure 6: Property State Horizontal Bar chart

Chart, bar chart

Description automatically generated

Figure 6 displays that an overwhelming amount of the data comes from California, Texas and Florida, which makes sense since those three states have the highest populations in the US ([The 50 US States Ranked By Population - WorldAtlas](https://www.worldatlas.com/articles/us-states-by-population.html)). We did notice that there seems to be a disproportionate amount of data for mortgages in New York when compared to the New York population. New York is the 4 largest state by population but is the 9th state in our data when ranked by number of observations. It also appears that there is none or close to no data for the US territories in the dataset. We will be removing the territories from our data for the heatmap.

Figure 7: Loan Purpose Barchart

Chart, bar chart

Description automatically generated

The x labels for figure 7 are respectively abbreviations for the following: Refinance cash out, Refinance no cash out, and Purchase. We can gather from figure 7 that a large portion of the loans in our data are for purchase, while there are smaller amount of refinance loans with cash out and refinance loans with no cash out.

# **Bivariate Analysis:**

Note:

1. We will be using a significance level of 5% for the following analyses
2. We will be using R-Squared/Multiple R-Squared to measure how well a variable(s) explains variation in the interest rate variable or any other dependent variable

Figure 8: Credit Score scatterplot

Chart

Description automatically generated

Our regression can be visualized as a scatterplot as shown in figure 8. Obviously not all of our data points can be plotted since we have over a million points so we took a sample of 1000 observations to make this plot and all other scatterplots for this analysis. We can see the fall in interest rates as credit score increases as mentioned in table 5. We also notice that there are only two observations with credit scores lower than 600 in our sample of 1000. What is interesting is that there are streaks of similar interest rates for mortgage holders with very different credit scores. We will try to see if we can uncover any more information from the data by separating credit scores into categories based on FICO credit ratings, however this could just be due to interest rates being rounded to two decimal places.

Table 3: Credit Score simple regression

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Coefficients: | |  |  |  |  |  |  |  |
| Estimate Std.Error t value Pr(>|t|) | | | | | | | | |
| (Intercept) 6.094 0.005964549 1021.9 <0.0000000000000002 | | | | | | | | |
| Credit\_Score -0.002563259 0.000007979 -321.3 <0.0000000000000002 | | | | | | | | |
|  |  |  |  |  |  |  |  |  |
| Multiple R-squared: 0.07187 | | |  |  |  |  |  |  |

There is a statistically significant negative relation between credit scores and interest rates due to a p-value of less than .05. We can reject the null hypothesis that the beta for credit score (beta1) is zero and conclude that we can expect an average decrease of 0.00002563 in interest rate for every one unit increase in credit score.

Table 4: First Time Homebuyer simple regression

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 4.175210 0.002266 1842.9 <2e-16 \*\*\***

**First\_Time\_HomebuyerY 0.064193 0.004791 13.4 <2e-16 \*\*\***

**Multiple R-squared: 0.00371, Adjusted R-squared: 0.003689**

We can conclude that the dummy variable for first time homebuyer flag is significant when fitted to original interest rate. This dummy variable can be interpreted as: the average expected difference in interest rate for a first-time homebuyer compared to a non-first-time homebuyer.

Figure 9: Credit Score Ratings boxplot

Chart, box and whisker chart

Description automatically generated

The boxplot that you see in figure 9 is the work of creating categorical data from a sample of the credit score ratings. The code classified an excellent score as a range from 800 to 850, a very good score was 740 to 800, a good score was 670 to 740, fair was a score between 580 and 670, and poor was 300 to 560. As mentioned earlier there are no poor credit scores since the NA values from debt to income got ride of most of the data below a credit score of 600. From observing figure 9 we see that there is almost no change in average interest rate between poor and fair credit scores but much more significant changes between fiar and good and good and very good.

Table 5: Debt to Income simple regression

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 3.973356 0.007743 513.18 <2e-16 \*\*\***

**Original\_Debt\_To\_Income 0.006154 0.000213 28.89 <2e-16 \*\*\***

**Multiple R-squared: 0.01701, Adjusted R-squared: 0.01699**

From table 5 we conclude that there is a significant positive relation between debt to income and interest rate (p-value > .05). This is interpreted as: a one unit increase in debt to income yields a 0.000061 increase in interest rate on average. Notice that the adjusted R-squared is significantly lower than the credit score regression in table 5, so it seems that credit score may be a better predictor of interest rates than debt to income.

Figure 10: Loan Purpose boxplot

Chart, box and whisker chart

Description automatically generated

Figure 10 validates our conclusions found in table 8 by visualizing the distribution and median of interest rates for each loan purpose category. It appears that mortgage holders who are refinancing with no cash out (N) have the lowest interest rates on average when compared to loans for purchases and loans for refinances with cash out.

Table 8: Loan Purpose simple regression

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 4.236934 0.004030 1051.32 < 2e-16 \*\*\***

**Loan\_PurposeN -0.227259 0.006255 -36.34 < 2e-16 \*\*\***

**Loan\_PurposeP -0.015166 0.004770 -3.18 0.00148 \*\***

**Multiple R-squared: 0.03429, Adjusted R-squared: 0.03425**

Featured above in table 8 is the summary of a simple linear regression of loan purpose on interest rate. The reference group for this dummy variable is Refinance – Cash out (C). We can infer that both of our dummy variables are significant since the p-value is lower than our significance level. We can conclude that on average we can expect for a -0.227 decrease in interest rates for a non-cash out refinancing loan when compared to a cash out refinance loan. We can also conclude that on average we can expect interest rates to be -0.015 lower for a purchase loan when compared to a cash out refinance loan.

Table 9: Loan Term simple regression

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 2.9526228 0.0078052 378.3 <2e-16 \*\*\***

**Original\_Loan\_Term 0.0037896 0.0000234 162.0 <2e-16 \*\*\***

**Multiple R-squared: 0.3523, Adjusted R-squared: 0.3523**

In table 9 we can come to the conclusion that for every one month increase in loan term we can expect a 0.0000379 increase in interest rate on average. This is in fact a significant relationship since our p-value is below .05. It is also important to note that the loan term regression has a much higher multiple R-squared value compared to any of our other variables.

Figure 12: US interest rate heat map

Map

Description automatically generated

From the heat map that we created based on the 2017 historical data(figure 12), we observed that Nevada has the highest average and North Dakota the lowest. We were more interested in this discovery and did some research to find that [Nevada](https://smartasset.com/mortgage/nevada-mortgage-rates) interest rates were .12% higher and [North Dakota](https://smartasset.com/mortgage/north-dakota-mortgage-rates) was .12% lower in comparison to the United States average (4.03%) in the year 2017, according to [smartasset.com](https://smartasset.com/) data.

# **multivariate Analysis:**

Figure 13: Credit Score vs interest rate by Credit rating scatterplot

Chart, scatter chart

Description automatically generated

This scatterplot illustrates the relationship between credit score and credit score ratings. Tthe average expected change in interest rate for one unit change in credit score may vary depending on credit score rating. We decided to test this with a regression with an effect modifier found in table 11.

Table 12: Multiple Linear Regression

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 4.586e+00 2.725e-02 168.29 <2e-16**

**Credit\_Score -2.242e-03 3.352e-05 -66.90 <2e-16**

**Original\_Loan\_To\_Value 2.101e-03 1.035e-04 20.30 <2e-16**

**Original\_Loan\_Term 3.842e-03 2.337e-05 164.38 <2e-16**

**Loan\_PurposeN -1.767e-01 4.773e-03 -37.02 <2e-16**

**Loan\_PurposeP -1.589e-01 4.289e-03 -37.05 <2e-16**

**First\_Time\_HomebuyerY -5.282e-02 4.089e-03 -12.92 <2e-16**

**Multiple R-squared: 0.4447, Adjusted R-squared: 0.4446**

Lastly, we have fit all our variables into a multiple linear regression as shown in table 12. We can conclude that all of our variables are significant when fitted to interest rate including our dummy variables for loan purpose and first-time homebuyer flag. For this multiple linear regression our adjusted R-squared was 0.4446, which seems like a great improvement compared to the adjusted R-squared values for the simple linear regressions. To make sure that there are no outliers in our multiple linear regression we did a Cooke’s distance test to test for influential points and somewhat influential points and found nothing that stood out. Therefore, we will retain all our data. We also plotted a histogram of the residuals to make sure the residuals are normally distributed and as shown in figure 14 below they are approximately normally distributed.

Figure 14: Multiple Linear Regression residual histogram

Chart, histogram

Description automatically generated

After performing a Jaque-Bera test we found that the residuals of the multiple regression are actually not normally distributed since they are skewed right. Therefore, our model needs to be tweaked by using a log function or quadratic to fix our models skewness.

# **Conclusion**

It seems that in the end our multiple regression model didn’t end up passing the test for normality of residuals. However, we did learn a lot about our chosen variables along the way in how they are distributed and how they correlate to interest rates.