Analysis of Rate spread of mortgage applications

James Kamau Waiharo, November 2019

Executive Summary

This document presents an analysis of data concerning mortgage applications and their respective rate-spread. This analysis is based on 200,000 mortgage applications, each containing property, loan, applicant, and census information associated with a mortgage application and the rate spread of the mortgage rate offered. The objective of the analysis is to predict the rate spread of mortgage applications.

After exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the data, several potential relationships between mortgage application features and rate spread were identified. After exploring the data, a predictive model to predict a mortgage applications rate spread from these features was created.

After performing the analysis, the following are the conclusions:

While many factors can help indicate the rate spread of the mortgage application, significant features found in this analysis were:

- **Loan type** - Indicates whether the loan granted, applied for, or purchased was conventional, government-guaranteed, or government-insured
- **Property type** - Indicates whether the loan or application was for a one-to-four-family dwelling (other than manufactured housing), manufactured housing, or multifamily dwelling
- **Loan purpose** - Indicates whether the purpose of the loan or application was for home purchase, home improvement, or refinancing
- **Loan amount** - Size of the requested loan in thousands of dollars

Initial Data Exploration

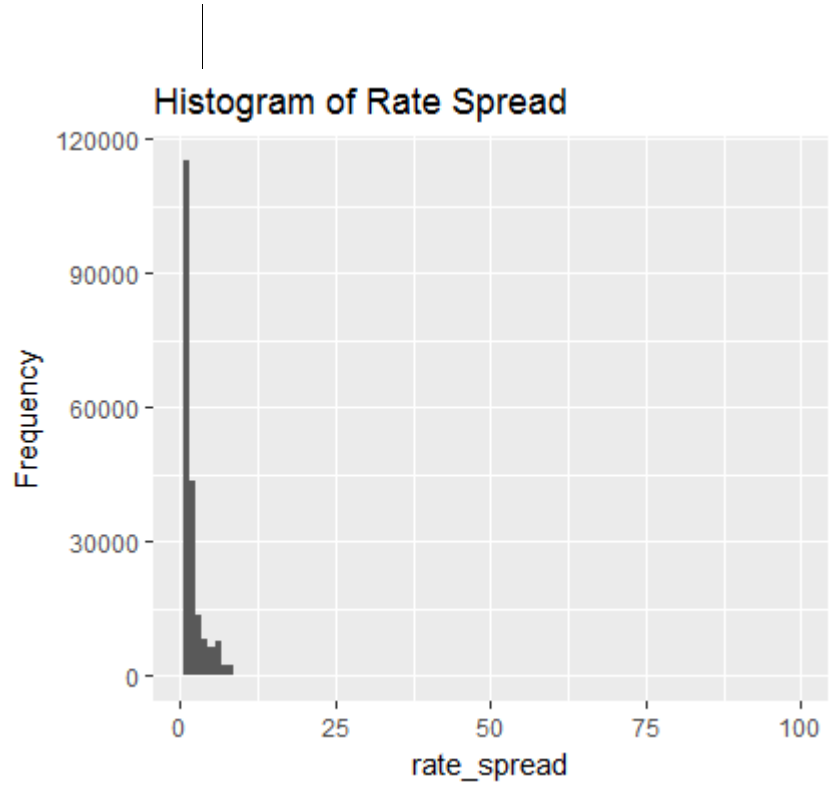
The initial exploration of the data began a look at the data, the actual data types of columns and some summary and descriptive statistics.

Individual Feature Statistics

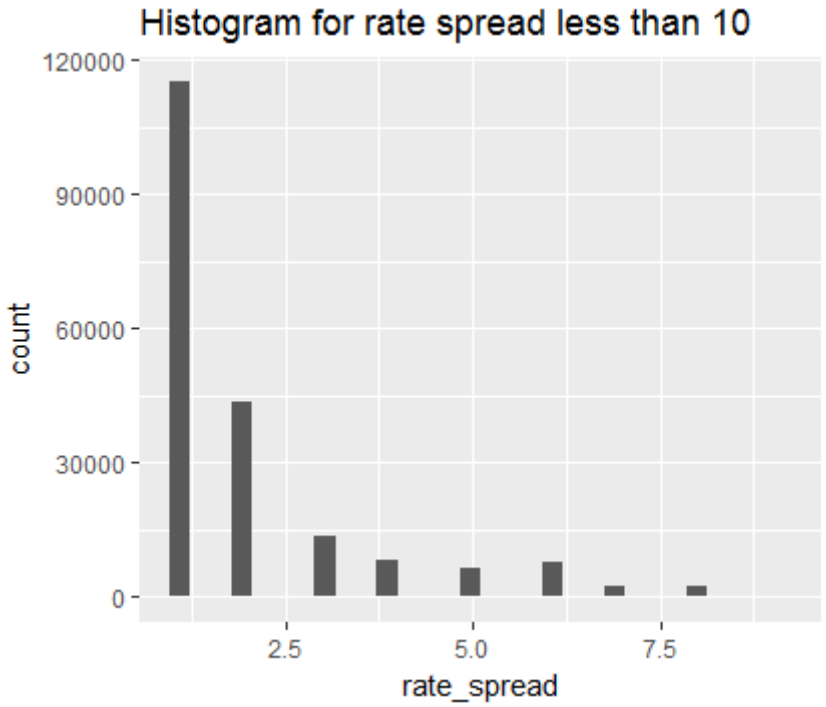
Summary statistics for minimum, maximum, mean, median, standard deviation, and distinct count were calculated for numeric columns, and the results taken from 200,000 observations are shown here:

Column	Min	Max	Median	Mean	Std. Dev	Count
Loan amount	1	11104	116	142.57	142.56	200000
population	7	34126	4959	5391.1	2669.03	198005
Minority population %	0.33	100	26	34.24	27.93	198005
Ffiecmedian family income	17860	125095	63485	64595.36	12724.51	198015
tract_to_msa_md_income_pct	6.19	100	98.96	89.28	15.06	197977
number_of_owner.occupied_units	3	8747	1304	1402.87	706.88	197988
number_of_1_to_4_family_units	6	13615	1799	1927.34	886.58	197984
rate_spread	1	99	1	1.98	1.66	200000

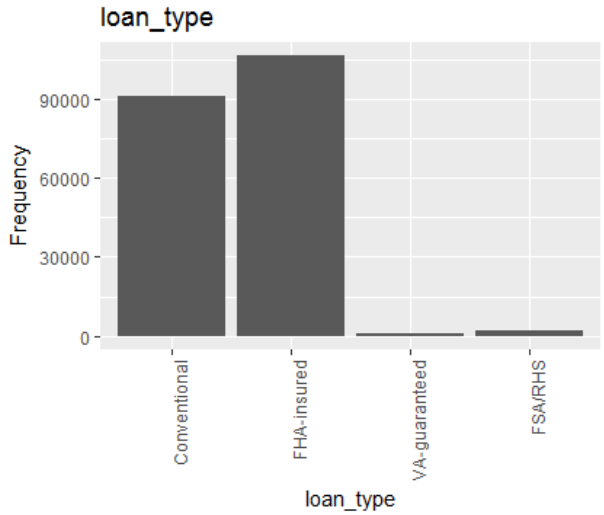
Since **rate_spread** is of interest in this analysis, it was noted that the difference between the median and the max i.e. 0 and the difference between the median and the max 98 is significantly different and that the comparatively small standard deviation indicates that there are extreme outliers in the **rate spread** column. A histogram of the **rate spread** column shows that the values are right-skewed with most of them being less 10

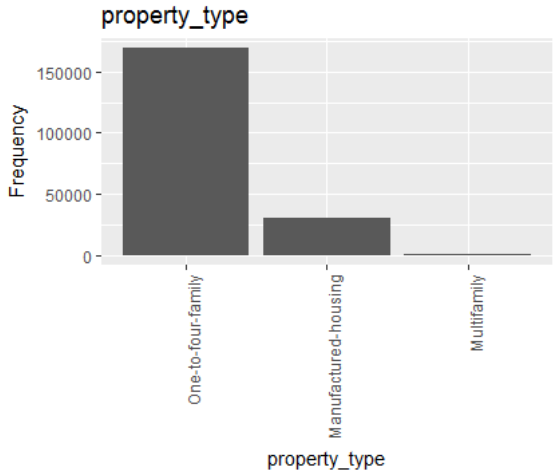


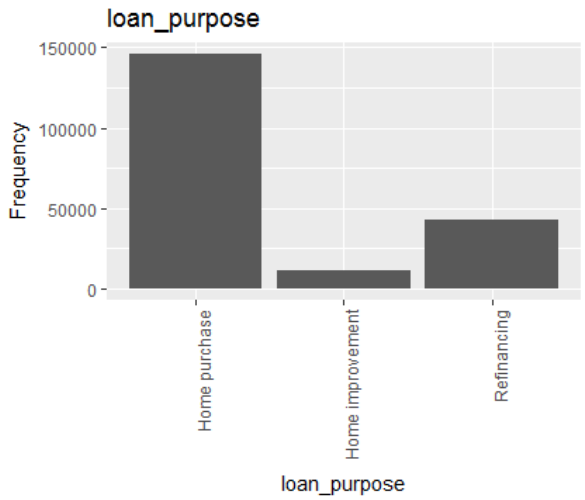
After removing outliers below is a histogram of the **rate spread** values less 10 show that rate spread is actually a discrete variable with only 0.1% of the observation with **rate spread** values greater than 8

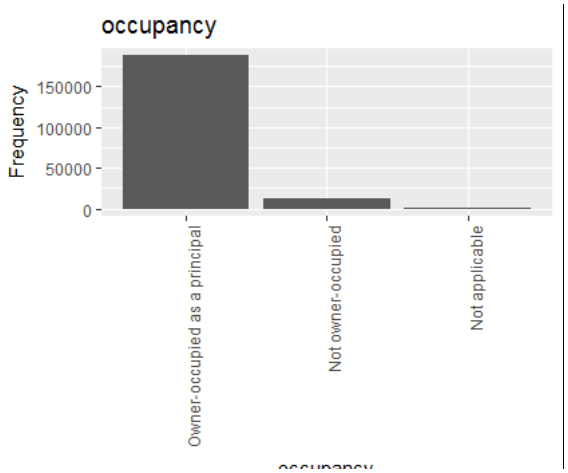


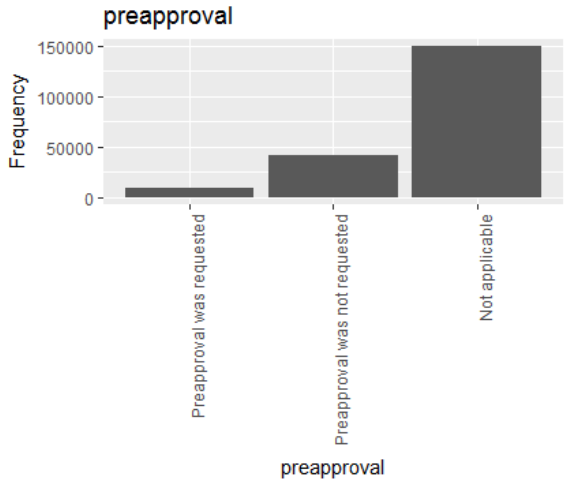
In addition to the numeric values, the mortgage application observations include categorical features, most of the categorical variables were encoded as integers. They were recoded showing their appropriate labels then bar charts created to show frequency of each:

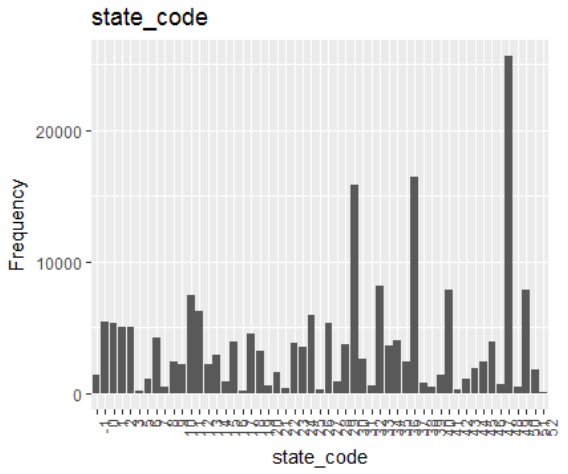


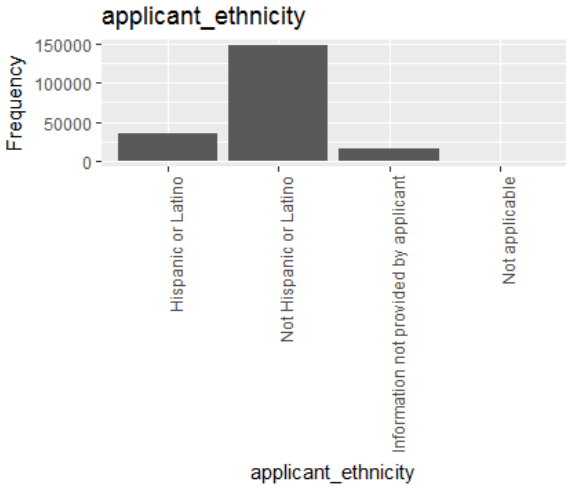


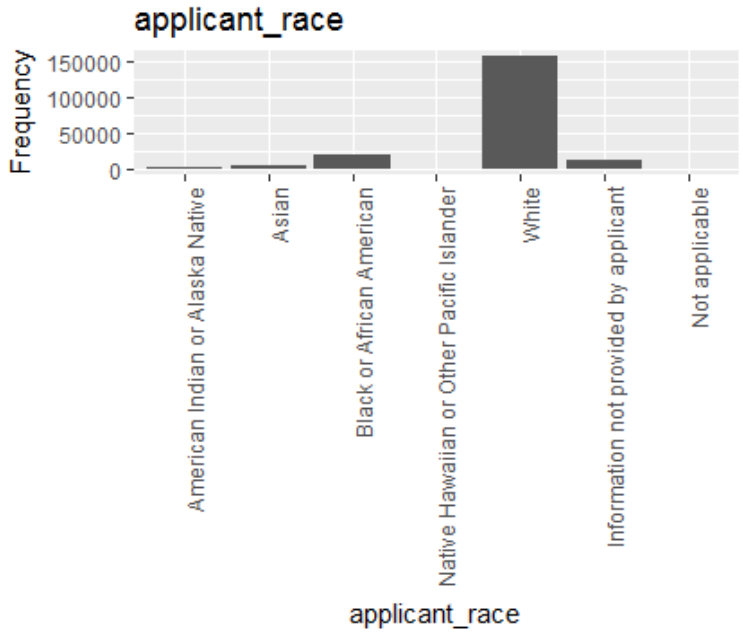


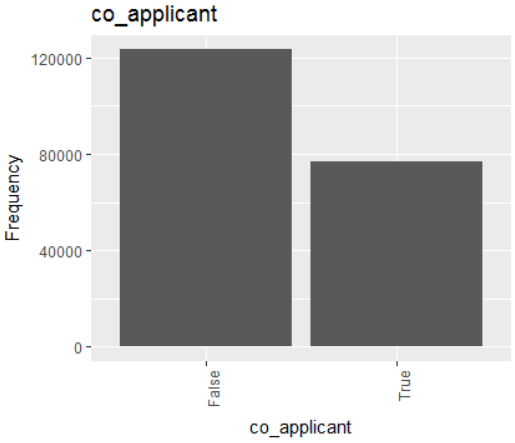

occupancy
Frequency
150000
100000
50000
0
Owner-occupied as a principal
Not owner-occupied
Not applicable

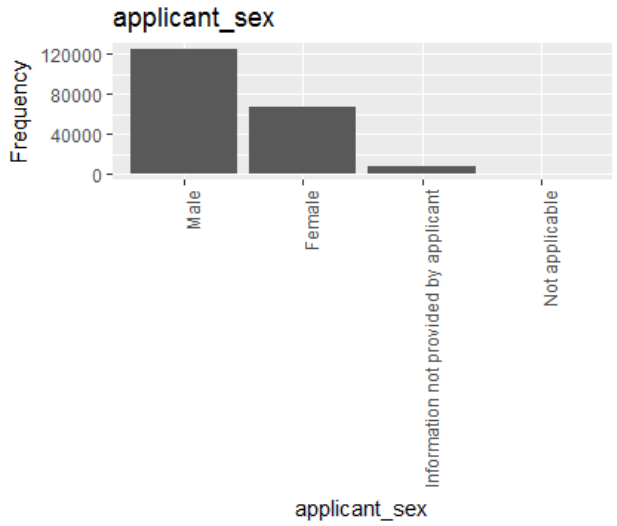

preapproval
Frequency
150000
100000
50000
0
Preapproval was requested
Preapproval was not requested
Not applicable
preapproval


state_code
Frequency
20000
10000
0
state_code


applicant_ethnicity
Frequency
150000
100000
50000
0
Hispanic or Latino
Not Hispanic or Latino
Information not provided by applicant
Not applicable
applicant_ethnicity


applicant_race
Frequency
150000
100000
50000
0
American Indian or Alaska Native
Asian
Black or African American
Native Hawaiian or Other Pacific Islander
White
Information not provided by applicant
Not applicable
applicant_race


co_applicant
Frequency
120000
80000
40000
0
False
True
co_applicant



Other categorical features that had too many members to visualize were:

- **msa_md** - A categorical with no ordering indicating Metropolitan Statistical Area/Metropolitan Division
- **county_code** - A categorical with no ordering indicating the county
- **Lender** - A categorical with no ordering indicating which of the lenders was the authority in approving or denying this loan

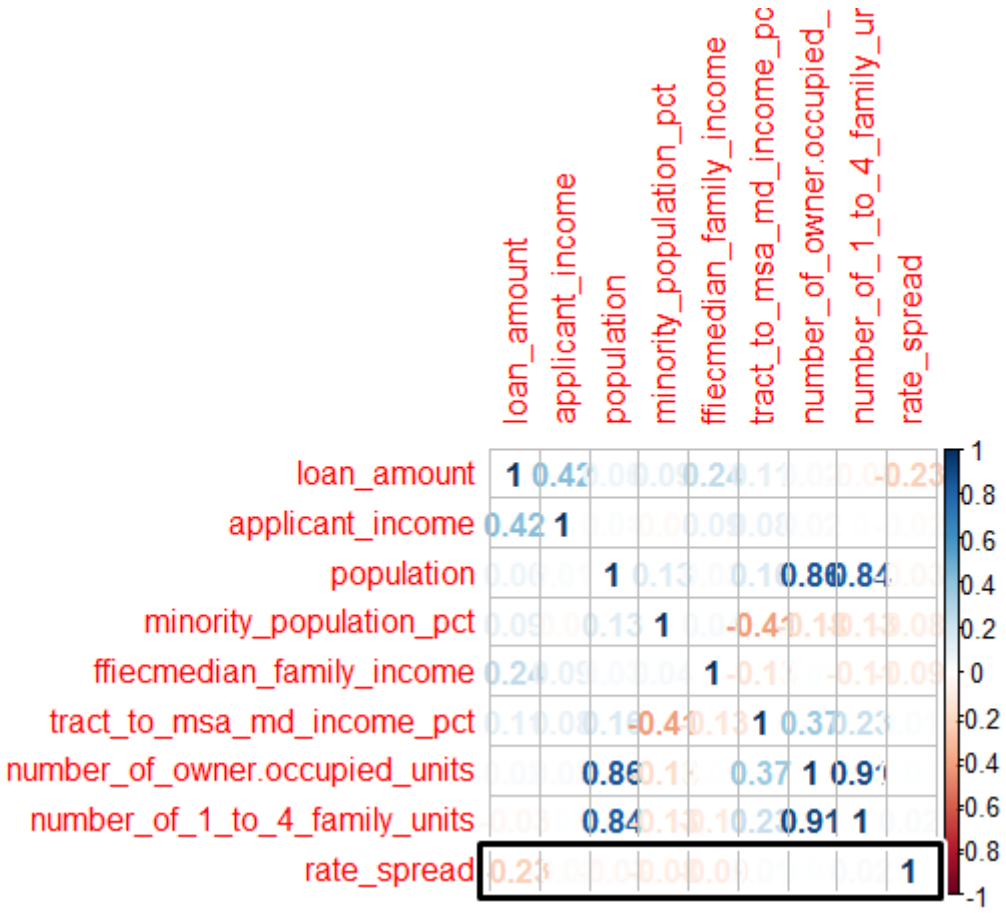
Correlation and Apparent Relationships

After exploring the individual features, an attempt was made to identify relationships between features in the data – in particular, between **rate spread** and the other features.

Most columns had very few missing values, therefore observations with missing variables were removed.

Numeric Relationships

The following corrplot was generated initially to check whether the numeric features are correlated with one another.


loan_amount
applicant_income
population
minority_population_pct
ffiecmedian_family_income
tract_to_msa_md_income_pc
number_of_owner.occupied_
number_of_1_to_4_family_ur
rate_spread
1
0.42
0.24
-0.23
0.86
0.84
-0.41
0.37
0.91
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Viewing the plot in the bottom row or the right-most column of this matrix shows only an inverse relationship between Loan amount and rate spread though a weak one. The correlation between rate spread and the other numeric features is extremely weak almost nonexistent.

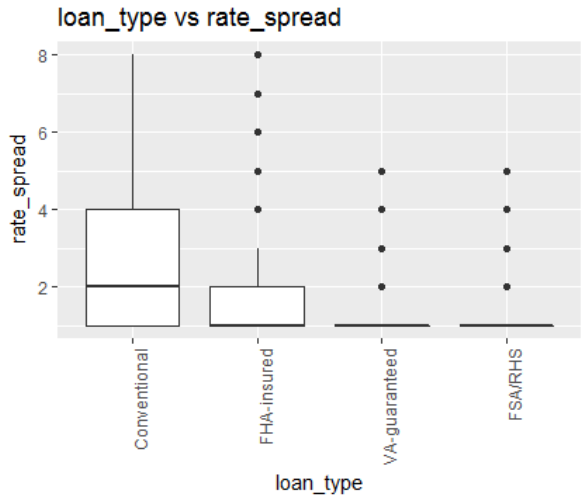
The correlation between the numeric columns and rate spread was calculated with the following results:

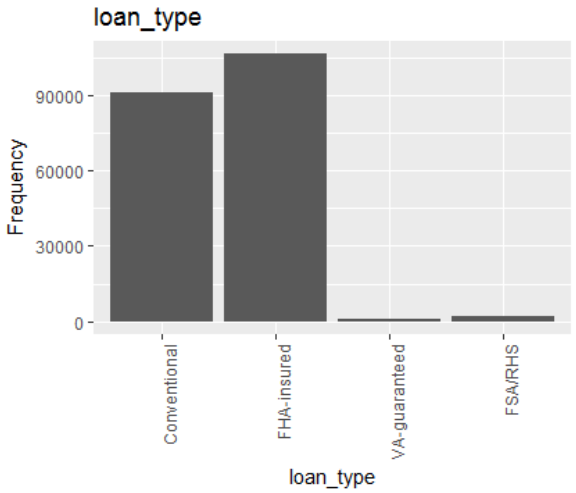
	rate_spread
loan amount	-0.23177632
applicant income	-0.01915714
population	-0.03478818
minority_population_pct	-0.08000118
ffiecmedian_family_income	-0.09238712
tract_to_msa_md_income_pct	0.01259316
number_of_owner.occupied_units	0.00627657
number_of_1_to_4_family_units	0.02397523
rate_spread	1

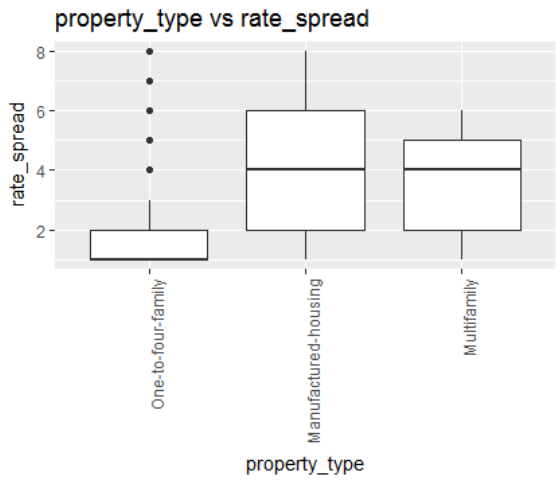
These correlations validate the plots by showing a negative correlation between Loan amount and rate spread and extremely weak correlations for the other numeric features.

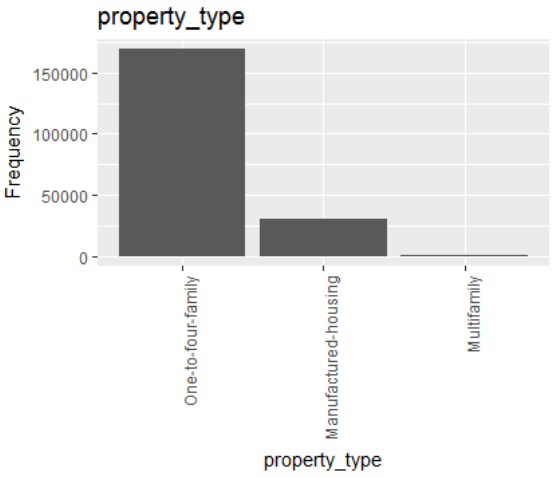
Categorical Relationships

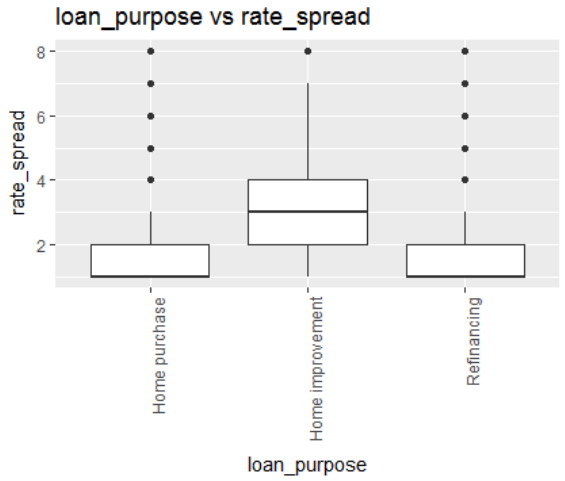
After exploring the relationship between categorical feature and rate spread, for most of the categorical features, the rate spread's median and range were similar for all the values of the categorical columns. The following box- plots and corresponding histogram show the categorical features that seem to exhibit a relationship with the rate spread:

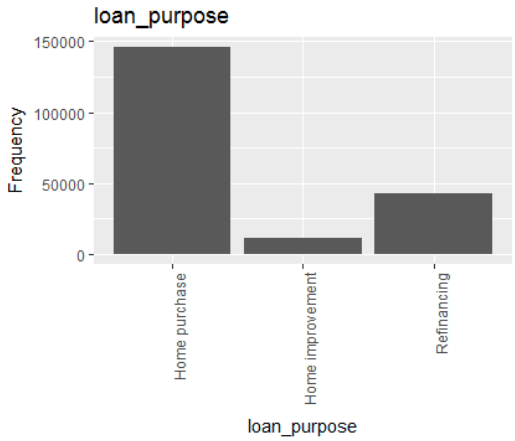






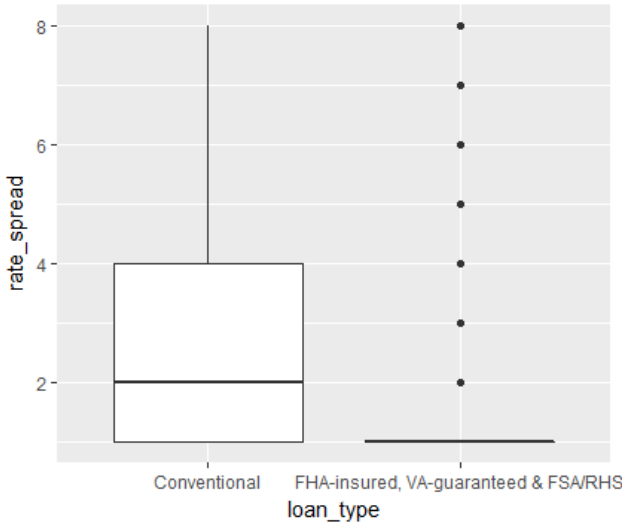


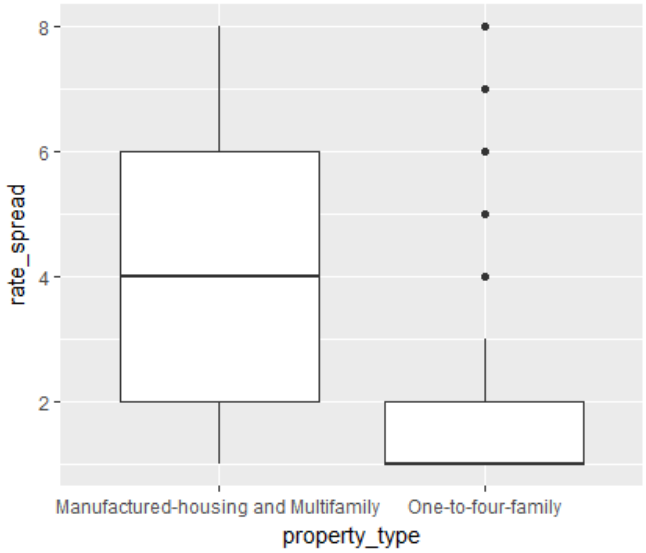


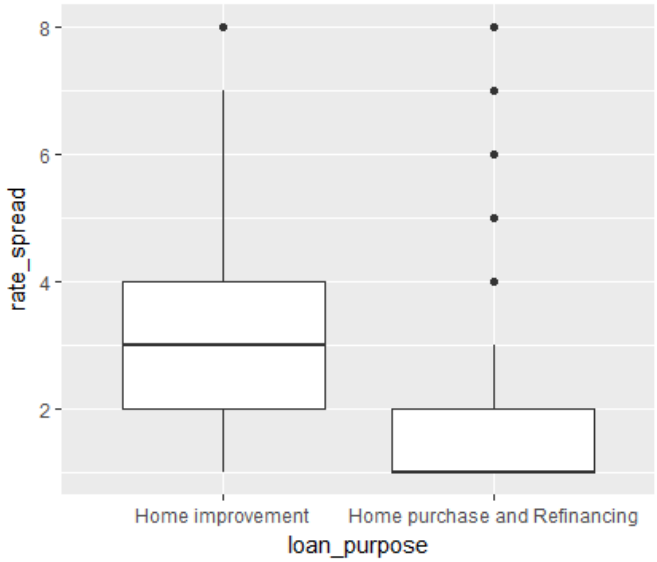


For the three features above, some values have low frequencies. The categorical values, similar in terms of the median and range of rate price values were combined into fewer categories.

This resulted in a smaller range of categories, as shown here:


8
6
4
2
rate_spread
Conventional
FHA-insured, VA-guaranteed & FSA/RHS
loan_type


8
6
4
2
rate_spread
Manufactured-housing and Multifamily
One-to-four-family
property_type


8
6
4
2
rate_spread
Home improvement
Home purchase and Refinancing
loan_purpose

The box plots show some clear differences in terms of the median and range of rate spread values for different categorical feature values.

Regression

Based on the analysis of the Mortgage application data, a predictive model to predict the actual rate spread of a mortgage application was created. Based on the apparent relationships identified when analyzing the data, a regression model was created to predict the value for rate spread.

The model was trained with 80% of the data. Testing the model with the remaining 20% resulted to an R squared coefficient of 0.498

Conclusion

This analysis has shown that around 50% of the variance in the rate spread of a mortgage application can be explained by the loan type, loan purpose, property type and loan amount. Therefore, to achieve our goal of accurately predicting the rate spread of a mortgage application, a further analysis of the features provided, some feature engineering and experimenting with more models will be necessary.