Analysis of Mortgage Rates from Government Data

Zarko Rashev, November 2019

# Executive Summary

This document represents an analysis of data concerning mortgage rates. The analysis is based on 200 000 (two hundred thousand) observations of mortgage data containing specific characteristics such as applicant, propery, loan and census information.

After exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the data, several potential relationships between applicant, location and property characteristics and rate spread were identified. After exploring the data, a regression model to predict a rate spread from its features was created.

After performing the analysis, the author presents the following conclusions:

More significant features found in this analysis were:

**loan\_amount** - Size of the requested loan in thousands of dollars

**minority\_population\_pct** - Percentage of minority population to total population for tract

**ffiecmedian\_family\_income** - FFIEC Median family income in dollars for the MSA/MD in which the tract is located (adjusted annually by FFIEC)

**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

**applicant\_ethnicity** - Ethnicity of the applicant

# Initial Data Exploration

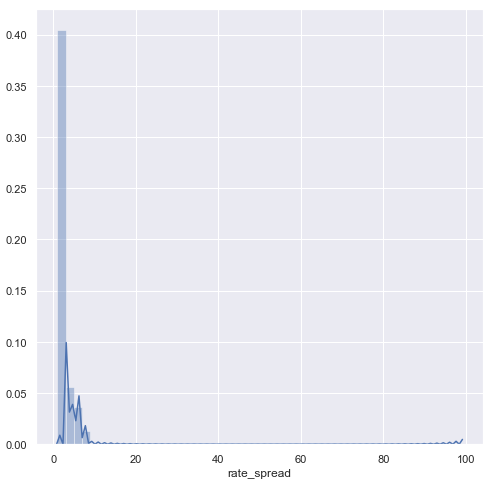
The initial exploration of the data began with some summary and descriptive statistics.

## 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:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Numeric Feature** | **min** | **max** | **mean** | **median** | **std** | **count** |
| **loan\_amount** | 1,00 | 11104 | 142.574940 | 116.000 | 142.559487 | 200000 |
| **applicant\_income** | 1,00 | 10042 | 73.617902 | 56.000 | 105.696934 | 189292 |
| **population** | 7,00 | 34126 | 5391.099099 | 4959.000 | 2669.028807 | 198005 |
| **minority\_population\_pct** | 0.326 | 100 | 34.238640 | 25.996 | 27.930882 | 198005 |
| **ffiecmedian\_family\_income** | 17860,00 | 125095 | 64595.355801 | 63485.000 | 12724.514485 | 198015 |
| **tract\_to\_msa\_md\_income\_pct** | 6.193 | 100 | 89.283022 | 98.959 | 15.059223 | 197977 |
| **number\_of\_owner-occupied\_units** | 3,00 | 8747 | 1402.872401 | 1304.000 | 706.880410 | 197988 |
| **number\_of\_1\_to\_4\_family\_units** | 6,00 | 13615 | 1927.336618 | 1799.000 | 886.576586 | 197984 |
| **rate\_spread** | 1,00 | 99 | 1.979110 | 1.000 | 1.656809 | 200000 |

Since **rate\_spread** is of interest in this analysis, it was noted that the mean and median of this value are significantly different and that the comparatively large standard deviation indicates that there is considerable variance in the rate\_spread of the mortages. A histogram of the **rate\_spread** column shows that the **rate\_spread** values are right-skewed – in other words, most mortgages are rated at the lower end of the rate\_spread range, as shown here:



In addition to the numeric values, the mortgage observations include categorical features, including:

**'msa\_md' -** indicating Metropolitan Statistical Area/Metropolitan, Division where a value of -1 indicates a missing value

**'state\_code' -** indicating the U.S. state where a value of -1 indicates a missing value

**'county\_code' -** indicating the county where a value of -1 indicates a missing value

**'lender' -** indicating which of the lenders was the authority in approving or denying this loan

**'loan\_type' -** Indicates whether the loan granted, applied for, or purchased was conventional, government-guaranteed, or government-insured; available values are:

1 -- Conventional (any loan other than FHA, VA, FSA, or RHS loans)

2 -- FHA-insured (Federal Housing Administration)

3 -- VA-guaranteed (Veterans Administration)

4 -- FSA/RHS (Farm Service Agency or Rural Housing Service)

**'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; available values are:

1 -- One to four-family (other than manufactured housing)

2 -- Manufactured housing

3 -- Multifamily

**'loan\_purpose' -** Indicates whether the purpose of the loan or application was for home purchase, home improvement, or refinancing; available values are:

1 -- Home purchase

2 -- Home improvement

3 -- Refinancing

**'occupancy' -** Indicates whether the property to which the loan application relates will be the owner's principal dwelling; available values are:

1 -- Owner-occupied as a principal dwelling

2 -- Not owner-occupied

3 -- Not applicable

**'preapproval' -** Indicate whether the application or loan involved a request for a pre-approval of a home purchase loan; available values are:

1 -- Preapproval was requested

2 -- Preapproval was not requested

3 -- Not applicable

**'applicant\_ethnicity' -** Ethnicity of the applicant; available values are:

1 -- Hispanic or Latino

2 -- Not Hispanic or Latino

3 -- Information not provided by applicant in mail, Internet, or telephone pplication

4 -- Not applicable

5 -- No co-applicant

**'applicant\_race' -** Race of the applicant; available values are:

1 -- American Indian or Alaska Native

2 -- Asian

3 -- Black or African American

4 -- Native Hawaiian or Other Pacific Islander

5 -- White

6 -- Information not provided by applicant in mail, Internet, or telephone application

7 -- Not applicable

8 -- No co-applicant

**'applicant\_sex' -** Sex of the applicant; available values are:

1 -- Male

2 -- Female

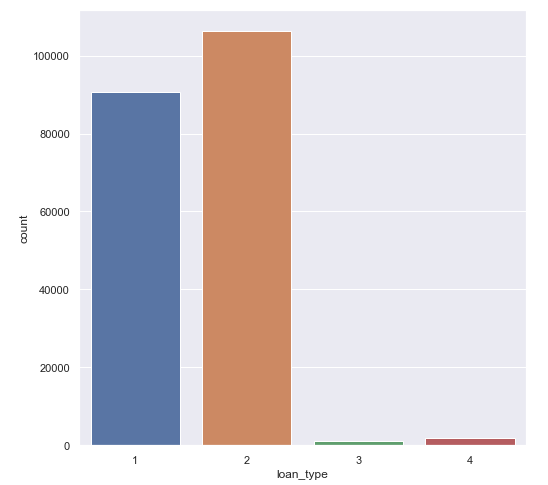
3 -- Information not provided by applicant in mail, Internet, or telephone application

4 or 5 -- Not applicable

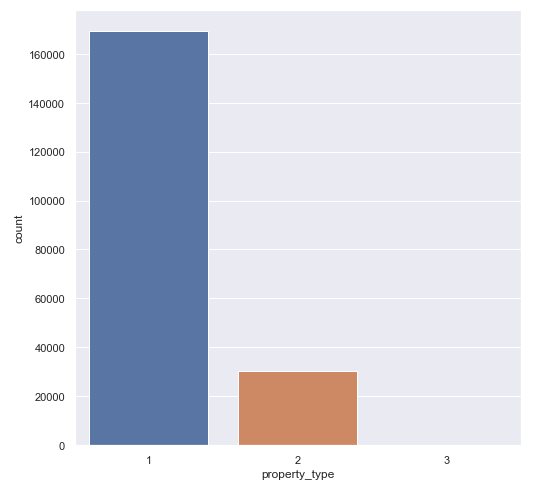
**'co\_applicant' -** Indicates whether there is a co-applicant (often a spouse) or not

Bar charts were created to show frequency of these features, and indicate the following:

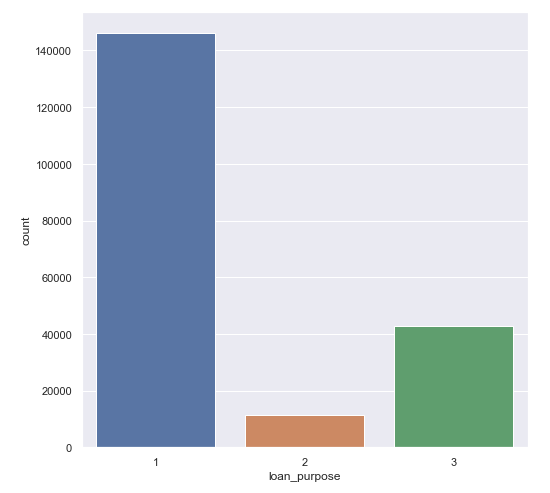
(graph first, conclusion follows second)



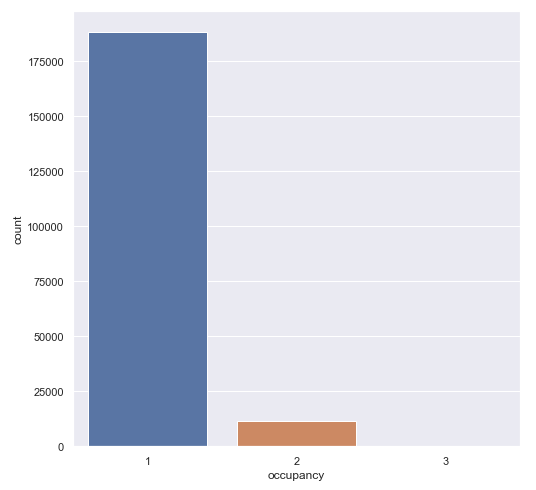
Most of the **loan types** are type 1 and 2 respectively Conventional (any loan other than FHA, VA, FSA, or RHS loans) and FHA-insured (Federal Housing Administration) with type 2 leading.



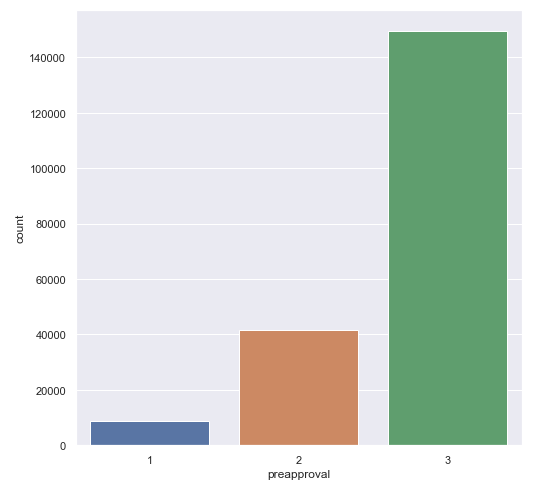
Most of the **property types** are of type 1 respectively One to four-family (other than manufactured housing) with significantly smaller amout of type 2 Manufactured housing and almost none Multifamily.



More of 140 000 of the loans had the **purpose** to purchase a home – 1 – Home purchase. Second most popular loan purpose was refinancing – 3 Refinancing, and least was to improve their home – 2 – Home improvement.

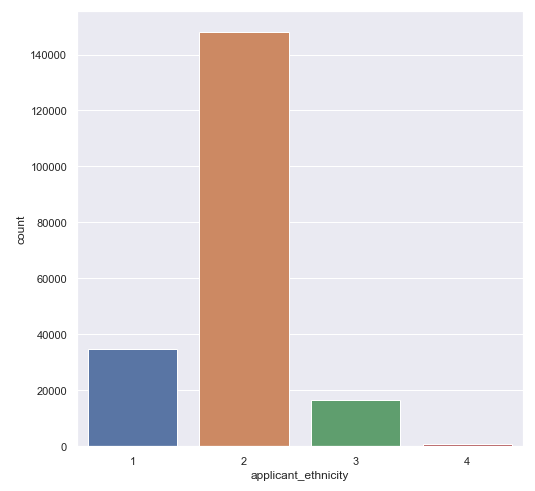


Most of the poperties financed by the mortgage will be occupied by the owner – 1 Owner-occupied as a principal dwelling. A very small amount of properties will not be occupied by the owner – 2 Now owner-occupied. For an insignificant part of the observations, the occupancy categorization is not applicable – 3 Not applicable.



For a greater part of the loans, preapproval was not applicable – 3 Not Applicable. For about 20% of all observations, preapproval was not requested – 2 Preapproval was not requested.

Preaproval was requested for a very small amount of observations, less than 5% - 1 Preapproval was requested.

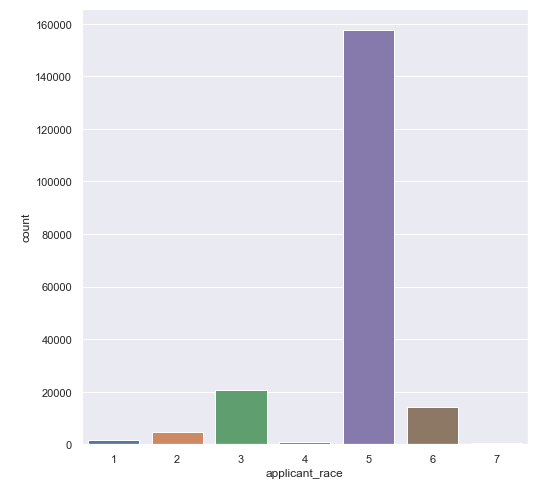


More than 70% of all applicants were of non-hispanic/non-latino ethnicity – 2 Not Hispanic or Latino.

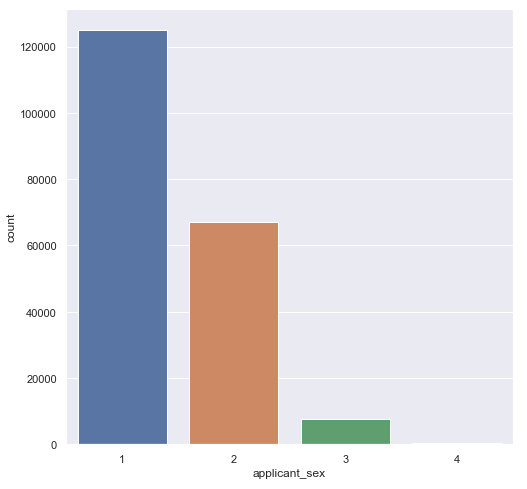
Second most popular ethnicity group was Hispanic/Latino – 1 Hispanic or Latino.

About 10% of applicants did not provide ethnicity information – 3 Information not provided

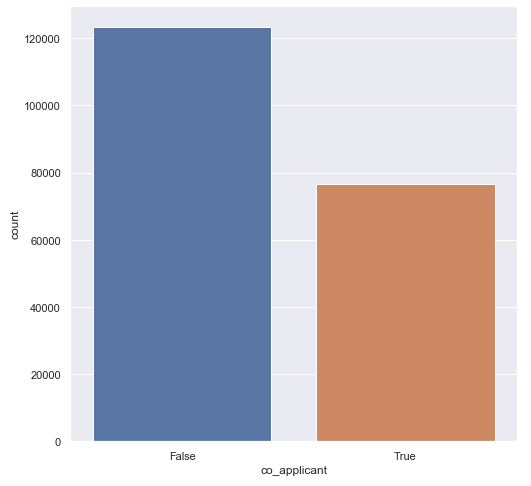
For an insignificant part of the applicants, ethnicity was not applicable – 4 Not applicable



Majority of applicants, about 80% were White – 5 White. Roughly 10% were of Black/African American – 3 Black or African American. About 7% did not provide information – 6 Information not provided. About 2% were observed to be Asian – 2 Asian. Rest of the observations are insignificant <=1%.



More than 60% of the applicants were male – 1 Male. About 30% were female – 2 Female. For very few observations, 3-4% informatio on sex was not provided – 3 Information not provided. Applicants where sex was not applicable are insignificant – 4 Not applicable.



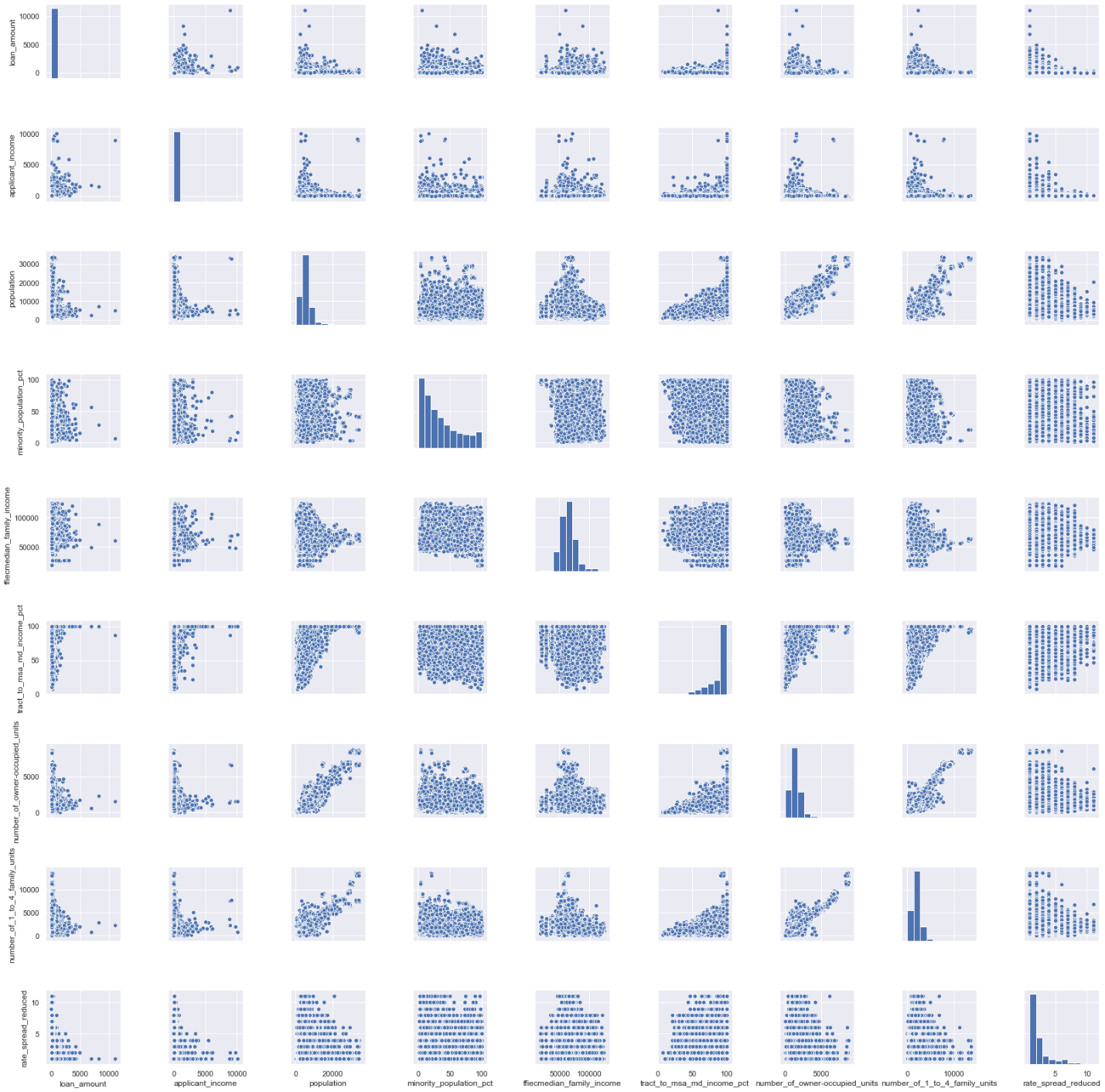
Majority of applicants had a co-applicant, more than 60% of all observations - False. Rest of the applicants had a co-applicant – True.

## 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.

## Numeric Relationships

The following scatter-plot matrix was generated initially to compare numeric features with one another. The key features in this matrix are shown here:



For better visualization, rate\_spread outliers (rate\_spread >12, totalling 58 observations) were dropped.

Missing values for rate\_spread were also dropped.

Viewing plots in the bottom row or the right-most column of this matrix shows some weak relationship but not at all apparent between **rate spread** and other numeric features. More apparent is the relationship with **loan\_amount** where rate spread increases as this feature decreases.

Less apparent relationship can be observed with **minority\_population\_pct** and **ffiecmedian\_family\_income.** Again, rate spread increasing as features decrease, but correlation is very weak, close to zero.

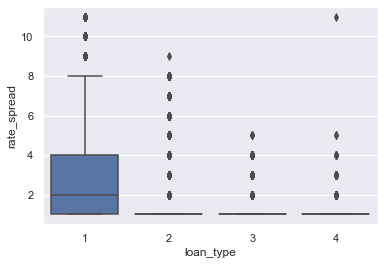
The correlation between the numeric columns was then calculated with the following results:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **loan\_amount** | **applicant\_income** | **population** | **minority\_population\_pct** | **ffiecmedian\_family\_income** | **tract\_to\_msa\_md\_income\_pct** | **number\_of\_owner-occupied\_units** | **number\_of\_1\_to\_4\_family\_units** | **rate\_spread** |
| **loan\_amount** | 1,00 | 0,45 | 0,07 | 0,10 | 0,25 | 0,11 | 0,03 | -0,02 | -0,26 |
| **applicant\_income** | 0,45 | 1,00 | 0,01 | -0,03 | 0,09 | 0,09 | 0,02 | 0,00 | -0,02 |
| **population** | 0,07 | 0,01 | 1,00 | 0,14 | 0,03 | 0,16 | 0,86 | 0,84 | -0,03 |
| **minority\_population\_pct** | 0,10 | -0,03 | 0,14 | 1,00 | 0,04 | -0,41 | -0,18 | -0,13 | -0,08 |
| **ffiecmedian\_family\_income** | 0,25 | 0,09 | 0,03 | 0,04 | 1,00 | -0,13 | 0,01 | -0,11 | -0,09 |
| **tract\_to\_msa\_md\_income\_pct** | 0,11 | 0,09 | 0,16 | -0,41 | -0,13 | 1,00 | 0,37 | 0,23 | 0,01 |
| **number\_of\_owner-occupied\_units** | 0,03 | 0,02 | 0,86 | -0,18 | 0,01 | 0,37 | 1,00 | 0,91 | 0,01 |
| **number\_of\_1\_to\_4\_family\_units** | -0,02 | 0,00 | 0,84 | -0,13 | -0,11 | 0,23 | 0,91 | 1,00 | 0,03 |
| **rate\_spread** | -0,26 | -0,02 | -0,03 | -0,08 | -0,09 | 0,01 | 0,01 | 0,03 | 1,00 |

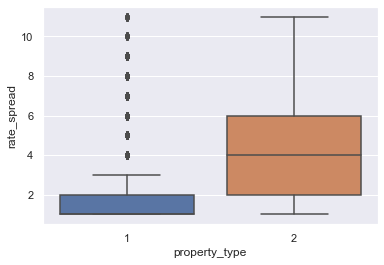
These correlations validate the plots by showing a weak-to-moderate negative correlation between **loan\_amount** and **rate\_spread**, and very weak negative correlations for **minority\_population\_pct** and **ffiecmedian\_family\_income**.

### Categorical Relationships

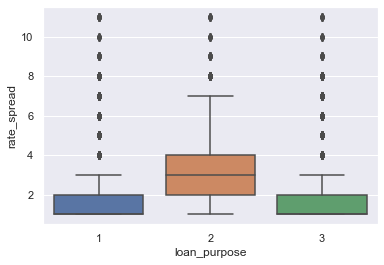
Having explored the relationship between rate\_spread and numeric features, an attempt was made to discern any apparent relationship between categorical feature values and rate\_spread. The following boxplots show the categorical columns and their relationship with the rate\_spread:



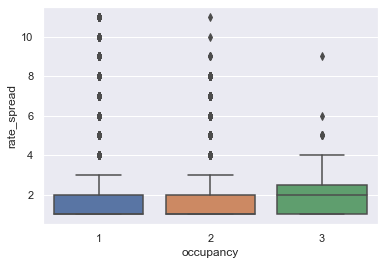
Loan type 1 - Conventional exhibits a wider range of rate\_spread.



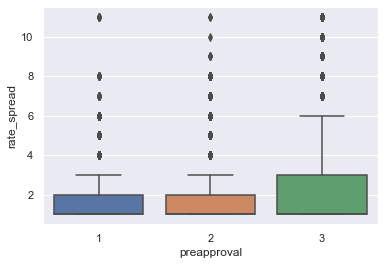
Property type 2 - Manufactured housing shows higher median value and wider range of rate\_spread



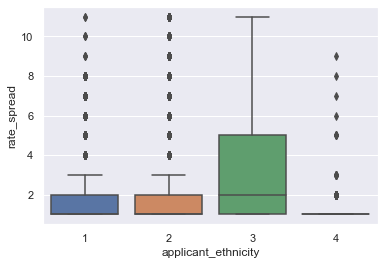
Loan for the purpose of 1 and 3 respectively Home purchase and Refinancing show similar rate spreads while range and median of loan purpose 2 – Home improvement exhibits higher range and median.



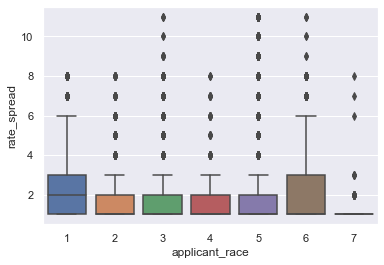
No significant difference between different occupancy categories with a little wider range and higher median for mortgages where occupancy was not applicable.



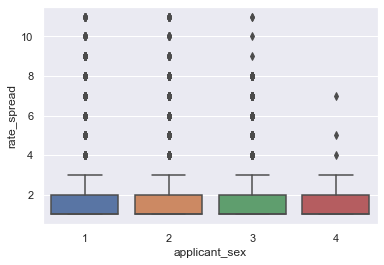
No significant difference between rate spread for mortgages where preapproval was requested and not requested.



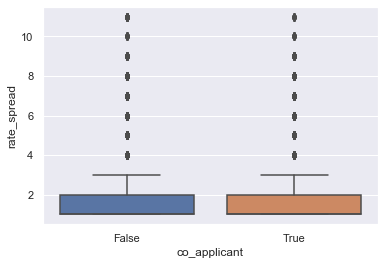
Applicants where ethincity was not provided have wider range of rate spread and higer median.



Applicant race did not observe big differences for rate spread except American Indian or Alaska Native where spreads where wider.



No difference for rate spread among sexes.

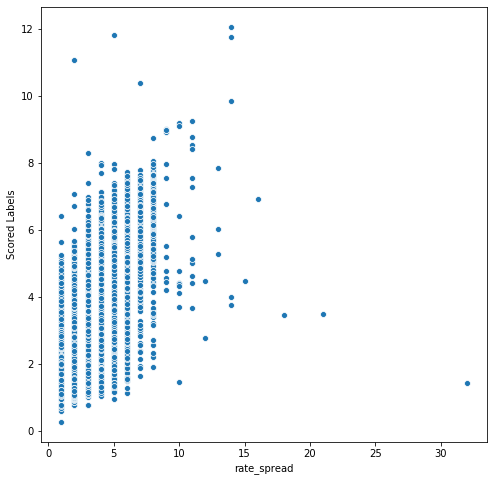


Finally, no difference for rate spreads among mortages with and without co-applicants.

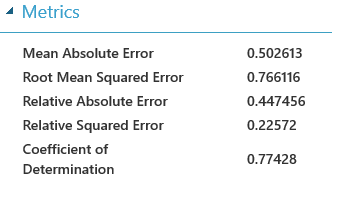
# Prediction of rate spread

A regression model to predict the actual rate spread of mortgages was created. Based on the apparent relationships identified when analyzing the data, a Boosted Decision Tree model was created to predict the value for rate spread.

The model was trained with 75% of the data and tested with the remaining 25%. A scatter plot showing the predicted rate spread and the actual rate spread is shown below:



## Metrics



This plot shows a linear relationship between predicted rate spread and actual values in the test dataset. R squared coefficient of determination is 0.77 which is the main score metric.

## Conclusion

This analysis has shown that the rate spread of a mortgage can be confidently predicted from its characteristics.