Loan Rate Spreads

Kurt Hinderer, Nov, 2019

Executive Summary

This document shows an analysis of loan data and the difference between the rate giving and the standard mortgage rate. The dataset was adapted from the Federal Financial Institutions Examination Council's (FFIEC).

After exploring the data, visuals were created. A predictive boosted tree model was created analysis the data.

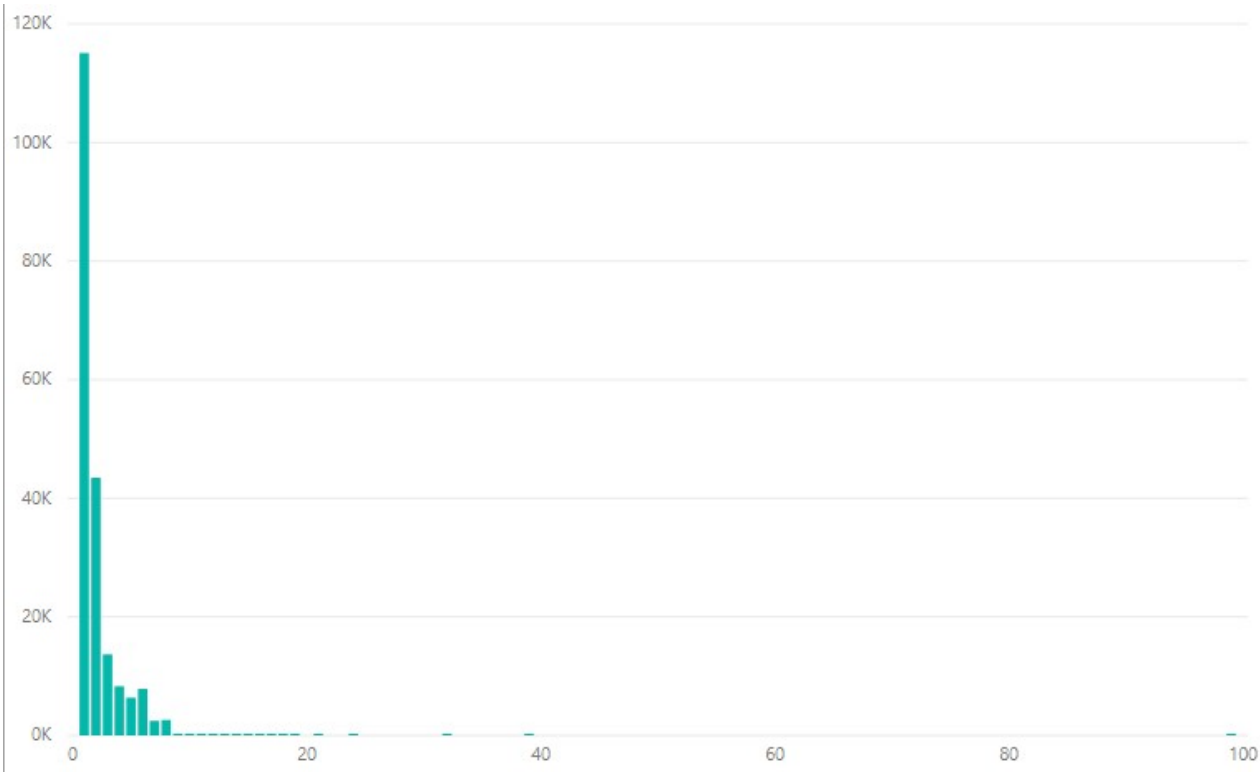
While there were many factors, there were two that were surprisingly useful in the prediction. The lender and the location both made a major difference in creating a predictive model.

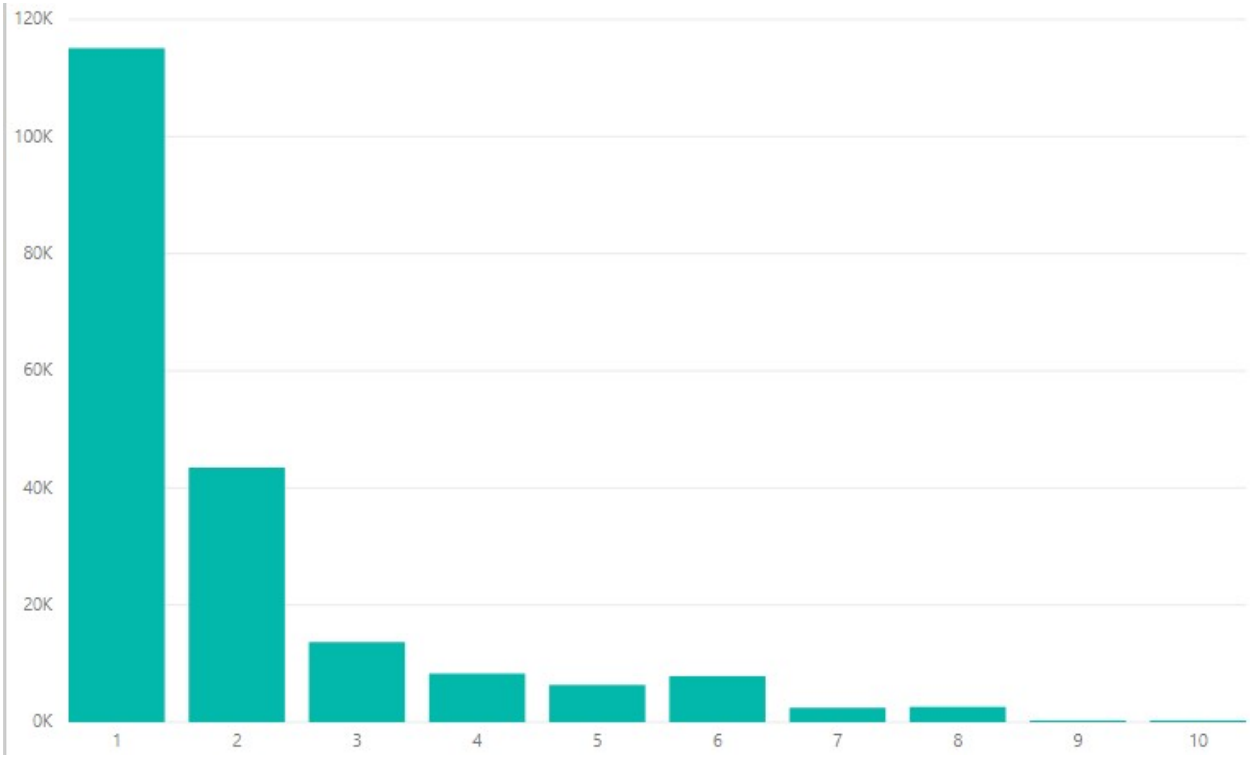
Data Analysis

The dataset had 200,000 observations. There were 99 outliers removed so only 199,901 observations used.

Rate Spread

The rate spread is the difference between the rate given for a mortgage and the standard mortgage rate for a comparable mortgage. This value in the training date ranged from 1 to 99. The majority of the rate spreads was 1 and very few were on the high end. There were only 99 values that had a rate spread above 10, so they were removed as outliers. A graph of rates spreads both with and without the outliers is shown.


120K
100K
80K
60K
40K
20K
0K
0
20
40
60
80
100


120K
100K
80K
60K
40K
20K
0K
1
2
3
4
5
6
7
8
9
10

Numerical Data

There are 8 numerical data values:

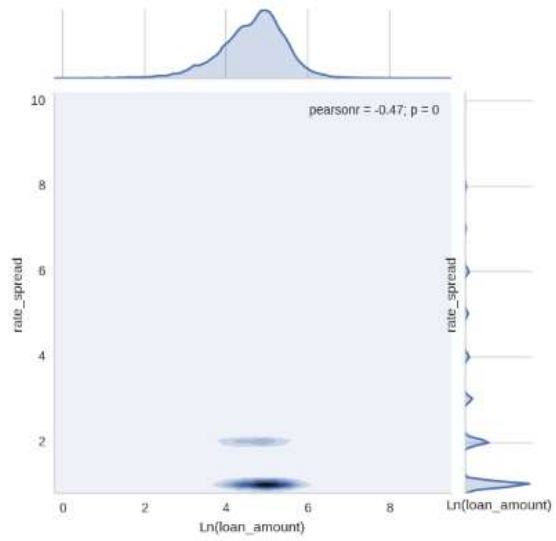
- Loan Amount – The size of the loan in thousands of dollars; the log of this value was used to compare the large differences in size.
- Applicant Income – The applicant's income in thousands of dollars; the log of this value was used to compare the large differences in size.
- Population – The total population of the tract.
- Minority Population Percent – The percentage of minority population to total population for the tract.
- FFIEC Median Family Income – The FFIEC Median family income in dollars for the MSA/MD in which the tract is located (adjusted annually by FFIEC).
- Tract to MSA/MD Income Percent – The percent of the tract median family income compared to MSA/MD median family income.
- Number of Owner-Occupied Units – The number of dwellings, including individual condominiums, that are lived in by the owner.
- Number of 1 to 4 family units – The number of dwellings that are built to house fewer than 5 families.

The summary statistics for these columns (mean, median, minimum, maximum, and standard deviation are shown in this chart.

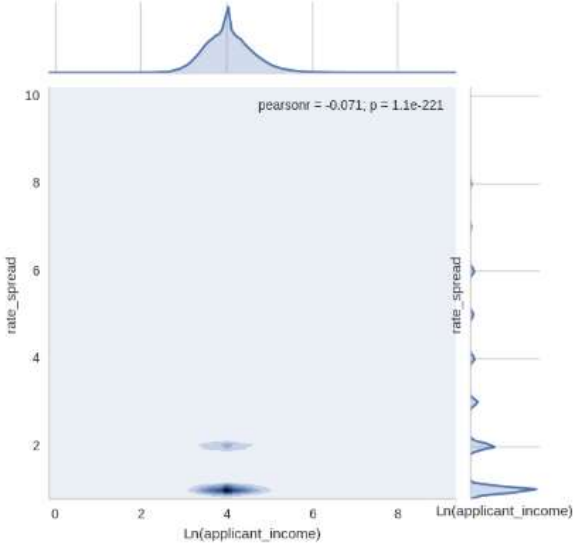
Column	Mean	Median	Minimum	Maximum	Standard Deviation
Loan Amount	142.6	116	1	11,104	142.5
Applicant Income	73.6	56	1	10,042	105.7
Population	5,391.4	4,959	7	34,126	2,669.1
Minority Population Percent	34.2	26.0	0.34	100	27.9
FFIEC Median Family Income	64,594	63,484	17,860	125,095	12,724
Tract to MSA/MD Income Percent	89.3	99.0	6.2	100	15.1
Number of Owner-Occupied Units	1,403	1,304	3	8,747	706.8
Number of 1 to 4 family units	1,927	1,799	6	13,615	886.6

There were some missing values which were replaced by the median. A 2D density plot vs. the rate spread is shown; unfortunately, it doesn't show much as the majority of rate spreads have a value of 1.

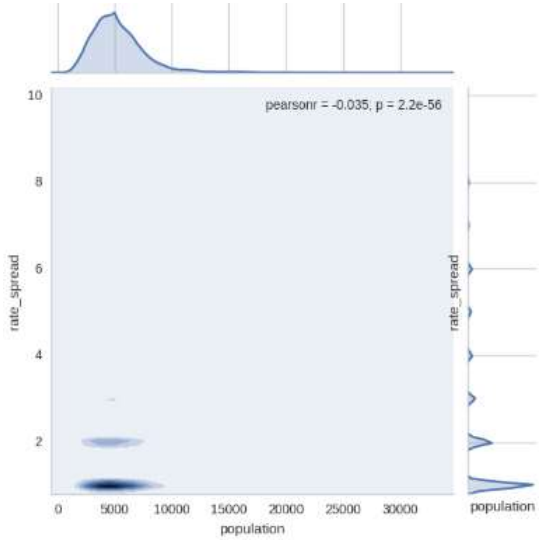
Loan Amount (natural log) vs. Rate Spread



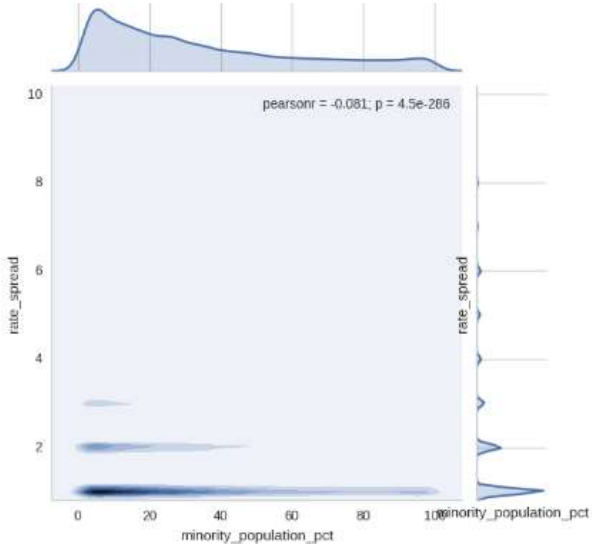
Applicant Income (natural log) vs. Rate Spread



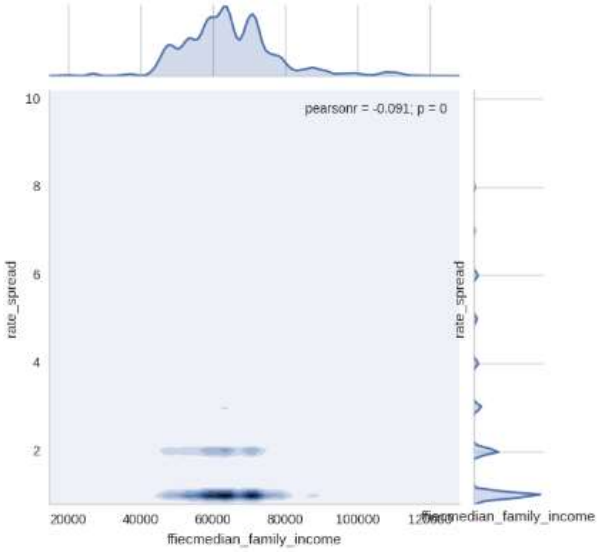
Population vs. Rate Spread



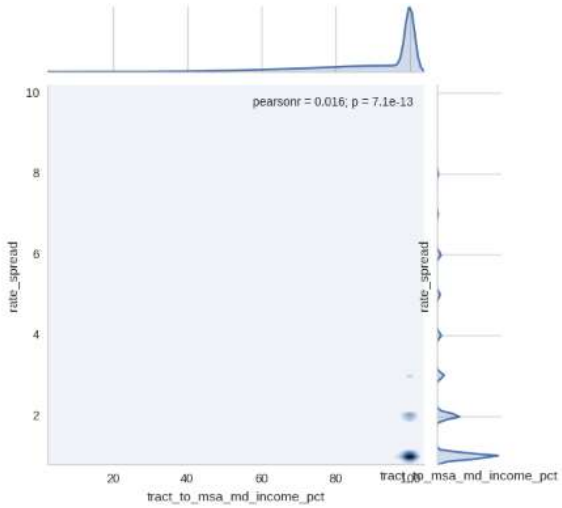
Minority Population % vs. Rate Spread



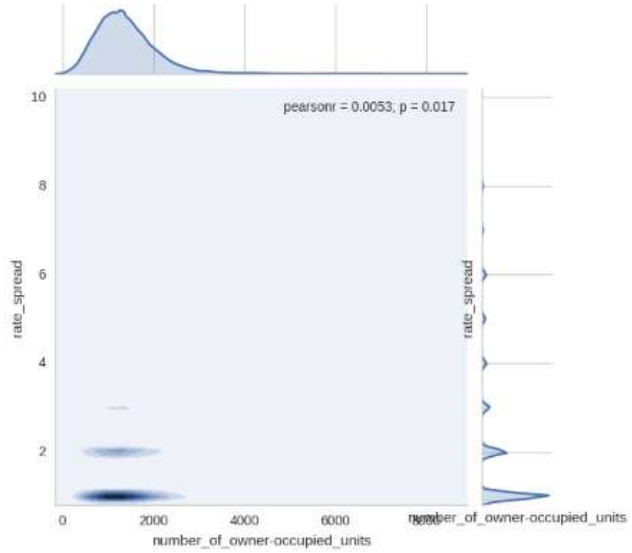
FFIEC Median Family Income vs. Rate Spread



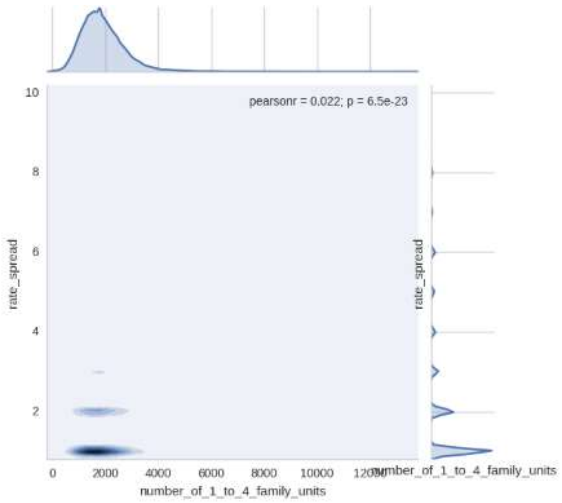
Tract to MSA/MD Income % vs. Rate Spread



Number of Owner-Occupied Units vs. Rate Spread



Number of 1 to 4 family units vs. Rate Spread



Categorical Data

There are 13 different categorical values but there are two different types of categorical data presented. One type has only a few different categories, this includes the demographic data. The other type has a massive amount of data, this includes geographical areas and individual lenders. Because of this, these data needed to be analyzed and visualized differently.

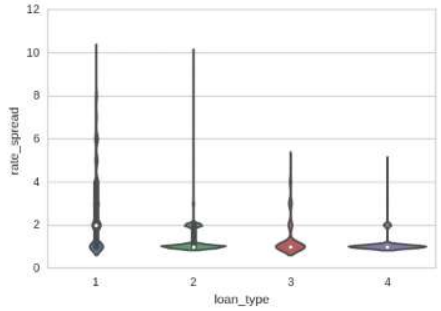
In the type that only has a few categories, the individual categories could be analyzed. The 9 different data values are:

- Loan Type
- Property Type
- Loan Purpose
- Occupancy
- Preapproval
- Applicant Ethnicity
- Applicant Race
- Applicant Sex
- Co-applicant

The number of each category in each value along with a violin plot are shown. The violin plots are much larger at the rate spread of 1 due to the large number of loans at that value.

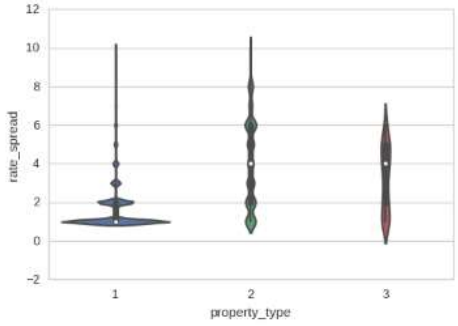
Loan Type

Value	ID #	Count
Conventional	1	90,610
FHA-insured	2	106,305
VA-guaranteed	3	1,082
FSA/RHS	4	1,904



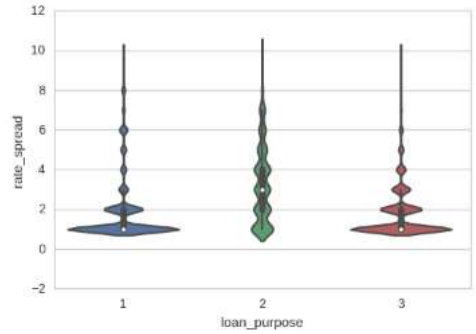
Property Type

Value	ID #	Count
1 to 4 – family	1	169,194
Manufactured Housing	2	30,469
Multi-family	3	238



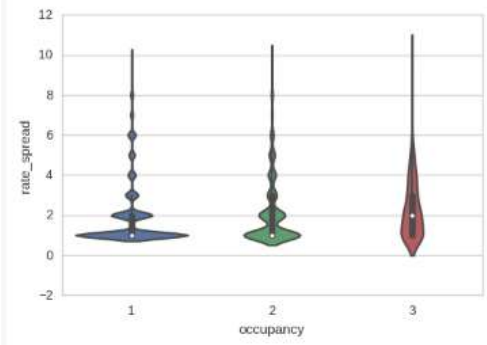
Loan Purpose

Value	ID #	Count
Home Purchase	1	146,072
Home Improvement	2	11,238
Refinancing	3	42,591



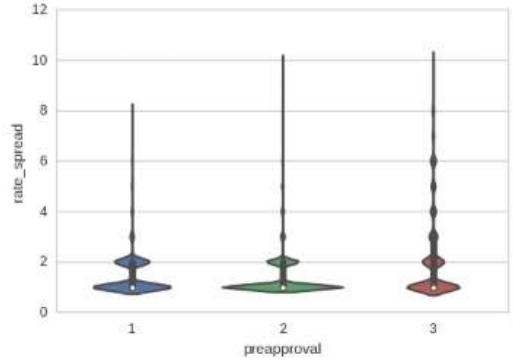
Occupancy

Value	ID #	Count
Owner-Occupied	1	187,923
Not Owner-Occupied	2	11,687
N/A	3	291



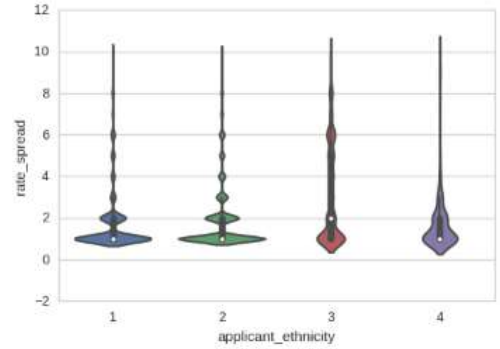
Preapproval

Value	ID #	Count
Preapproval Requested	1	8,886
Preapproval Not Requested	2	41,620
N/A	3	149,395



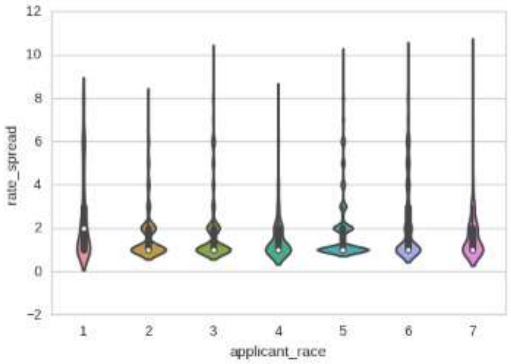
Ethnicity

Value	ID #	Count
Hispanic or Latino	1	34,805
Not Hispanic or Latino	2	147,938
Information Not Provided	3	16,445
N/A	4	713



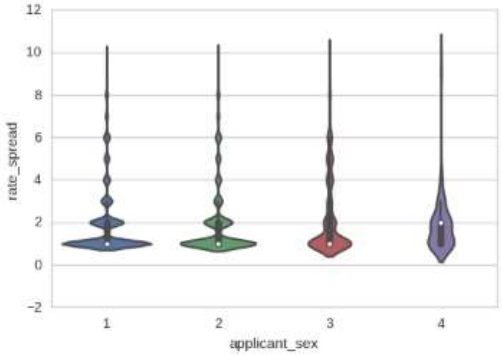
Race

Value	ID #	Count
American Indian or Alaskan Native	1	1,687
Asian	2	4,606
Black or African American	3	20,747
Native Hawaiian or Other Pacific Islander	4	684
White	5	157,535
Information Not Provided	6	14,001
N/A	7	641



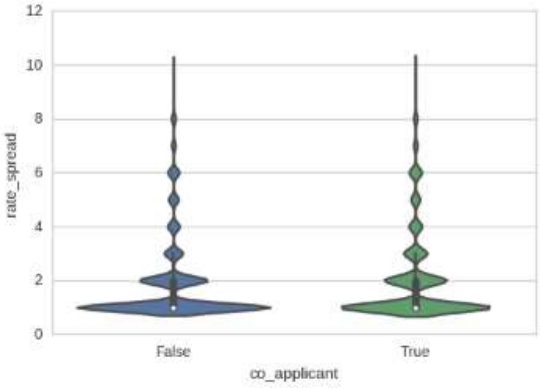
Sex

Value	ID #	Count
Male	1	124,947
Female	2	66,934
Information Not Provided	3	7,541
N/A	4	479



Co-Applicant

Value	ID #	Count
Co-Applicant	True	76,670
No Co-Applicant	False	123,231

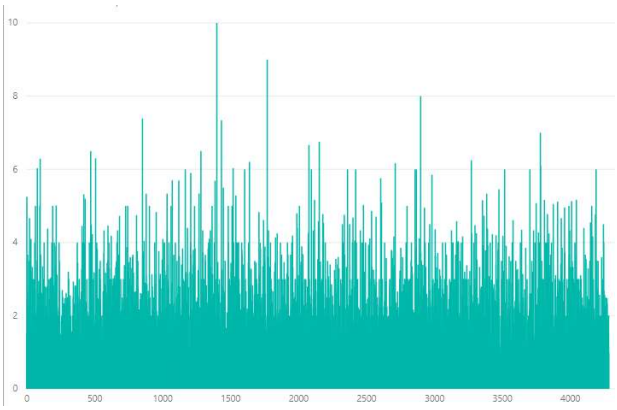


The other categorical data had a number of categories too large to easily see the differences with counts and violin plots. Instead, the difference in the rate spread averages was observed. Each category was identified by a number and there was no way to identify what that number represented. Each of these had rather random spreads for their categories. The categorical data with a large number of categories are:

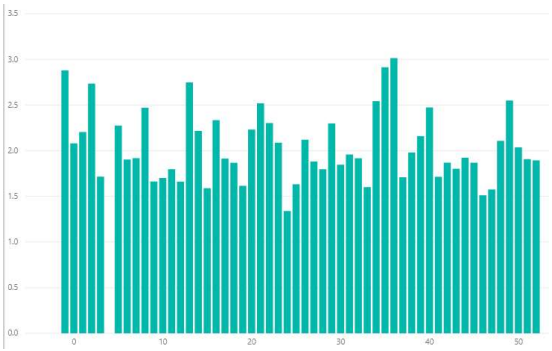
	Lender	State	County	Metropolitan Area
Number of Categories	3,892	53	306	409
Minimum Mean Rate Spread	1.00	1.33	1.00	1.13
Mean-Mean Rate Spread	2.01	2.04	2.32	2.02
Maximum Mean Rate Spread	10.00	3.01	5.04	3.60

The graphs of these values show a rather random change between each category

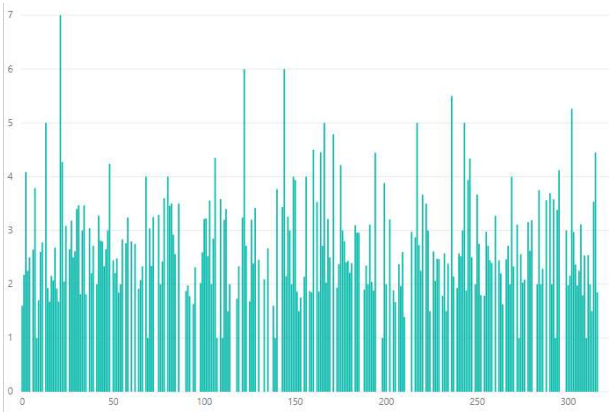
Lender



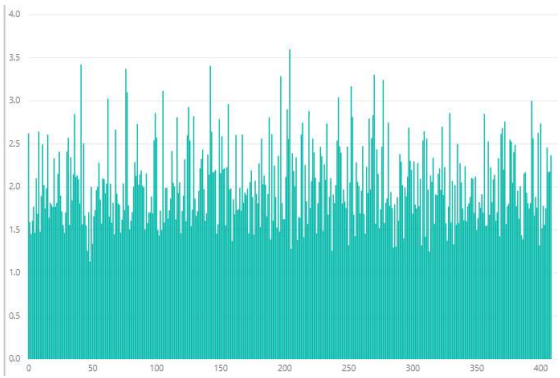
State



County



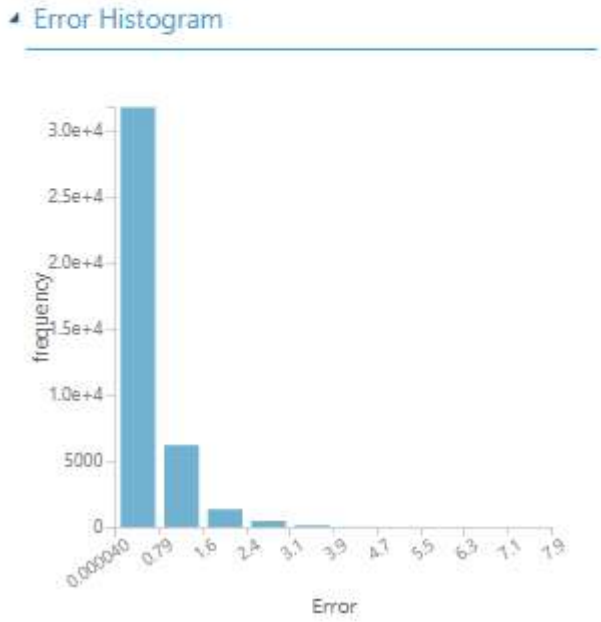
Metropolitan Area



Predictive Model

In selecting the features, the Population, Minority Population Percent, and the Co-Applicant did not appear to make a difference in the model and so were removed. The Boosted Decision Tree Regression model appeared to work the best in this situation. The data was split with 70% to train and 30% to test. The results of the test are shown.

Mean Absolute Error	0.535257
Root Mean Squared Error	0.762867
Relative Absolute Error	0.47694
Relative Squared Error	0.229193
Coefficient of Determination	0.770807



Conclusion

The rate spread of a loan can be predicted. While many features are useful in determining this, the lender and the location are very important factors in determining the rate spread.