THE REPORT FOR THE HMDA DATA

EXECUTIVE SUMMARY

This document presents an analysis of data concerning loans and mortgages. The analysis is based on 500,000 observations of HMDA data, each containing specific characteristics of loans and mortgages acquired in the past and its amount.

After exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the data, several potential relationships between loan characteristics and possibility of acquisition were identified. After exploring the data, a predictive model to classify loan possibility/acquisition categories was created, and finally a classification model to predict a loan grant from its features was created.

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

1.Most Acquisition of loans were on the lower side i.e, people who asked for loans were people that fell below 200,000 on the hist plot.

2. Major factors that helped to indicate the possibilities of a customer acquiring mortgages/loans are significant features found in this analysis :

a. Loan_Amount
b. Purpose of Loan
c.Applicant income
d. Applicant_sex
e. Co Applicant income
f. Applicant Ethnicity
g.Preapproval
h. Occupancy

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 500 observations are shown here:

	row_id	loan_type	property_type	loan_purpose	occupancy	loan_amount	preapproval	msa_md	state_code	county_cod
count	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.00000
mean	249999.500000	1.366276	1.047650	2.066810	1.109590	221.753158	2.764722	181.606972	23.726924	144.54206
std	144337.711634	0.690555	0.231404	0.948371	0.326092	590.641648	0.543061	138.464169	15.982768	100.24361
min	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	-1.000000	-1.000000	-1.00000
25%	124999.750000	1.000000	1.000000	1.000000	1.000000	93.000000	3.000000	25.000000	6.000000	57.00000
50%	249999.500000	1.000000	1.000000	2.000000	1.000000	162.000000	3.000000	192.000000	26.000000	131.00000
75%	374999.250000	2.000000	1.000000	3.000000	1.000000	266.000000	3.000000	314.000000	37.000000	246.00000
max	499999.000000	4.000000	3.000000	3.000000	3.000000	100878.000000	3.000000	408.000000	52.000000	324.00000

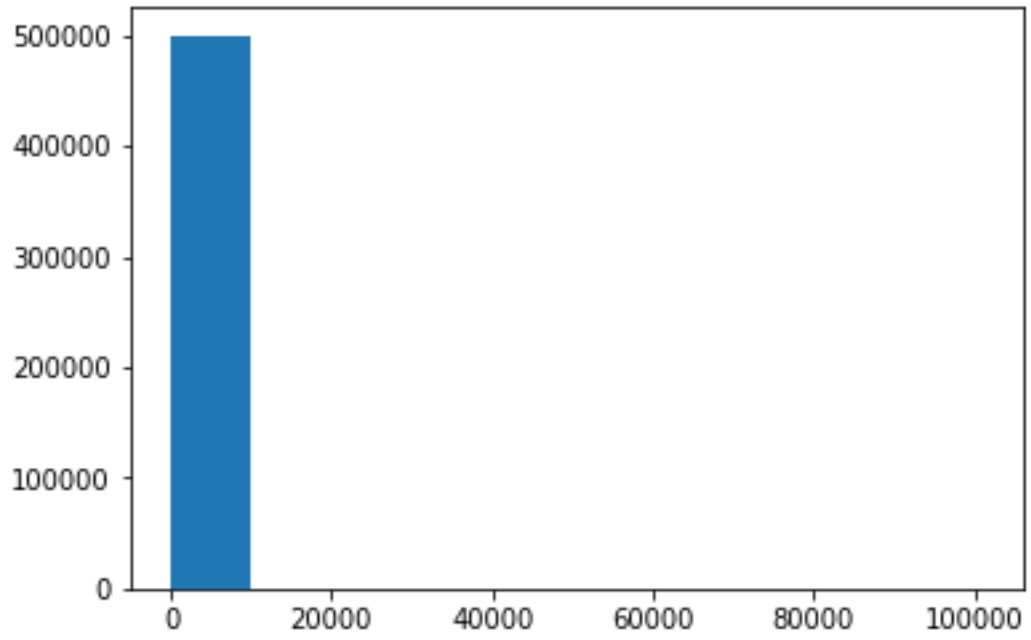
8 rows × 21 columns

And some informative statistics as well, as shown below:

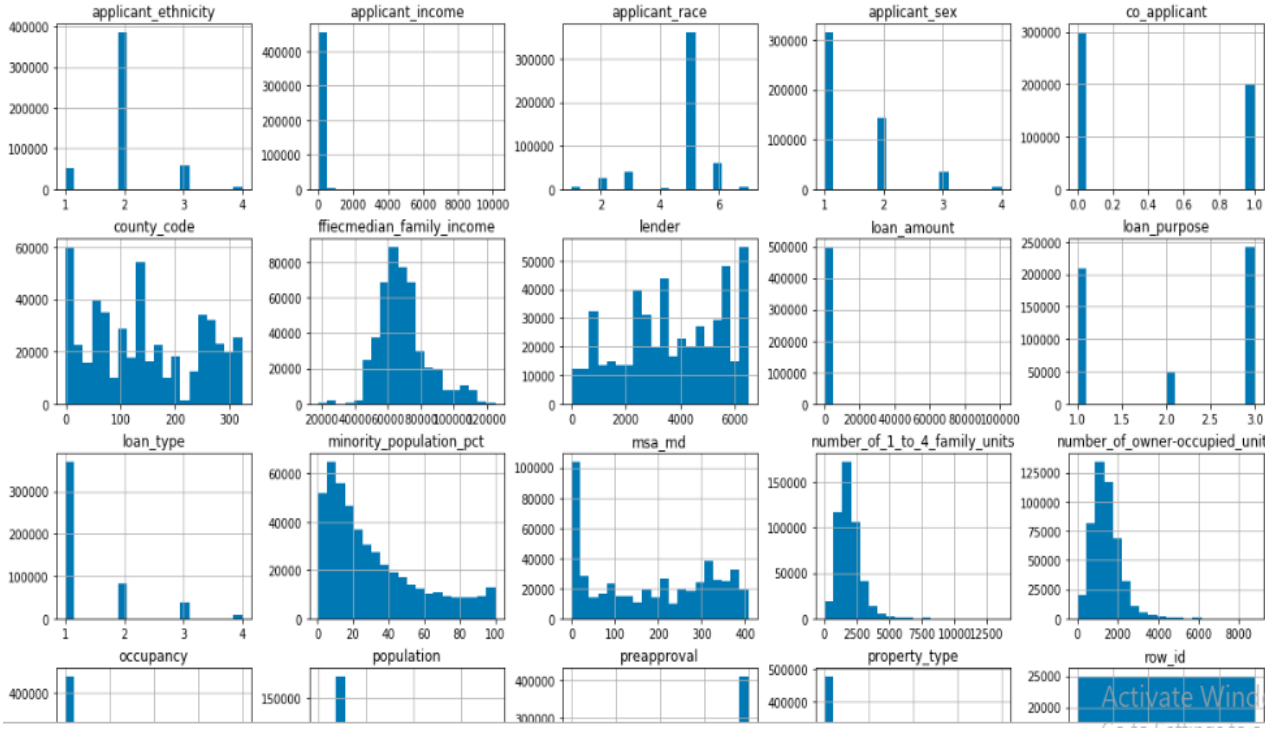
```
Data columns (total 22 columns):
row_id                              500000 non-null int64
loan_type                           500000 non-null int64
property_type                       500000 non-null int64
loan_purpose                        500000 non-null int64
occupancy                           500000 non-null int64
loan_amount                         500000 non-null float64
preapproval                         500000 non-null int64
msa_md                              500000 non-null int64
state_code                          500000 non-null int64
county_code                         500000 non-null int64
applicant_ethnicity                 500000 non-null int64
applicant_race                      500000 non-null int64
applicant_sex                       500000 non-null int64
applicant_income                    460052 non-null float64
population                          477535 non-null float64
minority_population_pct             477534 non-null float64
ffiecmedian_family_income           477560 non-null float64
tract_to_msa_md_income_pct          477486 non-null float64
number_of_owner-occupied_units      477435 non-null float64
number_of_1_to_4_family_units       477470 non-null float64
lender                              500000 non-null int64
co_applicant                        500000 non-null bool
dtypes: bool(1), float64(8), int64(13)
```

Since **Loan Amount** 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 (596.64) indicates that there is considerable variance in the loan amounts. A histogram of the **loan_amount** column shows that the values are right-skewed – as shown here:

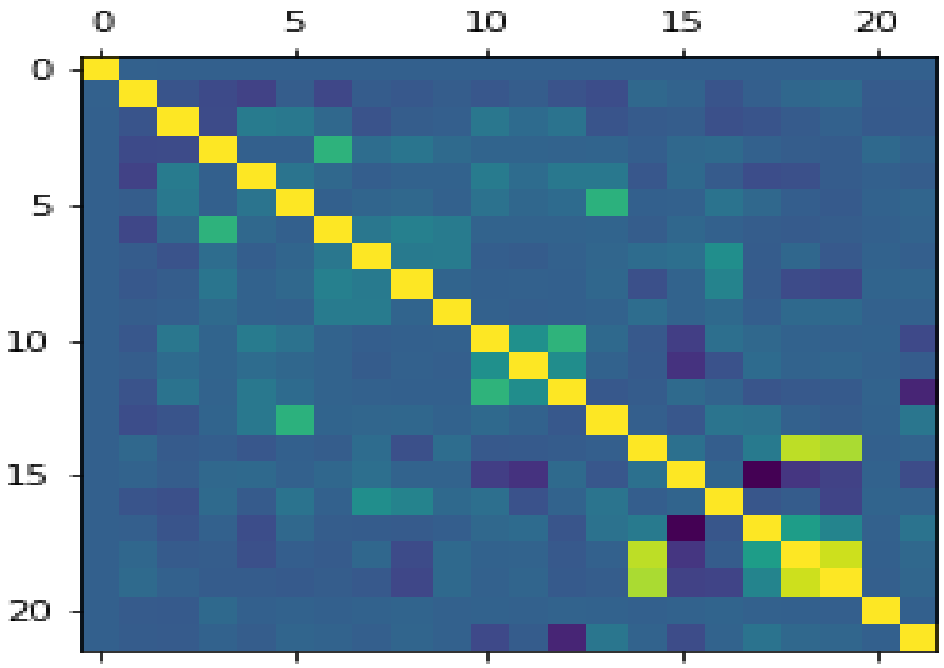


Bar charts were created to show frequency of these features, and indicate the following:



Correlation and Apparent Relationships

After exploring the individual features, an attempt was made to identify relationships between features in the data – in particular, between **Accepted** 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:

Classification of customers based on loan requisition acceptance

Based on the analysis of the HMDA, a predictive model to classify customers into two categories: 0(not accepted) and 1 (accepted)

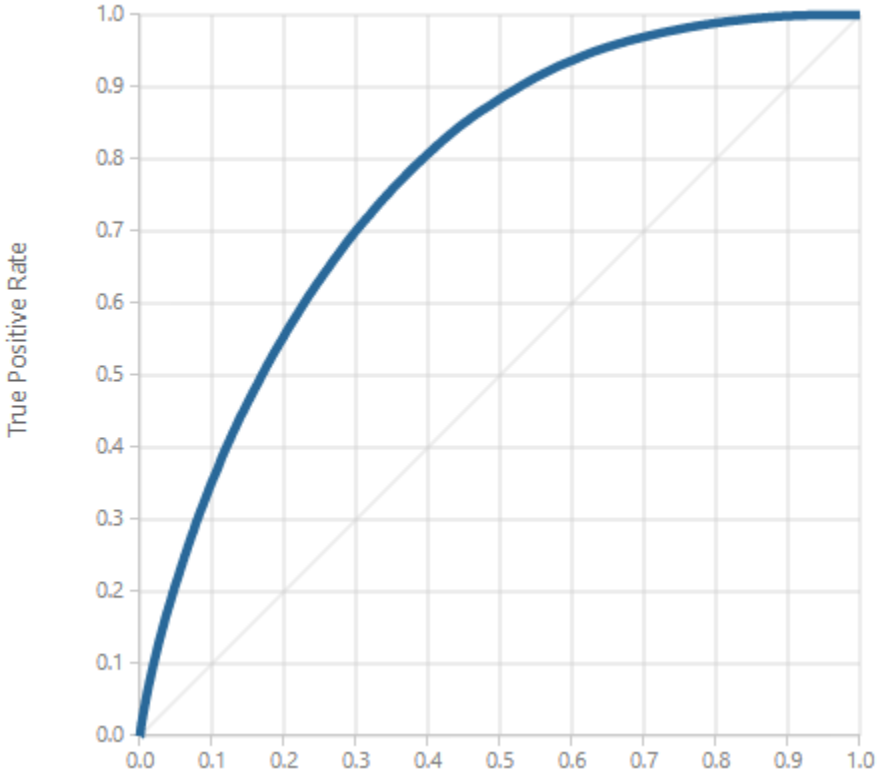
The model was created using the Two-Class Boosted Decision Trees algorithm and trained with 75% of the data. Testing the model with the remaining 25% of the data yielded the following results:

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
141577	45951	0.704	0.685	0.5	0.774

False Positive	True Negative	Recall	F1 Score
65147	122325	0.755	0.718

Positive Label	Negative Label
1	0

The Received Operator Characteristic (ROC) curve for the model is shown here, with the blue line indicating the model's performance at varying classification threshold values, and the diagonal line showing the expected results of a random guess:



Conclusion

This analysis has shown that the probability that a customer can receive loan can be confidently predicted from its characteristics including the demographics of the customers. In particular, loan_amount, coapplicant, purpose of loan, and applicant income have a significant effect the prediction. Secondary features, such as applicant sex and loan type can help further classify customers and determine which group they belong.