Analysis of Home Mortgage Disclosure Act

Taiwo Ogundare

Executive Summary

This document presents an analysis of data concerning Home Mortgage Disclosure Act and its applicants. the analysis is based on 500,000 observations and the result of thier applicant either accepted or rejected. After exploring the data by calculation summary and descriptive statistics, and by creating visualisation of the data, several potential relationships between HMDA characteristics and application were identified. After exploring the data, a predictive model to classify HMDA applicants into two categories was created.

After performing the analysis, the following conclusions were made:

while many factors can help indicate whether the application will be accepted, significant features found in this analysis were:

-**lender acceptance rate** - acceptance rate per lender. The acceptance rate of some lenders are very low compare to ohters.

-**Applicant Income** - the income of the applicants. Applicants with income greather than 1000 has a 53.6% acceptance rate.

-**Loan Amount** - amount the applicant is loaning. Applicants applying for greater than 2000 has a 63% accpetance rate.

-**Loan Purpose** - purpose of the loan. application with loan purpose home purchases tend to have high acceptance rate(59%) than home improvement(45%) and refinancing(33%).

-**Lender Unigue Count** - unique counts of lender.

-**Mean Applicant Income Per Lender** - mean applicant income per lender.

EXPLORATORY DATA ANALYSIS

The exploration of the data began with some summary and descriptive statistics. the Home Mortgage Disclosure Act data set has 500000 rows and 21 features with the target feature accepted for the training data and 500000 rows and 21 features for the testing data set.

```
## 'data.frame':    500000 obs. of  22 variables:
##  $ loan_type                    : Factor w/ 4 levels "1","2","3","4": 3 1
2 1 1 1 3 2 1 1 ...
##  $ property_type                : Factor w/ 3 levels "1","2","3": 1 1 1 1
1 1 1 1 1 1 ...
##  $ loan_purpose                 : Factor w/ 3 levels "1","2","3": 1 3 3 1
```

```
1 3 1 1 3 3 ...
##  $ occupancy                    : Factor w/ 3 levels "1","2","3": 1 1 1 1
1 1 1 1 2 1 ...
##  $ loan_amount                  : num  70 178 163 155 305 133 240 210 209
197 ...
##  $ preapproval                  : Factor w/ 3 levels "1","2","3": 3 3 3 1
3 3 3 3 3 3 ...
##  $ msa_md                       : Factor w/ 409 levels "-
1","0","1","2",..: 20 370 18 307 26 223 375 324 26 196 ...
##  $ state_code                   : Factor w/ 53 levels "-
1","0","1","2",..: 39 53 12 49 39 15 30 39 39 11 ...
##  $ county_code                  : Factor w/ 318 levels "-
1","1","2","3",..: 242 294 300 179 21 56 131 36 21 21 ...
##  $ applicant_ethnicity          : Factor w/ 4 levels "1","2","3","4": 2 1
2 2 2 2 1 1 2 2 ...
##  $ applicant_race               : Factor w/ 7 levels "1","2","3","4",..:
5 5 5 5 3 5 5 5 5 5 ...
##  $ applicant_sex                : Factor w/ 4 levels "1","2","3","4": 1 1
1 1 2 2 2 1 1 1 ...
##  $ applicant_income             : num  24 57 67 105 71 51 104 55 244 86
...
##  $ population                   : num  6203 5774 6094 6667 6732 ...
##  $ minority_population_pct      : num  44.23 15.9 61.27 6.25 100 ...
##  $ ffiecmedian_family_income    : num  60588 54821 67719 78439 63075 ...
##  $ tract_to_msa_md_income_pct   : num  50.9 100 100 100 82.2 ...
##  $ number_of_owner-occupied_units: num  716 1622 760 2025 1464 ...
##  $ number_of_1_to_4_family_units : num  2642 2108 1048 2299 1847 ...
##  $ lender                       : int  4536 2458 5710 5888 289 964 5488
2442 2118 3507 ...
##  $ co_applicant                 : Factor w/ 2 levels "FALSE","TRUE": 1 1
1 2 1 1 1 2 2 1 ...
##  $ accepted                     : Factor w/ 2 levels "0","1": 2 1 2 2 2 2
2 2 2 1 ...

## [1] "The training data has 500000 rows and 22 columns"
```

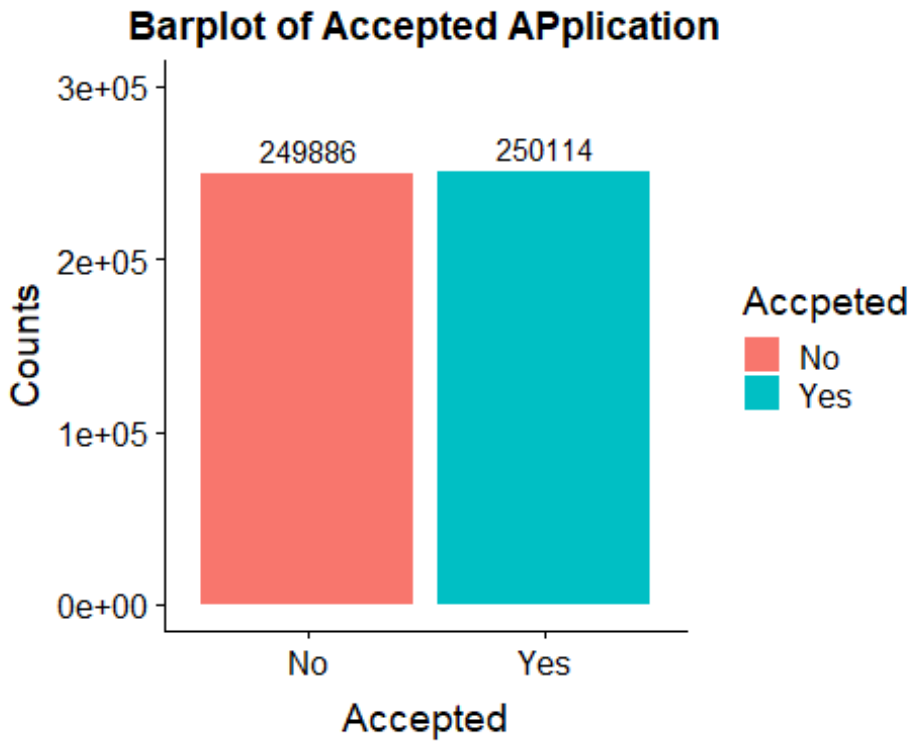
Summarise the missing values in the data

```
## [1] "there are 7 columns with missing values. below are the summarises for
each column."

## number_of_owner-occupied_units  number_of_1_to_4_family_units
##                          22565                           22530
##     tract_to_msa_md_income_pct        minority_population_pct
##                          22514                           22466
##                     population      ffiecmedian_family_income
##                          22465                           22440
```

Fixing Missing Values

For each feature with missing values, the NA's were replaced with the median of their respective distribution.

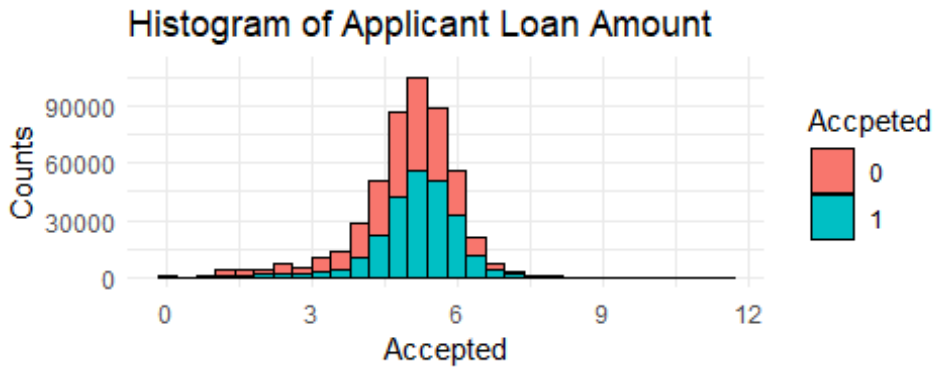


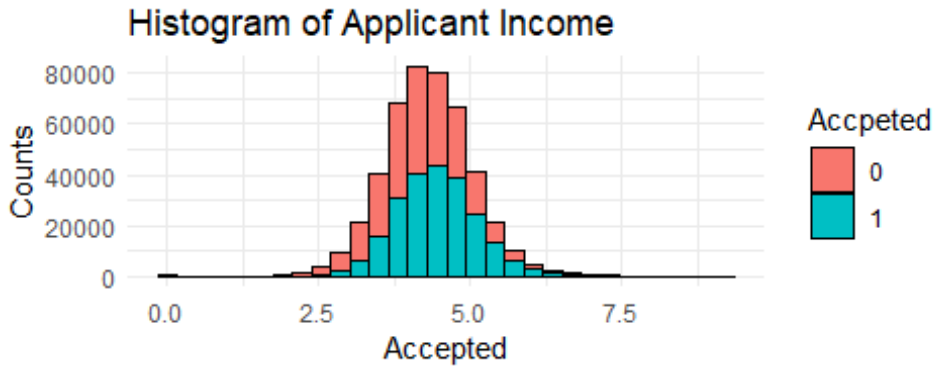
Target Variable

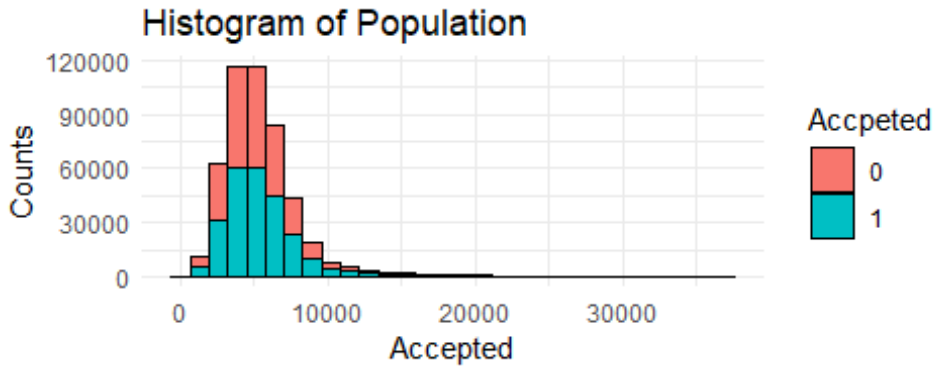
NUMERICAL FEATURES

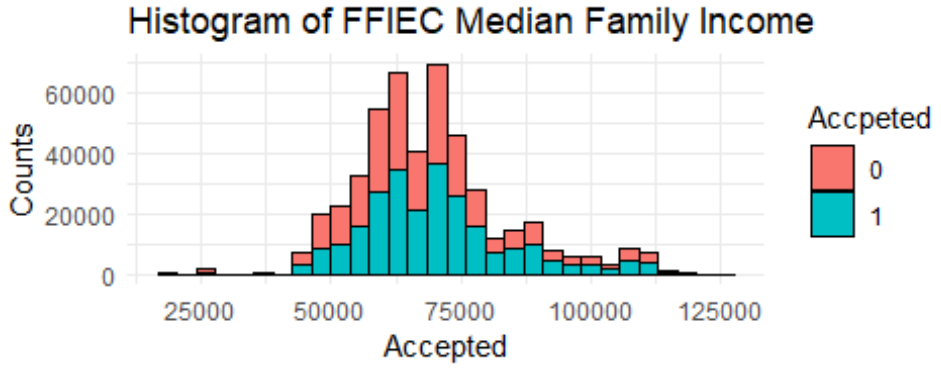
Numeric Relationships

The following plot was generated initially to analyse and compare numeric features. The key numeric features in the data are visualised below:





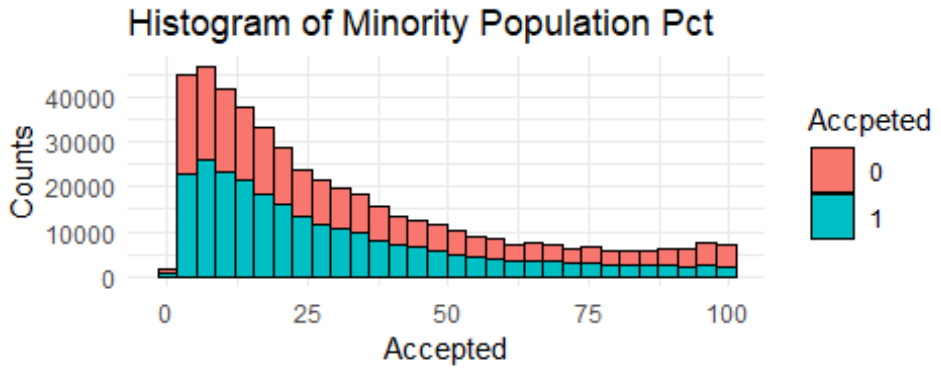


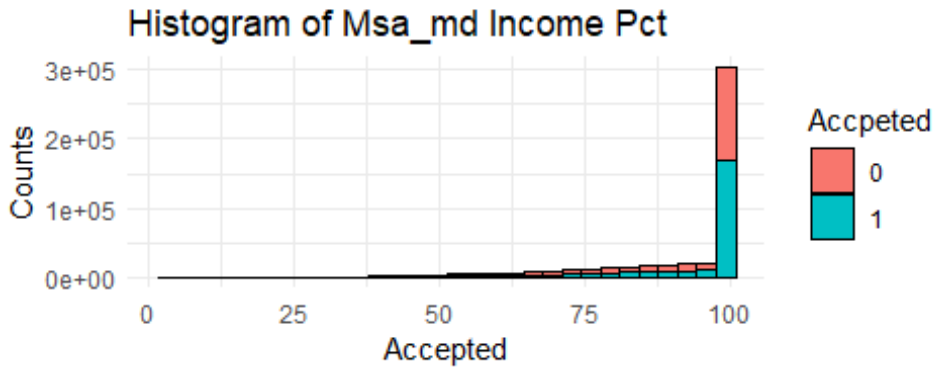


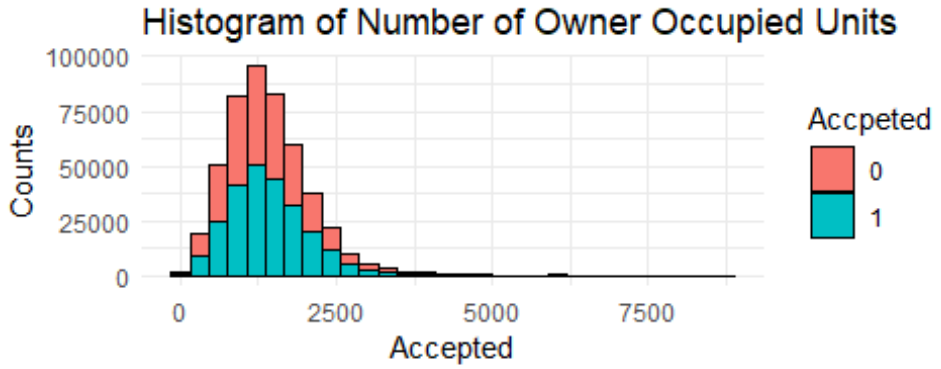
-**Loan Amount** initially has a left skewd distribution, taking log transformation of the features shows a normal distribution with few outliers and plot acertain that both accpeted and rejected applications has the same distribution.

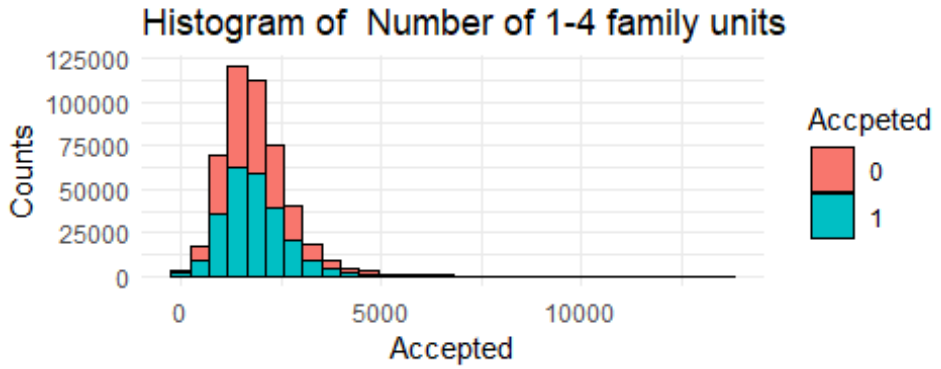
-**Population** has a left skewd distribution.

-**Applicant Income** initially has a left skwed distribution, taking log transformation of the feature show a normal distribution with few outliers.

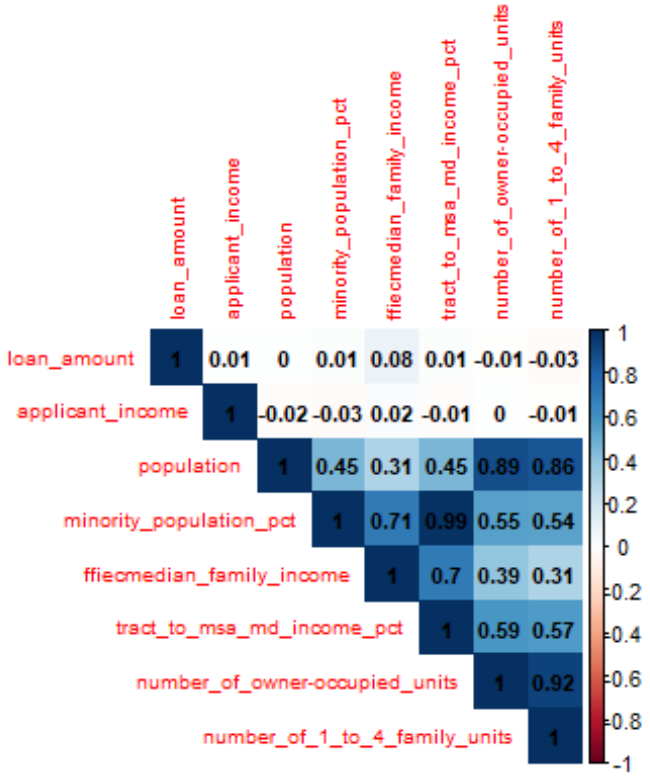








- **No of owner Occupied Units** has a left skewed distribution.
- **Msa_md Income Percentage** plot has a right skewed distribution and ranges from about 4% to 100% with majority of the applicants having a 100% msa_md income percentage.
- **Number of 1-4 family Units** plot shows a left skewed distribution.



- minority population percentage and tract to msa_md income percentage have a correlation of 0.99.
- population has a high correlation with both number of owner occupied units and nmuber of 1-4 family units which is a result of all three being counts of personnels. - loan amount and applicant income have extremely low correlation with other numeric features.
- number of owner occupied units has a high correlation with number of 1-4 family units.

CATEGORICAL FEATURES

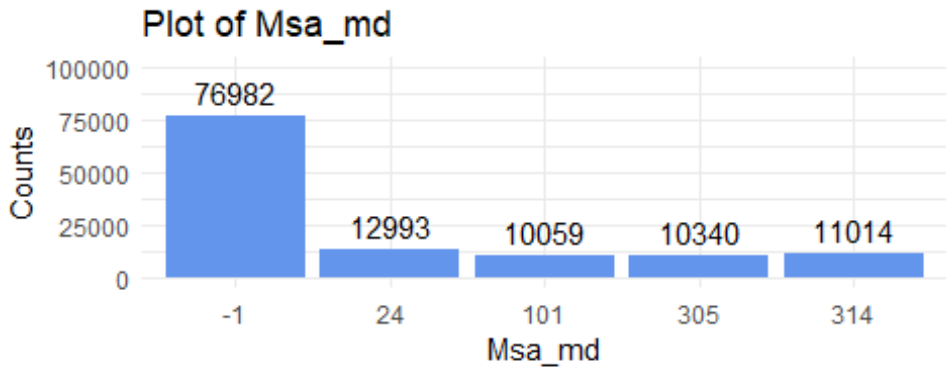
The following are the categorical features as in the data;

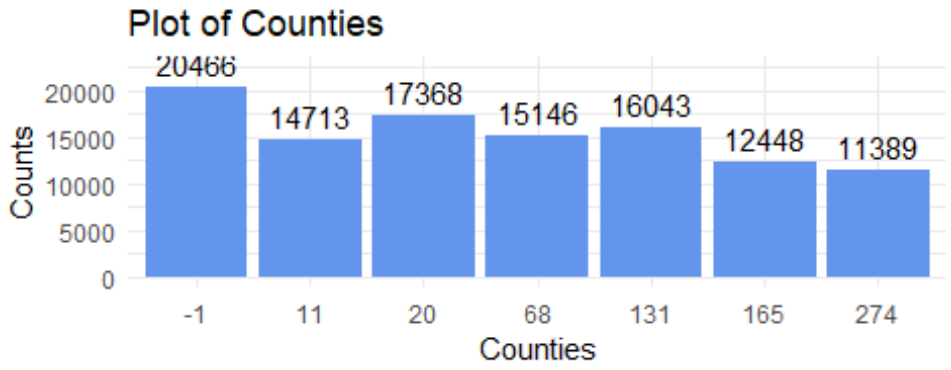
- **Loan type -** One of Conventional, Federal adminstration, veteran adminstration,farm service agency
- **Property type**
- **Loan Purpose**
- **Occupany**

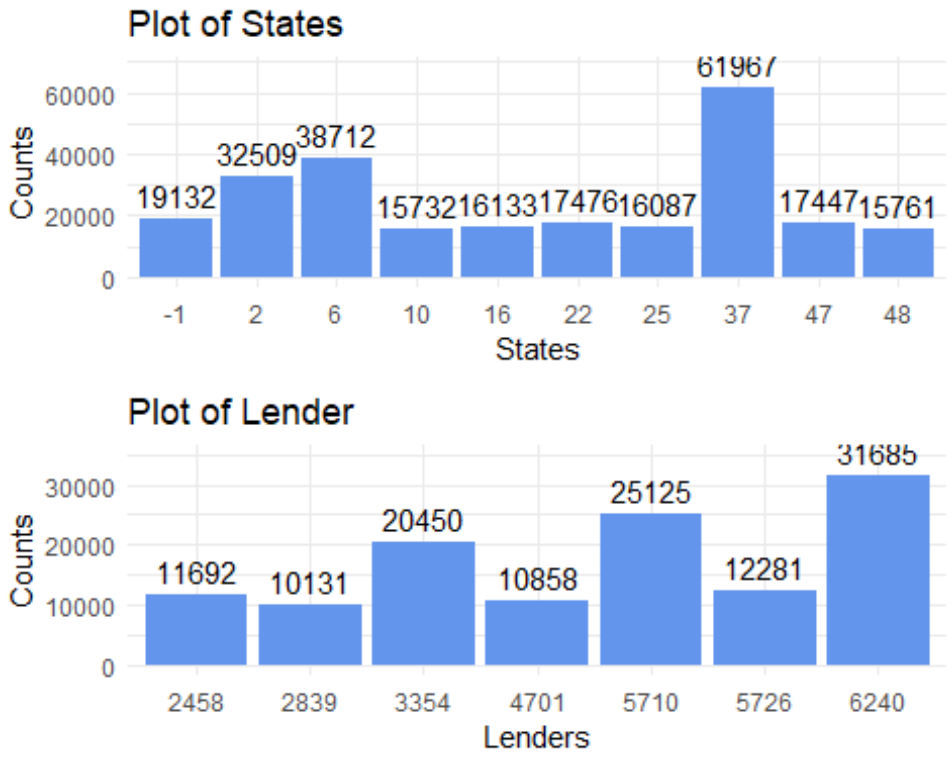
- **Preapproval**
- **Msa_md**
- **State_code**
- **county_code**
- **Applicant ethnicity**
- **Applicant race**
- **Applicant sex**
- **Lender**
- **Co Applicant**

categorical features were explored and visualised with charts, the following conclusion were made;

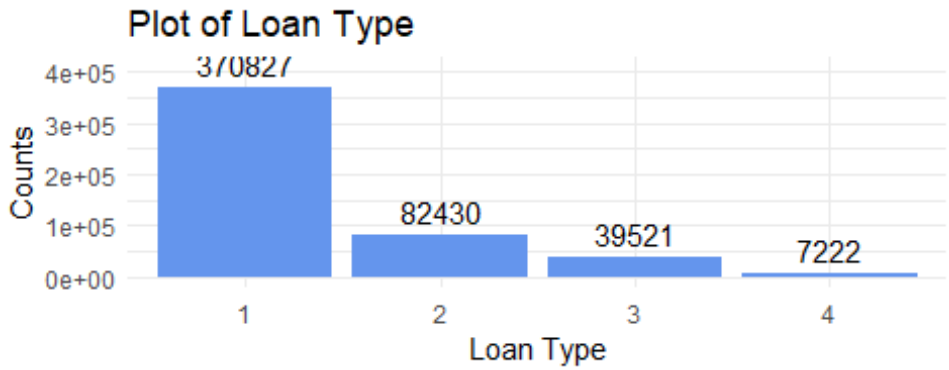
- Lender ID 6240 is the most common lender followed by lender 5710 and 3354 respecrtively
- State 37,6,2 are the most common states tallying about 26.64% of the population, there are 19132 missing values in the state value assinged value -1
- About 15.39% of the values in the msa_md column are missing.

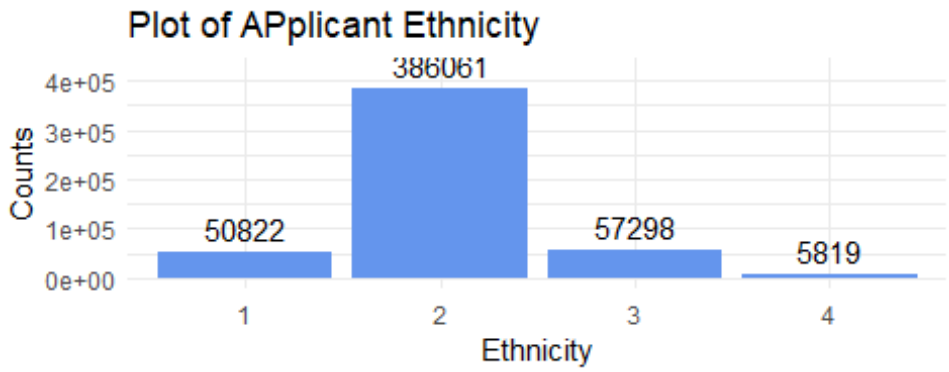


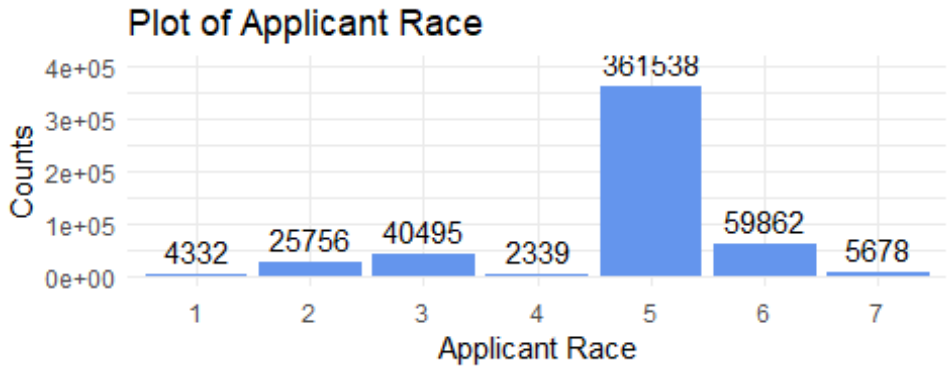


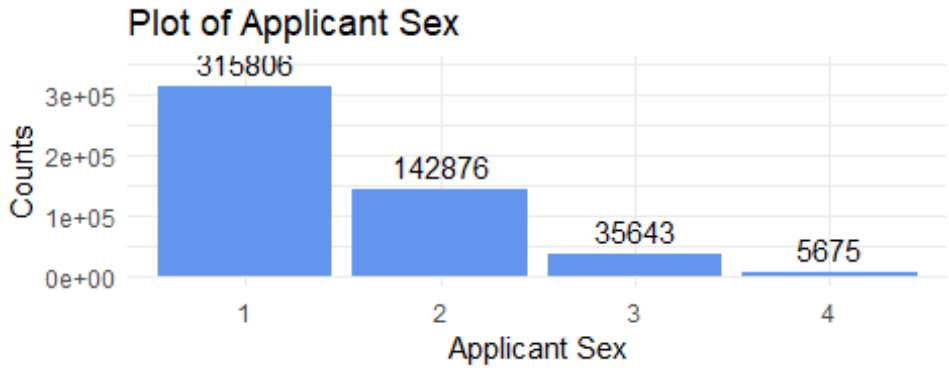


- Conventional loan type is the most common loan type, followed by federal housing adminstration. Veteran adminstration and Farm service agency are much less common.
- The majority of applicants are Male.
- Male applicants are more common than female applicants.
- Hispanic/Latino is the most common applicant ethnicity.
- The majority of applicant are white, with very small frequencies for each of the other values.



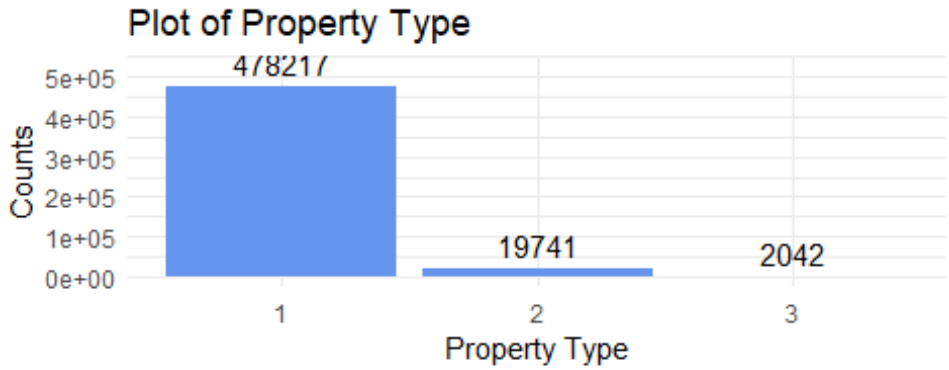


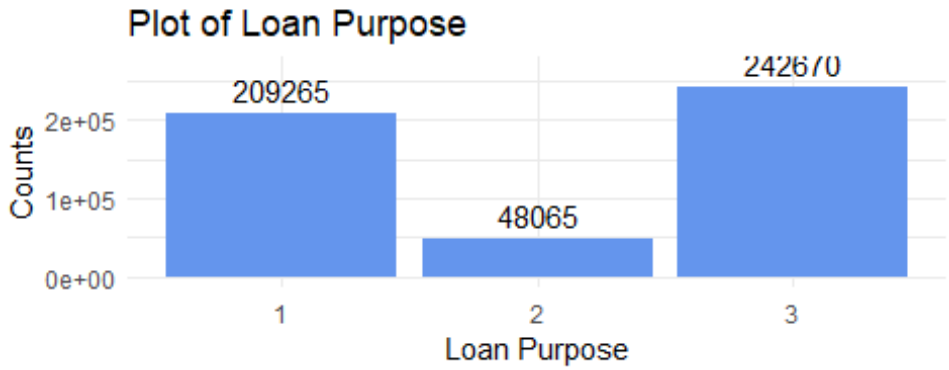


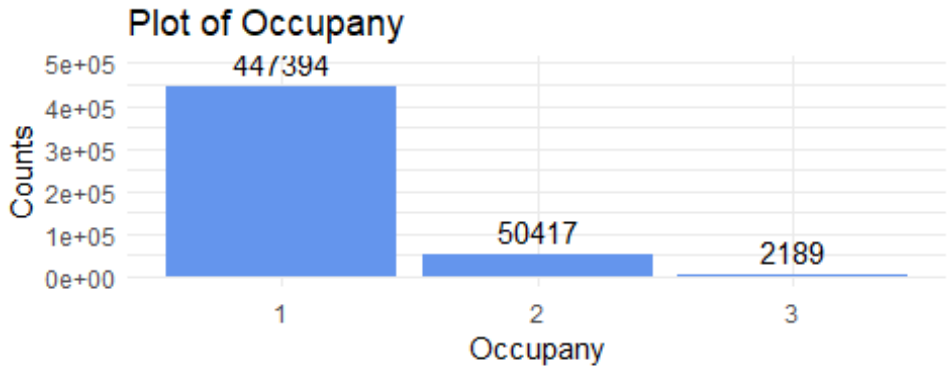


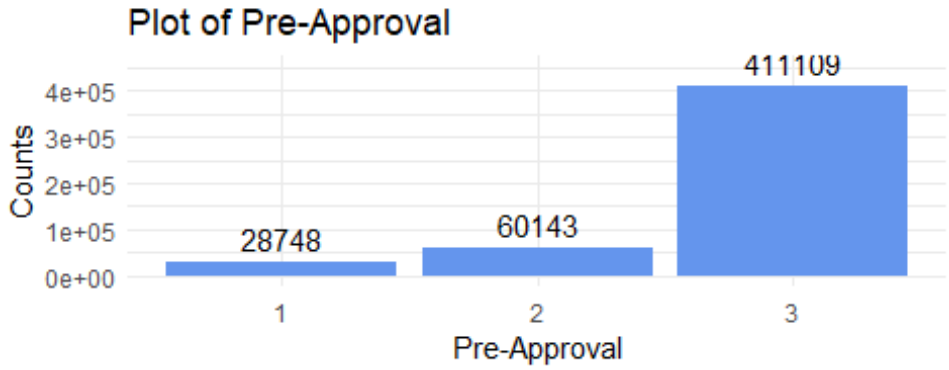
- MultiFamily is the most common property type.
- One - Four family are more common than manufactured housing property.

- Re-Financing is the more common than home improvemnet with home purchasing being relatively uncommon.
- Pre-Approval for loans are not applicable to most of the applicants.







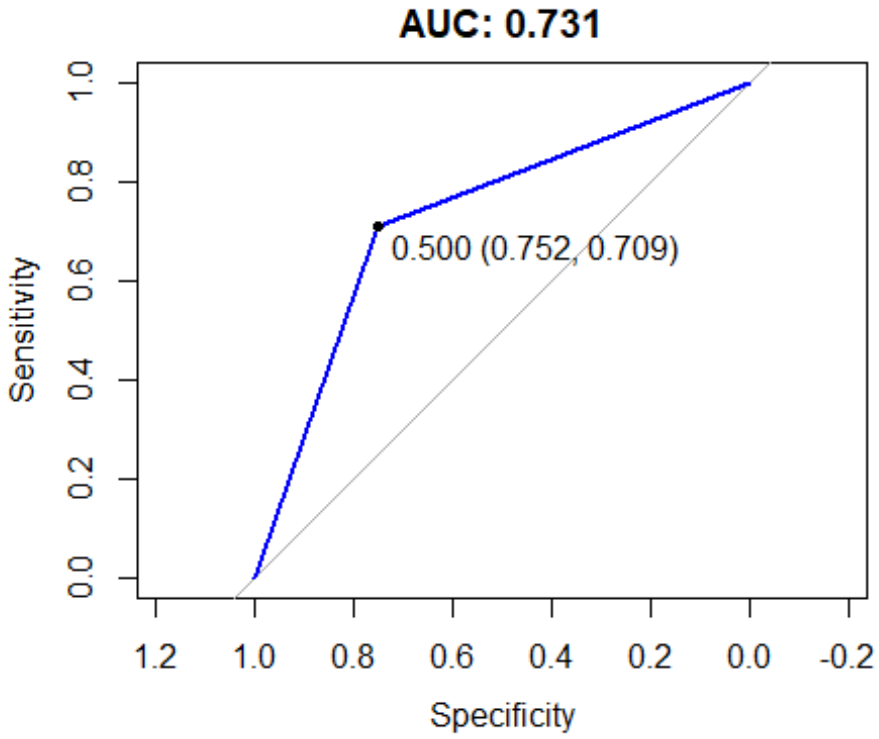


MODEL SUMMARY

Based on the analysis of the Home Mortgage Disclosure Act data, a predictive model to classify loan application into two categories: 1 (loan application is accepted) and 2 (loan application is rejected).

The model was created using the lightGBM machine learning algorithm with binary objective and trained with all the data yielded the following results:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 170355  56161
##          1  79531 193953
##
##                Accuracy : 0.7286
##                  95% CI : (0.7274, 0.7298)
##     No Information Rate : 0.5002
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4572
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.7755
##             Specificity : 0.6817
##          Pos Pred Value : 0.7092
##          Neg Pred Value : 0.7521
##              Prevalence : 0.5002
##          Detection Rate : 0.3879
##    Detection Prevalence : 0.5470
##       Balanced Accuracy : 0.7286
##
##        'Positive' Class : 1
##
```



Area under the curve: 0.731

Accuracy: 72.8%

Precision: 75.2%

Recall: 68.1%

F1 Score: 74.1%

CONCLUSION

This analysis has shown that the loan application of home mortage applicants can be confidently predicted from its characteristics. In particular, the lender accpetance rate, applicant income, mean applicant income per lender, unique counts of lender, loan amount, and loan purpose have a significant effect on the loan acceptance for each applicant and can help further classify loan application either to accept or reject the application.