# 2015 Texas HMDA Data Analysis

STAT 605, Group 11: Yuqi Wang, Chaoqi Ye, Yulin Li, Wenlu Dong, Zhuoer Xu

December 2, 2021

## Contents

# 1 Introduction

This report gives an overview of the analysis of 2015 Texas The Home Mortgage Disclosure Act(HMDA) Data that covers information provided by 6913 U.S financial institutions to show whether lenders are serving the housing needs of their communities and shed light on lending patterns that could be discriminatory.

The aim of the analysis is to find patterns in the lending process, and give some suggestion for improving the possibilities of loans approval.

To answer these questions, we explore the impact of each variable on the loan market by analyzing the relationship between the variables in HMDA and the approved-rate of mortgage loans. First, we analyze the differences in the number of loans and loan-success rates in different cities. Through doing plots and Chi-square test, we think that the loan approval rate may be related to county, income, race, gender, Ethnicity, loan type, loan purpose, and the integrity of the information providedetc. Then we used SQL to filter and group the statements about the purpose of home improvement in our second data set. Finally, we tried to apply some knowledge in machine learning, thus we fitted the model by using the logistics regression method.

Chapter 2 of this report gives an introduction of the data set and an explanation of some variables of it. Chapter 3 presents the relationship between several variables(county, lien status, loan purpose, property type, applicant income, etc) and the approval rate of mortgage loans by some plots and math methods. In chapter 4, we introduce secondary data set to analyze which part of house people are more likely to improve and focus on home improvement by using SQL in R, we also apply a machine learning Algorithm to our data.

# 2 Data Preparation

For main data set(2015 HMDA loan information of Texas ), we analyze this data set to find patterns in the lending process, and give some suggestion for improving the possibilities of loans approval. For secondary data set(2007 through 2018 lending club data), we only focus on the loan purpose, state, and description and use SQL and regular expression to analyze the data.

## 2.1 Main data set

The dimension of main data set is (1139573,78), not every single variable of them is useful. Some important variables can be grouped into the following subjects:

- Location describes the county, metro area and census tract of the property.
- Loan describes the action taken on the loan, loans lien status, purpose of the loan, type of the loan, loan amount.
- Property Type describes the property type and cccupancy of the property.
- Applicant describes the demographic information for the applicants. This has the applicant gender, race, ethnicity and income.

In order to help readers to understand the report better, the used variables are explained in detail as following:

```
##  action_taken_name  county_name        lien_status_name   property_type_name
##  Length:827878      Length:827878      Length:827878      Length:827878
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  applicant_income_000s loan_amount_000s  applicant_ethnicity_name
```

```
##  Min.   :   1.0        Min.   :    1.0   Length:827878
##  1st Qu.:  54.0        1st Qu.:   91.0   Class :character
##  Median :  84.0        Median :  150.0   Mode  :character
##  Mean   : 111.7        Mean   :  197.4
##  3rd Qu.: 130.0        3rd Qu.:  229.0
##  Max.   :9999.0        Max.   :99999.0
##  NA's   :72013
##  applicant_race_name_1 applicant_sex_name denial_reason_name_1
##  Length:827878         Length:827878      Length:827878
##  Class :character      Class :character   Class :character
##  Mode  :character      Mode  :character   Mode  :character
##
##
##
##
```

The lending process ends when the loan is approved or denied, so we firstly categorized data based on the variable *action_taken_name*, this variable includes 8 categories: Loan originated; Application approved but not accepted; Application denied by financial institution; Application withdrawn by applicant; File closed for incompleteness; Loan purchased by the institution; Pre-approval request denied by financial institution; Pre-approval request approved but not accepted.

| action_taken_name | n |
|---|---|
| Loan originated | 557266 |
| Application denied by financial institution | 188621 |
| File closed for incompleteness | 46331 |
| Application approved but not accepted | 35576 |
| Preapproval request approved but not accepted | 54 |
| Preapproval request denied by financial institution | 30 |

Table 1: Count number of action taken name

Explanation: "Loan originated" is a commonly used term in finance and refers to the loan application that has been approved. The primary market is where securities are created, while the secondary market is where those securities are traded by investors. In this project, we just focus on the loan in the primary market, which means the borrowers did not withdraw their application or purchase loan by financial institution(this is seen as loan in the secondary market). So we remove the rows where the applicant has withdrawn and the rows where the action taken is "Loan purchase by financial institution". Then, named loan originated as success; the others are categorized as fail.

## 2.2 Secondary data set

This data set(*accepted_2007_to_2018Q4*) contains the full LendingClub data available from their website. There are separate files for accepted and rejected loans, the accepted loans also include the FICO scores.Since the desc column contains the description of loan purpose, we can filter the part that is about mortgage loan as supplementary data to analysis.

# 3 Mian data set analysis

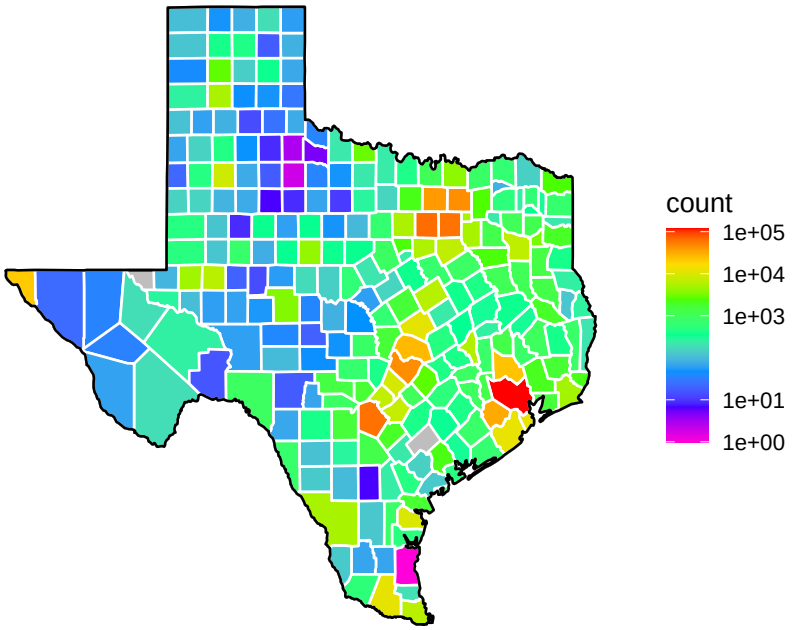
## 3.1 County

*County_name* lists which county does the applicant come from.

| subregion | count |
|---|---|
| harris | 118648 |
| tarrant | 64195 |
| bexar | 60706 |
| dallas | 60125 |
| travis | 45125 |
| kent | 7 |
| mcmullen | 7 |
| foard | 5 |
| cottle | 3 |
| king | 2 |
| kenedy | 1 |

Table 2: Count number of loan applications of counties

We can see the top5 county with the largest number of loan applications and the bottom five county with the smallest number of loan appliances. We could see that Harris county has the largest number of loan application and Kenedy county has the smallest number of loan applications. Considered this part is related to location, so we will use ggmap to make a map, that will make the result visualized.




We can see that the counties in the east and southeast of Texas has more loan applicants than other regions. And we notice that the number of loan applications varies a lot in different county. So we imported the population data of Texas and considered the influence of population.

Which county has the highest application rate?

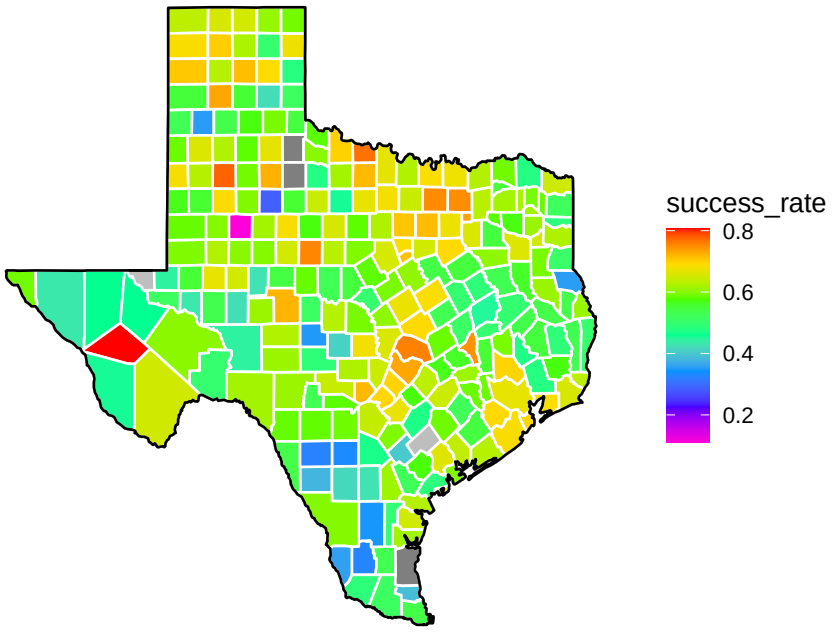
| subregion | application_rate |
|---|---|
| rockwall | 0.0577668 |
| comal | 0.0547519 |
| williamson | 0.0512041 |
| denton | 0.0494951 |
| collin | 0.0487974 |

Table 3: The application rate of counties

Except exploring the highest number of loan application, we will also see which county has the highest percentage of loan approve and explore the potential reason based on the result.

| county | fail | success | success_rate |
|---|---|---|---|
| jeff davis | 7 | 29 | 0.8055556 |
| lubbock | 1804 | 6207 | 0.7748096 |
| wichita | 681 | 2232 | 0.7662204 |
| williamson | 6338 | 19644 | 0.7560619 |
| taylor | 1026 | 3127 | 0.7529497 |
| denton | 9652 | 28933 | 0.7498510 |

Table 4: The condition of loan approval of counties



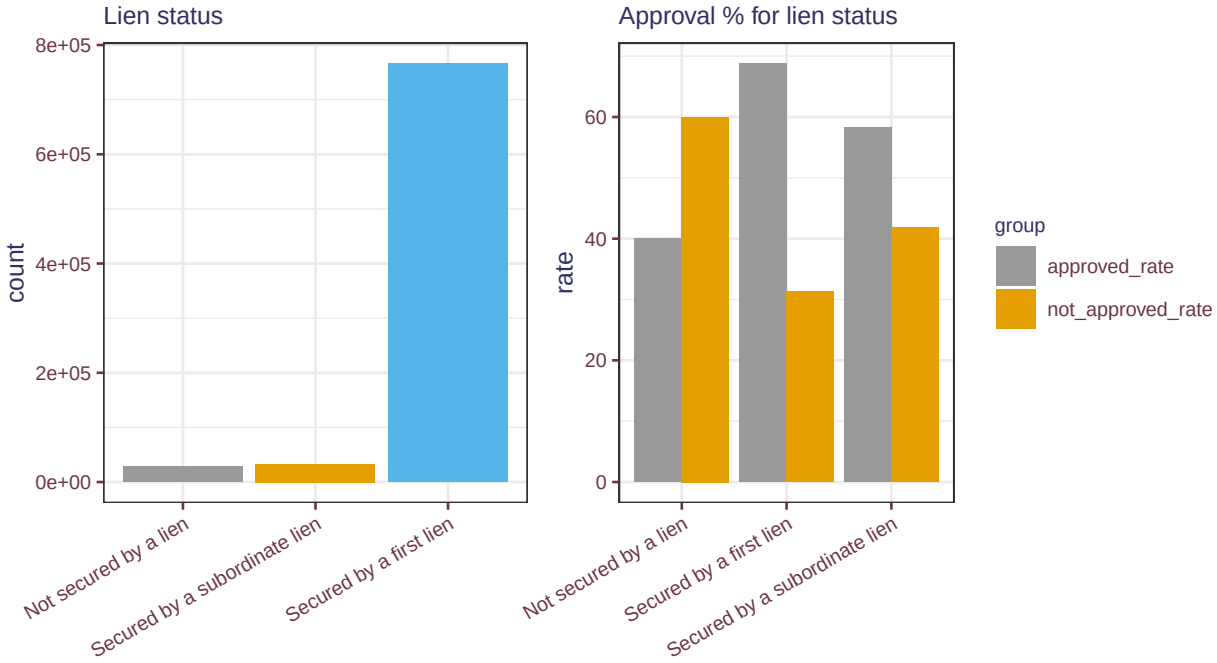

We notice that Jeff Davis county has the highest approval rate of 80.6 percentage. And the top five counties wit the highest loan-approved rate don't have large number of loan applicants. And we can see that counties in northwest and west of Texas have higher loan-Approved race. In our previous analysis we can see that the number of loan applications in these areas is small. Thus we speculate that the loan-approved rate may be related to the total number of loan applicants in the region.

## 3.2 Lien status, loan purpose and property type

### 3.2.1 Lien status

A lien serves to guarantee an underlying obligation of the repayment of a loan. If the underlying obligation is not satisfied, the creditor(lender) may be able to seize the asset that is the subject of the lien.Once executed, a lien becomes the legal right of a creditor to sell the collateral property of a debtor who fails to meet the obligations of a loan or other contract. Most mortgages are secured by a lien against the property. In the event of a forced liquidation, first lien holders will generally get paid before subordinate lien holders.

The variable *lien_status_name* indicates the type of lien status. There 3 categories included: secured by a first lien, secured by a subordinate lien, not secured by a lien. We produce the plot of count number of different categories and the plot of the approved rate of related category to explore the relation of lien status and approval rate of mortgage loan.

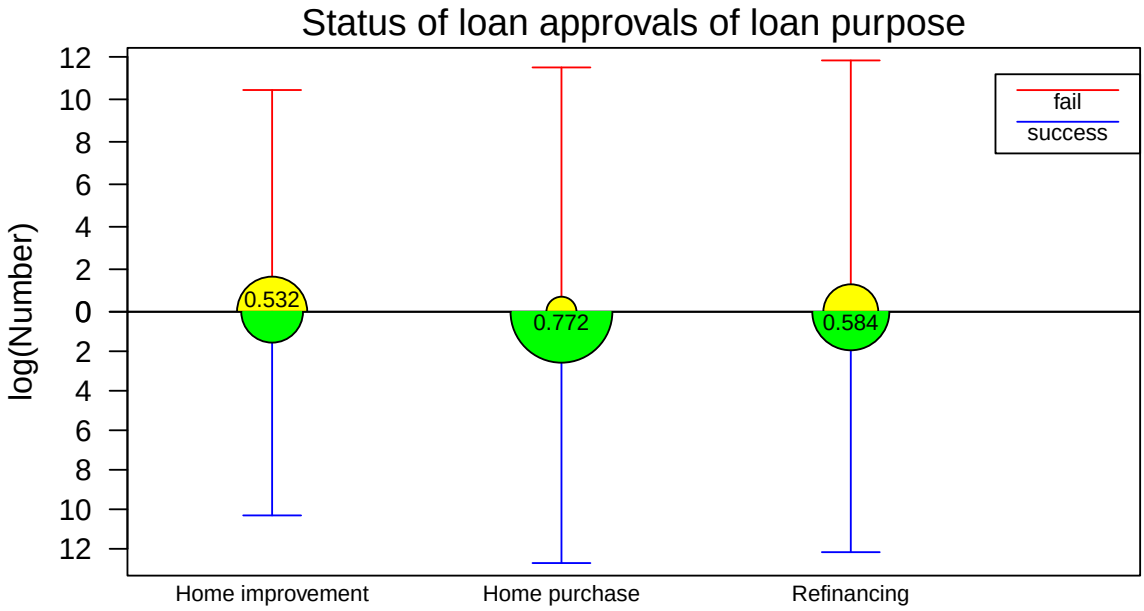



Most of the loans are secured by a first lien. Loans secured by a first lien get approved 68.7 percentage of the time. Loans not secured by a lien have the highest percentage of not getting approved, which is up to 60 percentage.

### 3.2.2 Loan purpose

The variable *loan_purpose_name* identifies what aspect does the applicants want to use the loan for, including 3 aspects: home purchase, home improvement, refinancing.

Here, we produce the original killer plot to display the relation of loan purpose and approval rate of the loan. The killer plot can display the count number of categories that are binary and their related probability together. In the plot, the height of lines represents the count number(or log(number)) of categories, the area or the radius of semicircle represents the probability of related binary category. Note, When we use log(number) as y scale, even the difference of height of lines is small, the true different is very large.

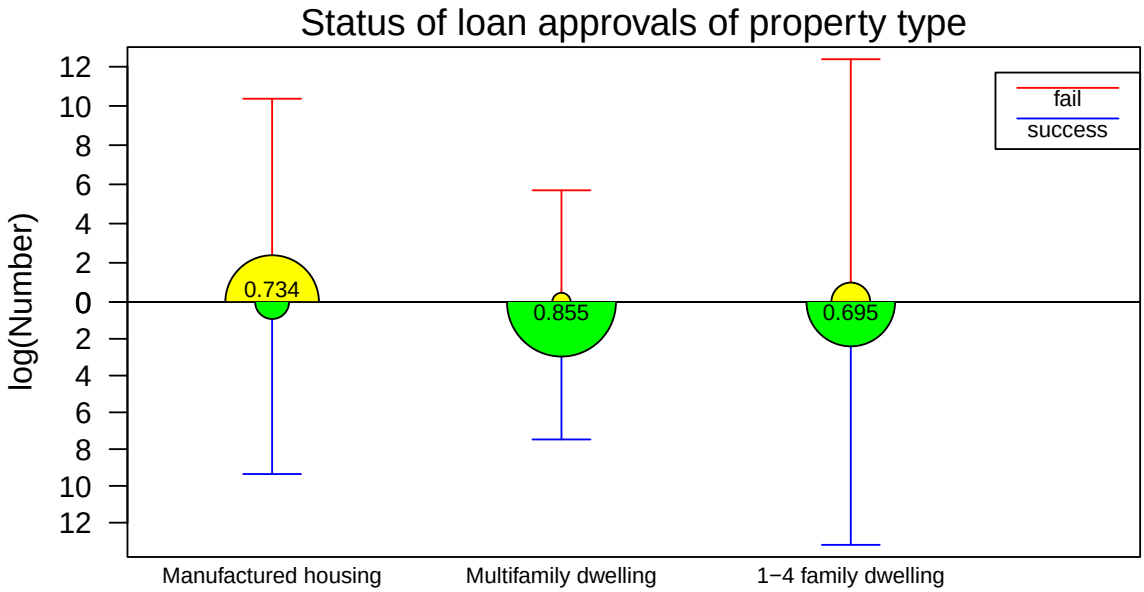



Those recordings show most people took out a mortgage loan to purchase home, a minority applied the loan to home improvement in 2015. In addition, most of the home purchase loans are approved while it's less likely for a refinancing loan or a home improvement loan to get approved.

### 3.2.3 Property type

The variable *property_type_name* indicates the type of properties. There 3 categories included: one to four-family, manufactured housing, multifamily.

Here, we produce the original killer plot to display the relation of property type and approval rate of the loan.

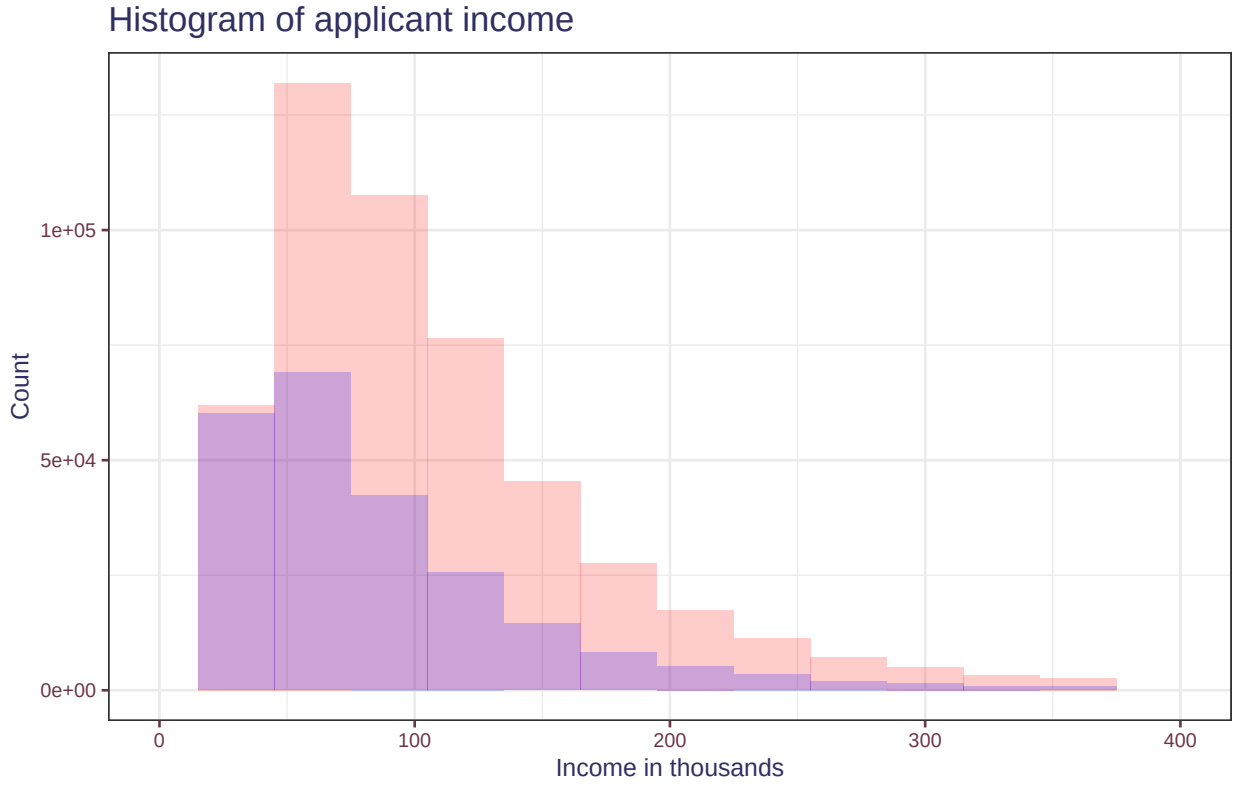


Multifamily-dwelling is a safe choice when going for house loans.Getting a higher residential mortgage on a multifamily dwelling is easier based on the rental income generated,which can cover or reduce the mortgage.Its tougher to get a loan for manufactured housing. Many manufactured home loan programs have strict guidelines about the property condition and age. Thats because manufactured housing tends to depreciate, while traditional home values tend to increase over time.

### 3.3 Income and loan amount

The variable *applicant_income_*000*s* and *loan_amount_*000*s* respectively represents the income of applicants and the amount of money that applicants get, both are in thousands of dollars.

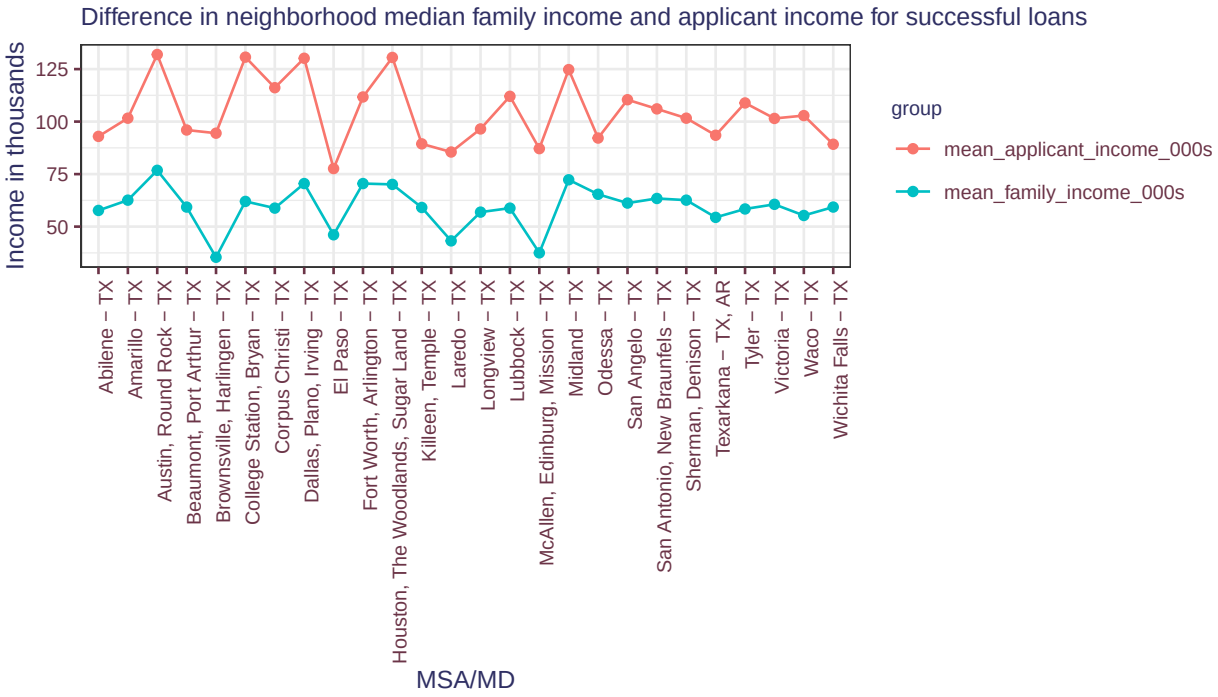
In this section, we first compare the histograms of applicants' income for successful and unsuccessful loans. And then, we draw two scatter plots to check the difference between neighborhood median family income and applicant income for successful and unsuccessful loans, respectively. At last, we use box-plot to show the loan amount pattern for people in different income categories.



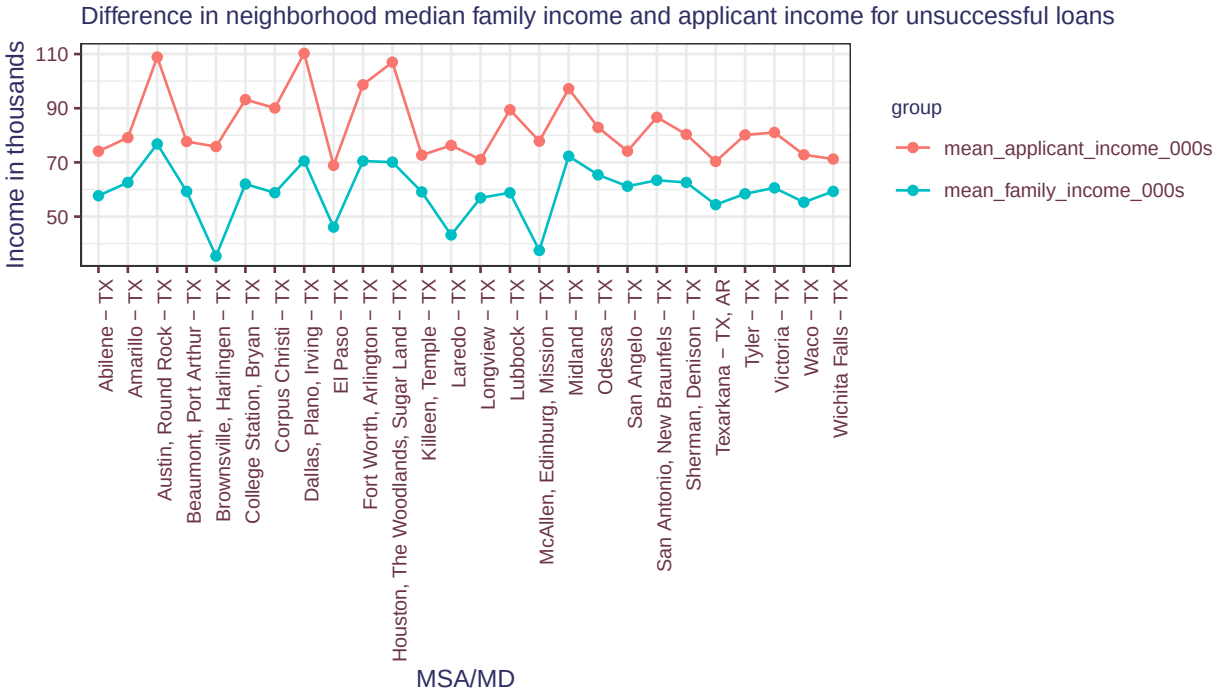

We can see that the applicants' income for successful loans (red) is generally larger than the income for unsuccessful loans (blue). Besides, both the histograms show that the distributions are positively skewed.

Does the neighborhood family income affect applicants' chances of getting a loan?

Instead of disclosing the address, lenders disclose the census tract, which is part of the community where the property is located. Each census tract is located in a Metropolitan Statistical Area/Metropolitan Division (MSA/MD). The *hud_median_family_income* is the median family income in dollars for the MSA/MD in which the tract is located.



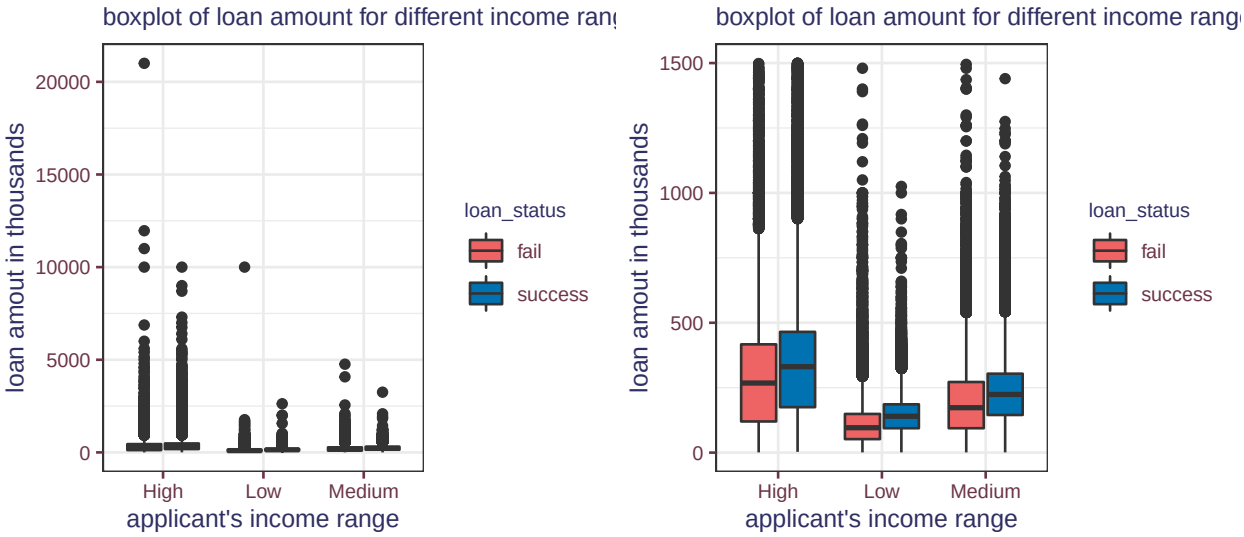

For a loan to be approved, the applicant's income is above the neighborhood median family income.




What's very unsettling is that for the loans not approved the applicant's income is still greater than the neighbourhood median family income.

What is the loan amount pattern for applicants in different income categories?

To explore this, we first categorize people as falling in the low (for applicant income less than 100k), middle(between 100k and 200k), and high(more than 200k) income range.

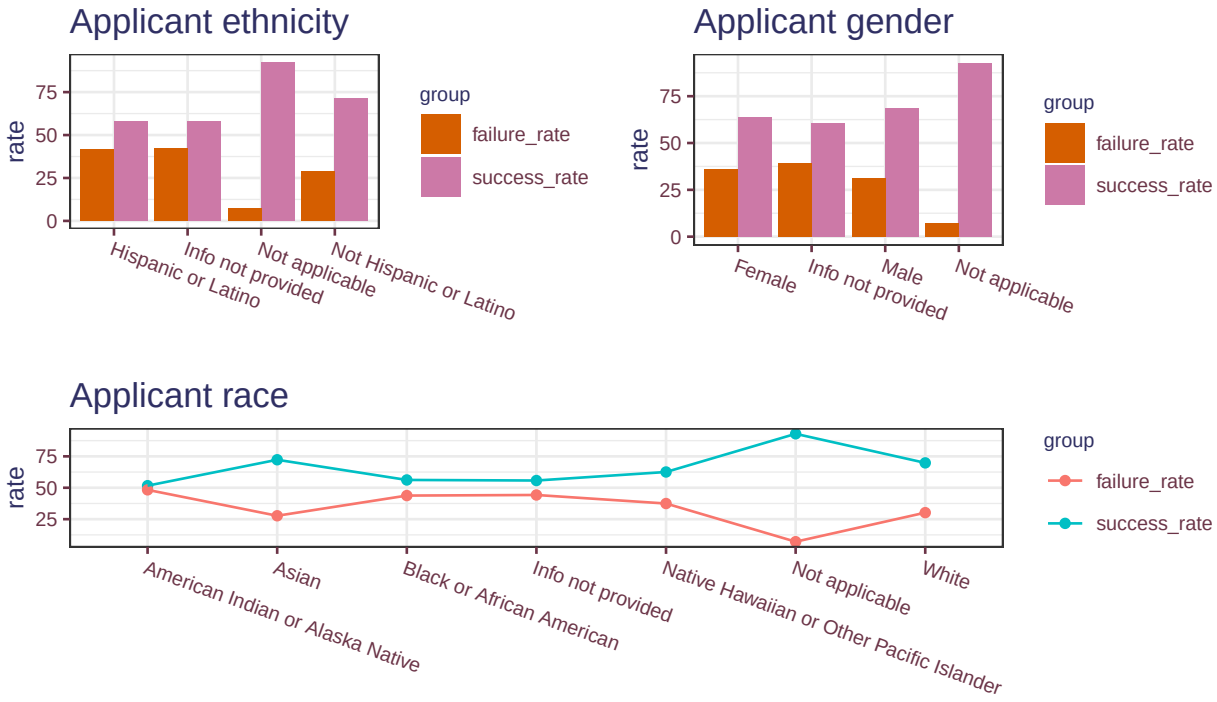



We can see that the left plot is very compressed. The reason for this is that loans with very high loan amounts are always rejected for the low and medium salary range applicant; while they may get accepted for applicants falling in the high-income range.

For the right boxplot, we limit the loan amount to less than 1,500,000. In all the income ranges, the successful loans have requested loan amounts greater than unsuccessful loans. A usual trend, people generally request loan amounts proportionate to their incomes. In most cases, applications where the loan amount far exceeds the applicant's income get rejected.

### 3.4 Ethnicity, gender and race

In this section, we use the bar plots and scatter plots to show the relationship between loan status and applicant ethnicit, gender and race. Besides, we use the Pearson Chi-squared test to check the independence between loan status and applicant ethnicity and gender respectively.



There isn't much difference between success rates for females and males, however, there seems to be a little discrimination based on ethnicity and race. We will compute the Pearson Chi-squared test statistic to check the independence hypothesis next.

The two-way contingency tables for gender and loan status, ethnicity and loan status, And ethnicity and loan status are displayed as below.

| | Female | Male |
|---|---|---|
| fail | 78328 | 169111 |
| success | 138248 | 373918 |

Table 5: loan approval condition of gender

| | Hispanic or Latino | Not Hispanic or Latino |
|---|---|---|
| fail | 74378 | 155954 |
| success | 103379 | 388586 |

Table 6: The condition of loan approval of ethnicity

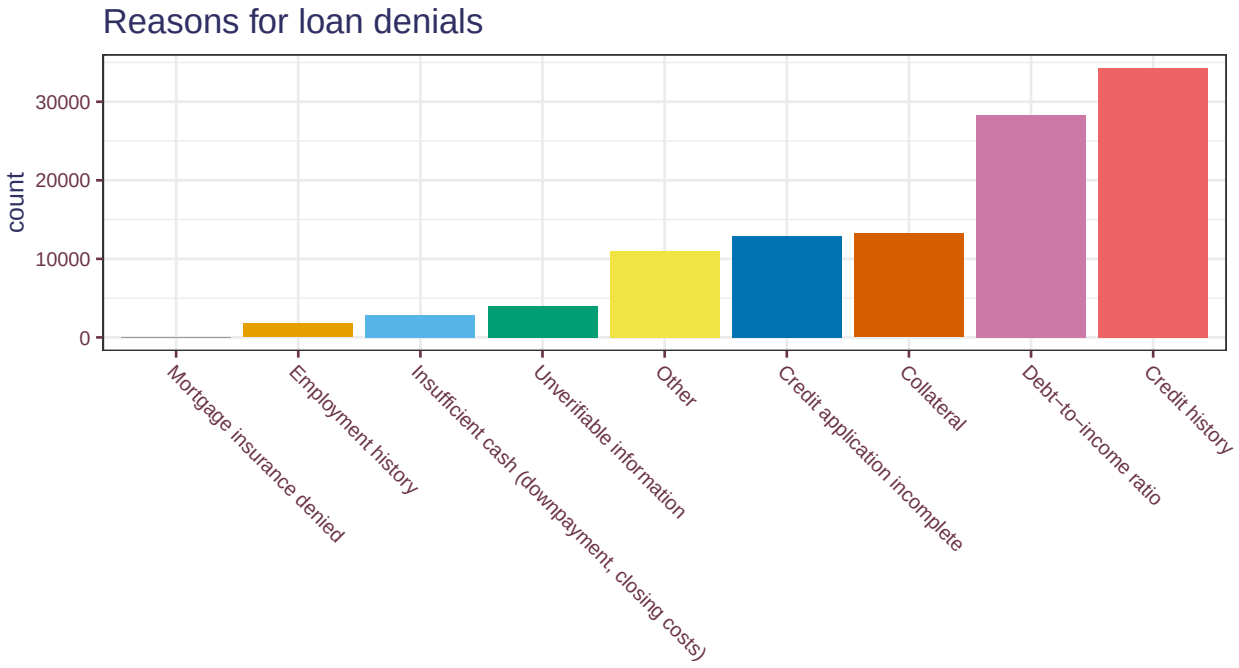
| | American Indian | Asian | Black | No info | Native Hawaiian | White |
|---|---|---|---|---|---|---|
| fail | 3232 | 13489 | 27373 | 46230 | 976 | 178490 |
| success | 3454 | 35180 | 35139 | 58261 | 1630 | 412792 |

Table 7: The condition of loan approval of race

The corresponding test statistics $X^2$'s equal 1779.3, 10755, and 12719, respectively. The first two test statistics follow the Chi-squared distribution with degree of freedom 1; the third test statistic follows the Chi-squared distribution with degree of freedom 5. All three tests have p-values $\approx 0$, therefore we reject the null hypothesis and none of the three variables is independent of the loan status.

## 3.5 Major reason of denial

In this section, we draw the bar-plot of the reasons for loan denials. Then, we make a summary for chapter 3 analysis.



Mostly applications are denied for poor credit history or high debt-to-income ratio. The most easily avoidable reason is incomplete credit application. Though a total of 12872 applications are denied because of it.

### 3.6 conclusion

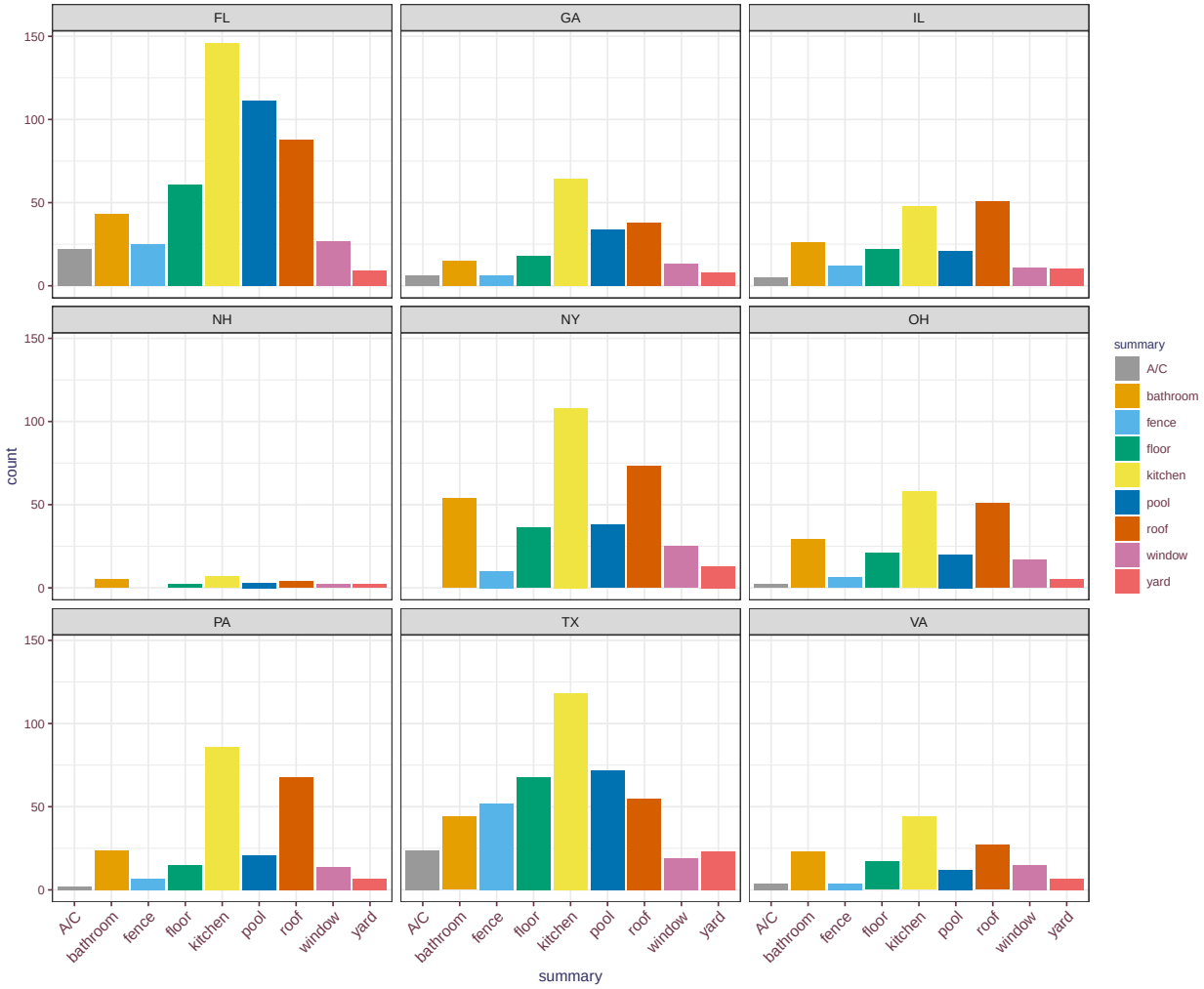
To increase the chances of getting loan success in Texas, applicants must keep the following points in mind:

- Get a home loan in Jeff Davis County (highest approval rate).
- Application for a home purchase loan has a better chance at getting success.
- In whichever MSA/MD you are buying a home, your income should be greater than or equal to the median family income of that MSA/MD.
- Apply for a multifamily or 1-4 family dwelling.
- Get your loan secured by a lien preferably by a first lien.
- Maintain good credit history and apply for loans proportional to your income.

# 4 Secondary data set analysis and logistic regression model

## 4.1 Home improvement analysis with secondary data set

We choose Taxes and other 8 counties which the applicant numbers are the most in the data set to see which part of house are more likely to be improved.



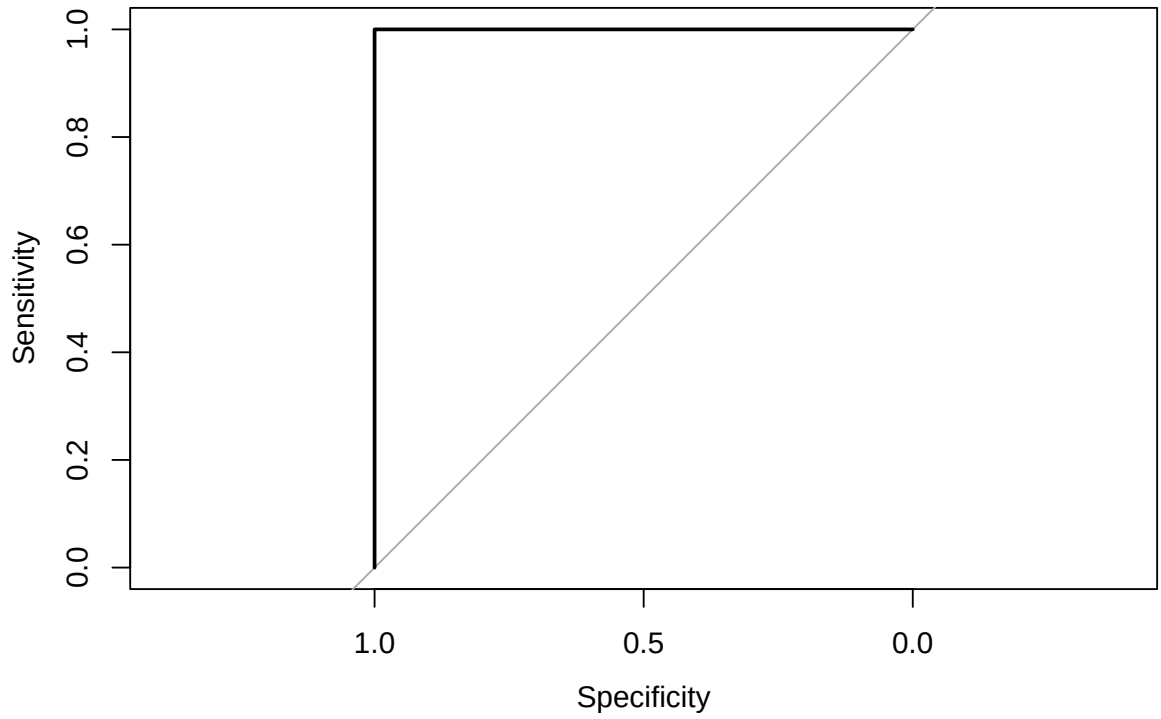

The plot indicates that for most counties, kitchen is the most popular part of house to be improved.

## 4.2 Logistic regression model

In this section, we choose logistical regression method to make the model.

At the very beginning, we select some variables from our original data, and change the categorical variables into factor (data type). And then we split the data into two data sets (train data and test data) to fit logistical model. In these two data sets, We use train data set to train the logistical regression model, and test set is used to evaluate model and convergences state.

After using the train set to apply in logistical regression model, we can get a fitted generalized model, which will be used to predict and see if it is Approved or Disapproved by using the threshold 0.5 to determine. At last, we get a ROC curve.




Because

$$FalsePositiveRate = Specificity = \frac{TrueNegatives}{TrueNegatives + FalsePositives} \tag{1}$$

$$TruePositiveRate = Sensitivity = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{2}$$

we can know from this ROC curve that logistical regression model has a great performance because TPR is higher while FPR is lower in value, at this time the logistical regression model has a great performance.

# 5 Conclusion

In summary, we conclude some patterns of mortgage loans, which may improves the approval rate of applicants.

For properties of loans:

- Apply a "home purchase" loan.
- Apply for a multifamily or 1-4 family dwelling.
- Get the loan secured by a lien, preferably by a first lien.

For applicants:

- Get a home loan in Jeff Davis County (highest approval rate).
- Income should not less than the median family income of that MSA/MD.
- Apply for loans proportional to the income.
- Maintain good credit history.