# Exploratory Data Analysis Home Mortgage NY

Code

## 1. Introduction

The Home Mortgage Disclosure Act (HMDA) requires many financial institutions to maintain, report, and publicly disclose information about mortgages.This dataset covers all mortgage decisions made in 2015 for the state of New York.

## 2. Understanding the problem - Business Perspective

Before we dive into solve the problem, let us first understand the business related to this dataset.

## 2.1 What is HMDA ?

Each year thousands of banks and other financial institutions report data about mortgages to the public, thanks to the Home Mortgage Disclosure Act, or “HMDA” for short. These public data are important because:

–Help show whether lenders are serving the housing needs of their communities; –Give public officials information that helps them make decisions and policies; and –Shed light on lending patterns that could be discriminatory

## 2.2 Loan Origination Journey

My friend Rose wants to buy a home but she doen’t have enough money to pay in cash, so she applies for a loan at her bank. Bank collects all the information related to her finances and the property she is wiiling to buy.These information helps the bank to make a decison whether or not to lend her money, and the terms of the loan. The bank reviews Rose’s background and decides that she mets their criteria, and she gets approved.Once all the papers are signed, Rose closes the loan. or in mortgage-speak, the loan is “originated.”.Therefore the last stage of the loan is Loan Origination.

## In the following steps we will learn how to build a classification tree to find out the deciding variables or the most important variables on which a loan application depends.

Step1- It includes loading the dataset and reading.

Hide

```
hmda<- read.csv('D:/Rutgers Study Material/Rutgers Study Material/DADM/Project/ny-home-mortgage/
ny_hmda_2015.csv')
# Just to check if dataset is properly loaded or not, we will use head
head(hmda,2)
```

| | action_taken <int> | action_taken_name <fctr> | agency_code <int> | agency_abbr <fctr> | agency_name <fctr> |
|---|---|---|---|---|---|
| 1 | 1 | Loan originated | 9 | CFPB | Consumer Financial Protection Bure |
| 2 | 1 | Loan originated | 9 | CFPB | Consumer Financial Protection Bure |

2 rows | 1-6 of 78 columns

Hide

```
mortgage<-head(hmda,45000)
```

Step2- Gaining some insights about the data.

Hide

```
# 'names' will return all the column names in the dataset
names(mortgage)

# we will look at the structure and dimension of the dataset
str(mortgage)
dim(mortgage)
```

There are 78 coulmns and 45000 rows of data.

Step3- This is the most important step, data cleansing.

Hide

```
#Function to check if there are any NA values in the dataset
sapply(mortgage,function(x) sum(is.na(x)))
```

Lots of NA values are recorded.

Hide

```
# The most up-to-date method for handling missing data is to use multiple imputations.
# Load the mice package
library(mice)
# pattern of missing data
md.pattern(mortgage)
```

Hide

```
# multiple imputations
imp <- mice(mortgage, m=1, maxit=2, method='cart', seed=500)
```

```
 iter imp variable
  1   1  applicant_income_000s  census_tract_number  msamd  hud_median_family_income  number_of_
1_to_4_family_units  number_of_owner_occupied_units  minority_population  population  rate_sprea
d  tract_to_msamd_income
  2   1  applicant_income_000s  census_tract_number  msamd  hud_median_family_income  number_of_
1_to_4_family_units  number_of_owner_occupied_units  minority_population  population  rate_sprea
d  tract_to_msamd_income
```

Hide

```
# completed data
home <- mice::complete(imp)
```

Here we use ‘cart’(classification and regression trees) as the imputation method. Now R does not need to do any X matrix inversion.

Hide

```
summary(home)
```

```
 action_taken
Min.   :1
1st Qu.:1
Median :1
Mean   :1
3rd Qu.:1
Max.   :1

                                        action_taken_name
Application approved but not accepted              :    0
Application denied by financial institution        :    0
Application withdrawn by applicant                 :    0
File closed for incompleteness                     :    0
Loan originated                                    :45000
Loan purchased by the institution                  :    0
Preapproval request denied by financial institution:    0
 agency_code   agency_abbr
Min.   :1.00   CFPB:15885
1st Qu.:5.00   FDIC: 1087
Median :7.00   FRS : 1829
Mean   :6.55   HUD :13424
3rd Qu.:9.00   NCUA: 9529
Max.   :9.00   OCC : 3246

                                     agency_name
Consumer Financial Protection Bureau       :15885
Department of Housing and Urban Development:13424
Federal Deposit Insurance Corporation      : 1087
Federal Reserve System                     : 1829
National Credit Union Administration       : 9529
Office of the Comptroller of the Currency  : 3246

applicant_ethnicity
Min.   :1.000
1st Qu.:2.000
Median :2.000
Mean   :2.081
3rd Qu.:2.000
Max.   :4.000

                                                                                applicant_ethnicity_name
Hispanic or Latino                                                                             : 2253
Information not provided by applicant in mail, Internet, or telephone application: 4247
Not applicable                                                                                 :  816
Not Hispanic or Latino                                                                         :37684



applicant_income_000s applicant_race_1 applicant_race_2
Min.   :   1.0        Min.   :1.000    Min.   :1.00
1st Qu.:  60.0        1st Qu.:5.000    1st Qu.:5.00
Median :  93.0        Median :5.000    Median :5.00
Mean   : 161.5        Mean   :4.809    Mean   :4.42
```

```
3rd Qu.: 149.0        3rd Qu.:5.000    3rd Qu.:5.00
Max.   :9999.0        Max.   :7.000    Max.   :5.00
                                       NA's   :44853
applicant_race_3 applicant_race_4 applicant_race_5
Min.   :1.00     Min.   :4        Min.   :5
1st Qu.:3.00     1st Qu.:4        1st Qu.:5
Median :4.00     Median :4        Median :5
Mean   :3.67     Mean   :4        Mean   :5
3rd Qu.:5.00     3rd Qu.:4        3rd Qu.:5
Max.   :5.00     Max.   :4        Max.   :5
NA's   :44994    NA's   :44998    NA's   :44998
                                                                                applicant_race_name_1
American Indian or Alaska Native                                                              :  134
Asian                                                                                         : 3108
Black or African American                                                                     : 2282
Information not provided by applicant in mail, Internet, or telephone application: 4358
Native Hawaiian or Other Pacific Islander                                                     :  118
Not applicable                                                                                :  788
White                                                                                         :34212
                              applicant_race_name_2
                                         :44853
American Indian or Alaska Native         :    4
Asian                                    :   12
Black or African American                :   13
Native Hawaiian or Other Pacific Islander:    7
White                                    :  111

                              applicant_race_name_3
                                         :44994
American Indian or Alaska Native         :    1
Asian                                    :    0
Black or African American                :    2
Native Hawaiian or Other Pacific Islander:    0
White                                    :    3

                              applicant_race_name_4
                                         :44998
Asian                                    :    0
Native Hawaiian or Other Pacific Islander:    2
White                                    :    0




                              applicant_race_name_5
                                         :44998
American Indian or Alaska Native         :    0
Native Hawaiian or Other Pacific Islander:    0
White                                    :    2



applicant_sex
Min.   :1.000
1st Qu.:1.000
```

```
Median :1.000
Mean   :1.493
3rd Qu.:2.000
Max.   :4.000

                                                                                applicant_sex_name
Female                                                                                      :14200
Information not provided by applicant in mail, Internet, or telephone application: 2818
Male                                                                                        :27199
Not applicable                                                                              :  783



application_date_indicator   as_of_year   census_tract_number
Min.   :0                  Min.   :2015   Min.   :   1
1st Qu.:0                  1st Qu.:2015   1st Qu.: 122
Median :0                  Median :2015   Median : 235
Mean   :0                  Mean   :2015   Mean   :1355
3rd Qu.:0                  3rd Qu.:2015   3rd Qu.:1228
Max.   :0                  Max.   :2015   Max.   :9811

co_applicant_ethnicity
Min.   :1.000
1st Qu.:2.000
Median :5.000
Mean   :3.661
3rd Qu.:5.000
Max.   :5.000

                                                                          co_applicant_ethnicity_name
Hispanic or Latino                                                                          : 1016
Information not provided by applicant in mail, Internet, or telephone application: 2181
No co-applicant                                                                             :24456
Not applicable                                                                              :  106
Not Hispanic or Latino                                                                      :17241



co_applicant_race_1 co_applicant_race_2 co_applicant_race_3
Min.   :1.000       Min.   :1.00        Min.   :3.00
1st Qu.:5.000       1st Qu.:4.00        1st Qu.:4.00
Median :8.000       Median :5.00        Median :5.00
Mean   :6.556       Mean   :4.35        Mean   :4.33
3rd Qu.:8.000       3rd Qu.:5.00        3rd Qu.:5.00
Max.   :8.000       Max.   :5.00        Max.   :5.00
                    NA's   :44946       NA's   :44997
co_applicant_race_4 co_applicant_race_5
Min.   :4           Min.   :5
1st Qu.:4           1st Qu.:5
Median :4           Median :5
Mean   :4           Mean   :5
3rd Qu.:4           3rd Qu.:5
Max.   :4           Max.   :5
NA's   :44999       NA's   :44999
                                                                            co_applicant_race_name_1
```

```
No co-applicant                                                                 :24456
White                                                                           :16024
Information not provided by applicant in mail, Internet, or telephone application: 2258
Asian                                                                           : 1387
Black or African American                                                       :  691
Not applicable                                                                  :   86
(Other)                                                                         :   98
                            co_applicant_race_name_2
                                           :44946
American Indian or Alaska Native           :    1
Asian                                      :    7
Black or African American                  :    3
Native Hawaiian or Other Pacific Islander:    4
White                                      :   39

                            co_applicant_race_name_3
                                           :44997
American Indian or Alaska Native           :    0
Asian                                      :    0
Black or African American                  :    1
Native Hawaiian or Other Pacific Islander:    0
White                                      :    2

                            co_applicant_race_name_4
                                           :44999
Native Hawaiian or Other Pacific Islander:    1
White                                      :    0




co_applicant_race_name_5 co_applicant_sex
     :44999              Min.   :1.000
White:    1              1st Qu.:2.000
                         Median :5.000
                         Mean   :3.553
                         3rd Qu.:5.000
                         Max.   :5.000

                                                                      co_applicant_sex_name
Female                                                                          :13940
Information not provided by applicant in mail, Internet, or telephone application: 1442
Male                                                                            : 5076
No co-applicant                                                                 :24456
Not applicable                                                                  :   86


 county_code              county_name    denial_reason_1
Min.   :  1.00   Nassau County  : 4011   Min.   : NA
1st Qu.: 53.00   Suffolk County : 3828   1st Qu.: NA
Median : 63.00   Monroe County  : 3574   Median : NA
Mean   : 66.58   New York County: 3178   Mean   :NaN
3rd Qu.: 85.00   Queens County  : 3110   3rd Qu.: NA
Max.   :123.00   Kings County   : 2659   Max.   : NA
```

```
NA's   :18       (Other)        :24640   NA's   :45000
denial_reason_2 denial_reason_3
Min.   : NA     Min.   : NA
1st Qu.: NA     1st Qu.: NA
Median : NA     Median : NA
Mean   :NaN     Mean   :NaN
3rd Qu.: NA     3rd Qu.: NA
Max.   : NA     Max.   : NA
NA's   :45000   NA's   :45000
                   denial_reason_name_1
                             :45000
Collateral                   :    0
Credit application incomplete:    0
Credit history               :    0
Debt-to-income ratio         :    0
Employment history           :    0
(Other)                      :    0
                   denial_reason_name_2
                             :45000
Collateral                   :    0
Credit application incomplete:    0
Credit history               :    0
Debt-to-income ratio         :    0
Employment history           :    0
(Other)                      :    0
                   denial_reason_name_3  edit_status
                             :45000     Min.   :6
Collateral                   :    0     1st Qu.:6
Credit application incomplete:    0     Median :6
Credit history               :    0     Mean   :6
Debt-to-income ratio         :    0     3rd Qu.:6
Employment history           :    0     Max.   :6
(Other)                      :    0     NA's   :37597
                 edit_status_name  hoepa_status
                         :37597   Min.   :1
Quality edit failure only: 7403   1st Qu.:2
                                  Median :2
                                  Mean   :2
                                  3rd Qu.:2
                                  Max.   :2

       hoepa_status_name  lien_status
HOEPA loan      :    7   Min.   :1.000
Not a HOEPA loan:44993   1st Qu.:1.000
                         Median :1.000
                         Mean   :1.156
                         3rd Qu.:1.000
                         Max.   :3.000

                     lien_status_name  loan_purpose
Not applicable               :    0   Min.   :1.000
Not secured by a lien        : 2563   1st Qu.:1.000
Secured by a first lien      :40563   Median :1.000
Secured by a subordinate lien: 1874   Mean   :1.765
```

```
                                    3rd Qu.:3.000
                                    Max.   :3.000

       loan_purpose_name   loan_type
Home improvement: 5239   Min.   :1.00
Home purchase   :25163   1st Qu.:1.00
Refinancing     :14598   Median :1.00
                         Mean   :1.23
                         3rd Qu.:1.00
                         Max.   :4.00

           loan_type_name      msamd
Conventional      :37103   Min.   :10580
FHA-insured       : 5938   1st Qu.:35004
FSA/RHS-guaranteed:  472   Median :35614
VA-guaranteed     : 1487   Mean   :34312
                           3rd Qu.:40380
                           Max.   :48060

                                      msamd_name
New York, Jersey City, White Plains - NY, NJ:14062
Nassau County, Suffolk County - NY          : 7834
Rochester - NY                              : 5676
                                            : 3554
Syracuse - NY                               : 3454
Albany, Schenectady, Troy - NY              : 3159
(Other)                                     : 7261
owner_occupancy
Min.   :1.0
1st Qu.:1.0
Median :1.0
Mean   :1.1
3rd Qu.:1.0
Max.   :3.0

                                owner_occupancy_name
Not applicable                                :  344
Not owner-occupied as a principal dwelling: 3797
Owner-occupied as a principal dwelling    :40859




 preapproval                          preapproval_name
Min.   :1.000   Not applicable               :32304
1st Qu.:2.000   Preapproval was not requested: 9325
Median :3.000   Preapproval was requested    : 3371
Mean   :2.643
3rd Qu.:3.000
Max.   :3.000

property_type
Min.   :1.000
1st Qu.:1.000
```

```
Median :1.000
Mean   :1.032
3rd Qu.:1.000
Max.   :3.000

                                                    property_type_name
Manufactured housing                                         :  581
Multifamily dwelling                                         :  427
One-to-four family dwelling (other than manufactured housing):43992




purchaser_type
Min.   :0.000
1st Qu.:0.000
Median :1.000
Mean   :2.578
3rd Qu.:6.000
Max.   :9.000

                                                                        purchaser_type_name
Loan was not originated or was not sold in calendar year covered by register:18471
Fannie Mae (FNMA)                                                              : 6579
Freddie Mac (FHLMC)                                                            : 4427
Affiliate institution                                                          : 3551
Ginnie Mae (GNMA)                                                              : 3533
Commercial bank, savings bank or savings association                           : 3176
(Other)                                                                        : 5263
   respondent_id   sequence_number     state_code state_abbr
476810    : 4276   Min.   :      1   Min.   :36   NY:45000
451965    : 3471   1st Qu.:    707   1st Qu.:36
33-0941669: 2252   Median :   3234   Median :36
852218    : 2051   Mean   :  44690   Mean   :36
16-1566654: 1536   3rd Qu.:  30860   3rd Qu.:36
4735      : 1324   Max.   :1206812   Max.   :36
(Other)   :30090
   state_name    hud_median_family_income loan_amount_000s
New York:45000   Min.   : 57200           Min.   :    1.0
                 1st Qu.: 69000           1st Qu.:   93.0
                 Median : 71300           Median :  182.0
                 Mean   : 77286           Mean   :  320.2
                 3rd Qu.: 82700           3rd Qu.:  348.0
                 Max.   :109000           Max.   :83240.0

number_of_1_to_4_family_units number_of_owner_occupied_units
Min.   :   6                  Min.   :   5
1st Qu.:1006                  1st Qu.: 823
Median :1562                  Median :1252
Mean   :1530                  Mean   :1269
3rd Qu.:2024                  3rd Qu.:1679
Max.   :6345                  Max.   :6454

minority_population   population     rate_spread
```

```
Min.   :  0.34      Min.   :    1   Min.   : 1.500
1st Qu.:  6.69      1st Qu.: 3522   1st Qu.: 1.580
Median : 14.68      Median : 4616   Median : 1.730
Mean   : 24.48      Mean   : 4808   Mean   : 2.273
3rd Qu.: 30.77      3rd Qu.: 5926   3rd Qu.: 2.340
Max.   :100.00      Max.   :26588   Max.   :14.640

tract_to_msamd_income
Min.   :  8.31
1st Qu.: 91.34
Median :109.91
Mean   :122.27
3rd Qu.:135.61
Max.   :367.61
```

## 3.Understanding the problem - Data Perspective

The data provided can be grouped into the following subjects

Location describes the State, metro area and census tract of the property

Property Type describes the Property Type and Occupancy of the property.Property type values include One-to-four family dwelling,Manufactured housing and Multifamily dwelling. This also answers the question “Will the owner use the property as their primary residence ?” . The values include Owner occupied as principal dwelling , Not owner occupied as principal dwelling and Not Applicable.

Loan describes the action taken on the Loan, purpose of the Loan , Type of the loan ,Loan’s lien status.

Lender describes the lender associated with the loan and the Federal agency associated with the loan.

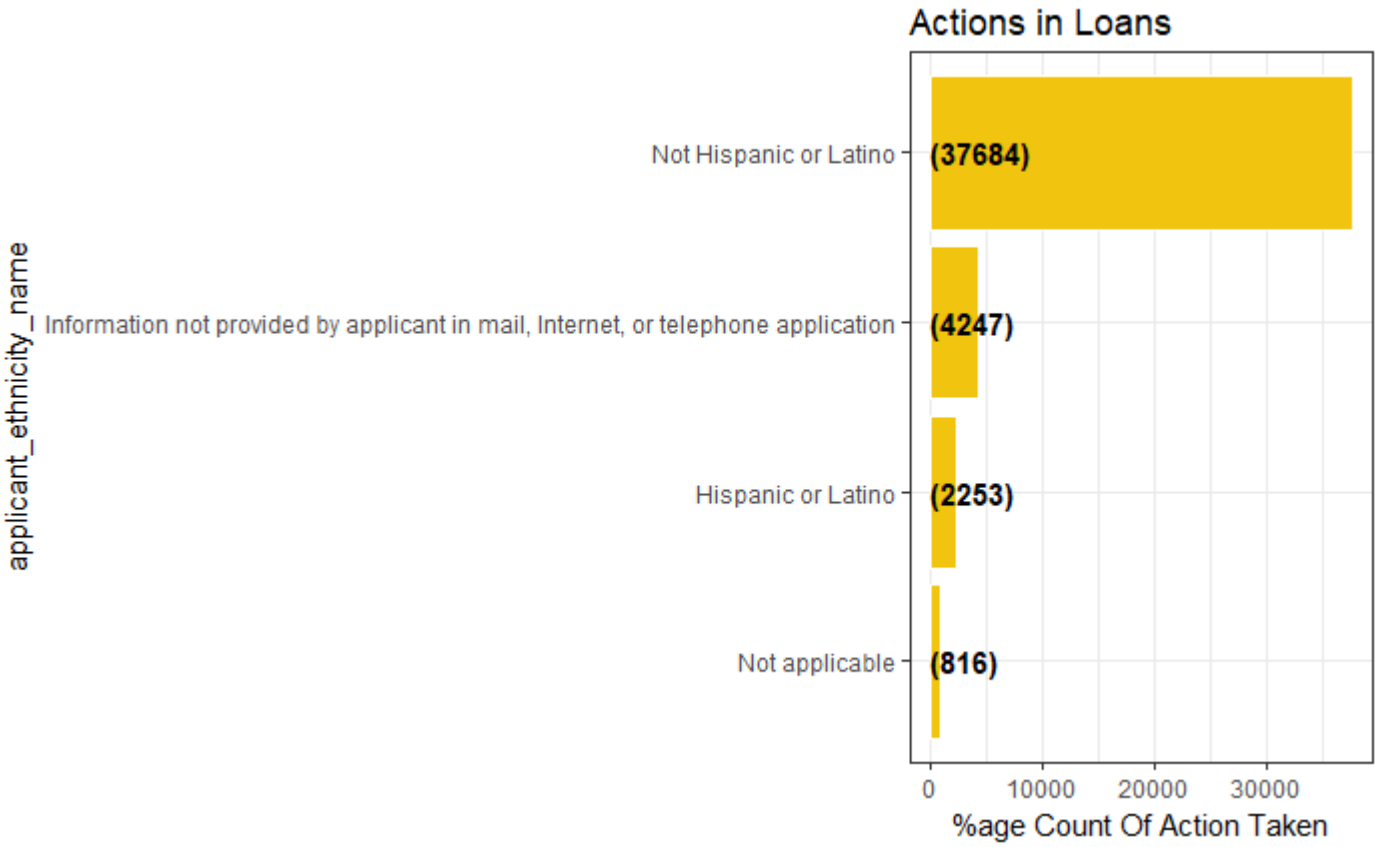
Applicant describes the demographic information for the applicants and the co-applicants.This has the applicant sex , co- applicant sex , applicant race and ethnicity, co- applicant race and ethnicity.

## Analyzing the data with the power of visualization

In this section, we examine the distribution of the various Actions on Loans. As discussed in the previous section, we would be interested in the loan action Loan Origination since this status signifies that the loan has been flagged off to be given to the applicant.

Hide

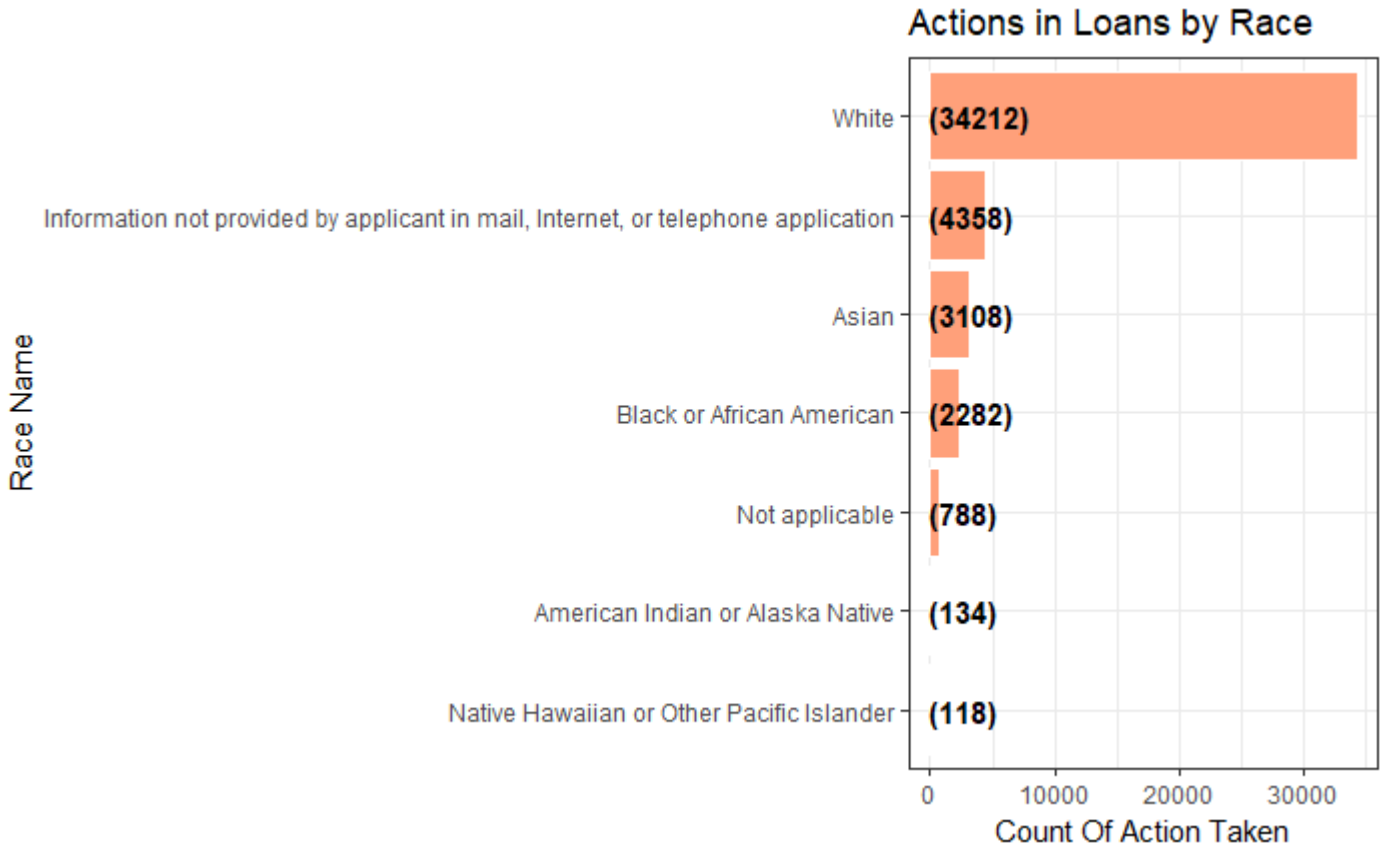
```
homeMortgageStatus_ethnicity = home %>% group_by(action_taken_name,applicant_ethnicity_name) %>%
  summarise(CountOfActionTaken = n()) %>%
  arrange(desc(CountOfActionTaken))
homeMortgage_ethnicity = home %>% group_by(applicant_ethnicity_name) %>%
  summarise(CountOfEthnicity = n()) %>%
  arrange(desc(CountOfEthnicity))
ggplot(homeMortgage_ethnicity, aes(x = reorder(applicant_ethnicity_name, CountOfEthnicity),
                                   y = CountOfEthnicity)) +
  geom_bar(stat='identity',colour="white", fill =fillColor2) +
  geom_text(aes(x = applicant_ethnicity_name, y = 1, label = paste0("(",round(CountOfEthnicity),
")",sep="")),
            hjust=0, vjust=.5, size = 4, colour = 'black',
            fontface = 'bold') +
  labs(x = 'applicant_ethnicity_name', y = '%age Count Of Action Taken', title = 'Actions in Loa
ns') +
  coord_flip() +
  theme_bw()
```




The Not Hispanic or Latino ethnic community applies for the largest percentage of the loans.

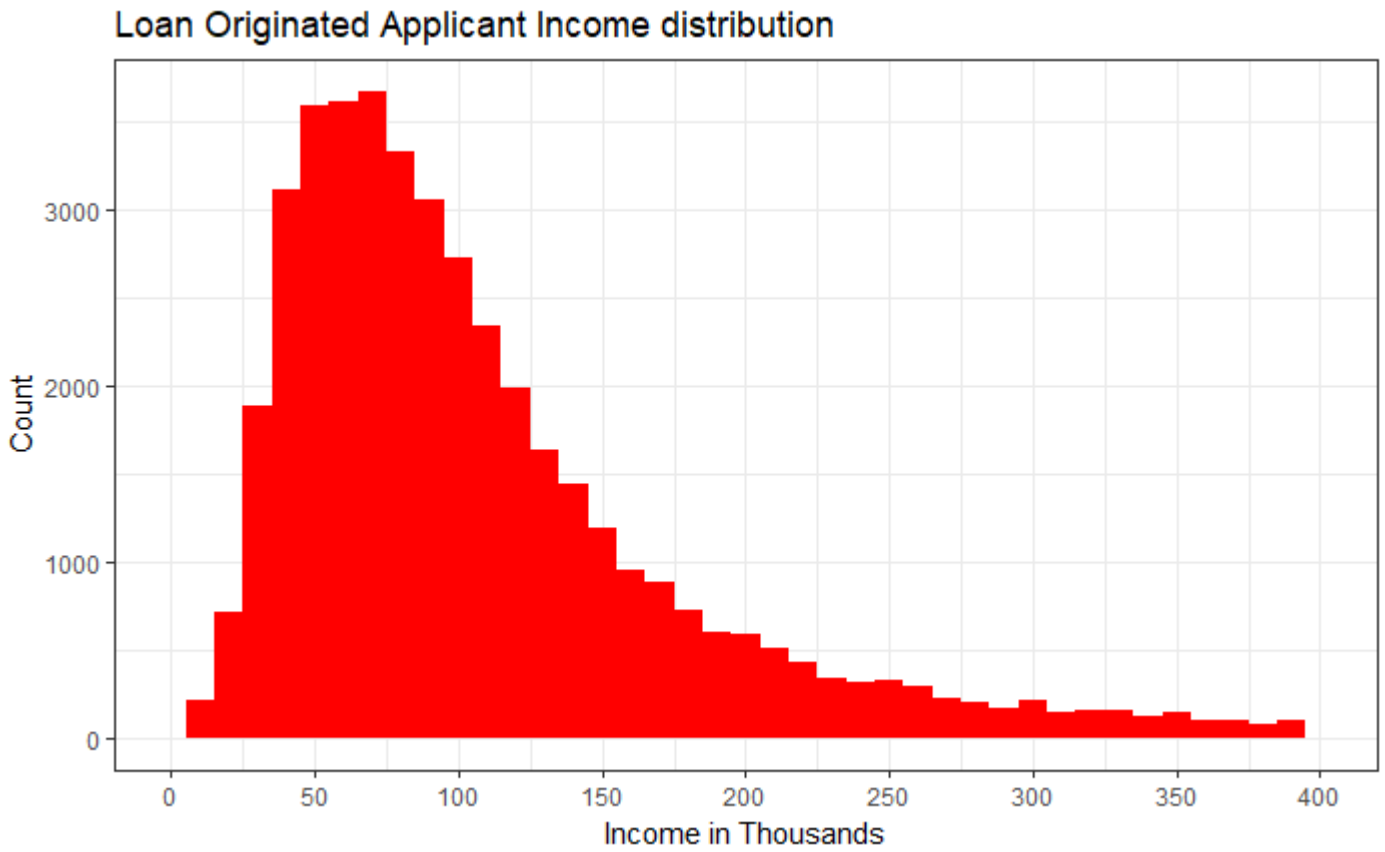
Hide

```
homeMortgageStatus_applicant_race1 = home %>% group_by(action_taken_name,applicant_race_name_1)
%>%
  summarise(CountOfActionTaken = n()) %>%
  arrange(desc(CountOfActionTaken))
homeMortgage_applicant_race1 = home %>% group_by(applicant_race_name_1) %>%
  summarise(CountOfRace1 = n()) %>%
  arrange(desc(CountOfRace1))
ggplot(homeMortgage_applicant_race1, aes(x = reorder(applicant_race_name_1, CountOfRace1),
                                  y = CountOfRace1)) +
  geom_bar(stat='identity',colour="white", fill =fillColor) +
  geom_text(aes(x = applicant_race_name_1, y = 1, label = paste0("(",round(CountOfRace1),")",sep
="")),
            hjust=0, vjust=.5, size = 4, colour = 'black',
            fontface = 'bold') +
  labs(x = 'Race Name', y = 'Count Of Action Taken', title = 'Actions in Loans by Race') +
  coord_flip() +
  theme_bw()
```




Hide

```
actionStatus = "Loan originated"
breaks = seq(0,400,50)
home %>%
  filter(action_taken_name == actionStatus ) %>%
ggplot(aes(applicant_income_000s)) +
  scale_x_continuous(limits = c(0, 400),breaks=breaks ) +
  geom_histogram(binwidth = 10,,fill = c("red")) +
  labs(x = 'Income in Thousands', y = 'Count', title = 'Loan Originated Applicant Income distrib
ution') +  theme_bw()
```




We observe that MOST of the loans which are originated have applicants with income around Sixty Thousand to Seventy Five thousand dollars.

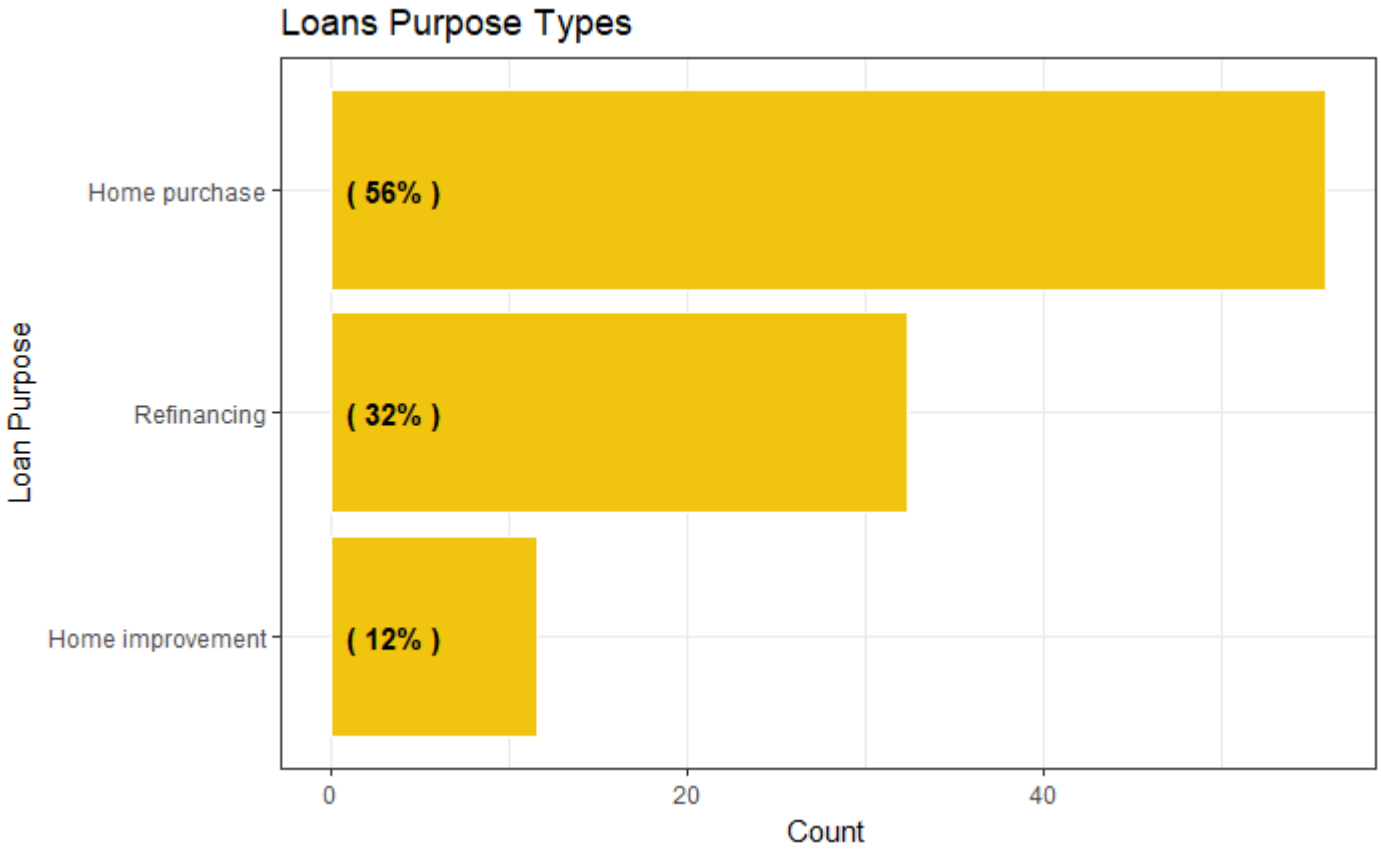
## Loan Purpose Types

We investigate the different loan Purpose Types associated with the loans. Loan Purpose Types distribution

Hide

```
home %>%
  filter(!is.na(loan_purpose_name)) %>%
  group_by(loan_purpose_name) %>%
  summarise(CountLoanPurpose = n() ) %>%
  mutate(percentage = ( CountLoanPurpose/sum(CountLoanPurpose) ) *100 ) %>%
  mutate(loan_purpose_name = reorder(loan_purpose_name, percentage)) %>%

  ggplot(aes(x = loan_purpose_name,y = percentage)) +
  geom_bar(stat='identity',colour="white", fill =fillColor2) +
  geom_text(aes(x = loan_purpose_name, y = 1, label = paste0("( ",round(percentage),"% )",sep=""
)),
            hjust=0, vjust=.5, size = 4, colour = 'black',
            fontface = 'bold') +
  labs(x = 'Loan Purpose', y = 'Count', title = 'Loans Purpose Types') +
  coord_flip() +
  theme_bw()
```




Home Purchase and Refinancing are the major Loan Purpose types.

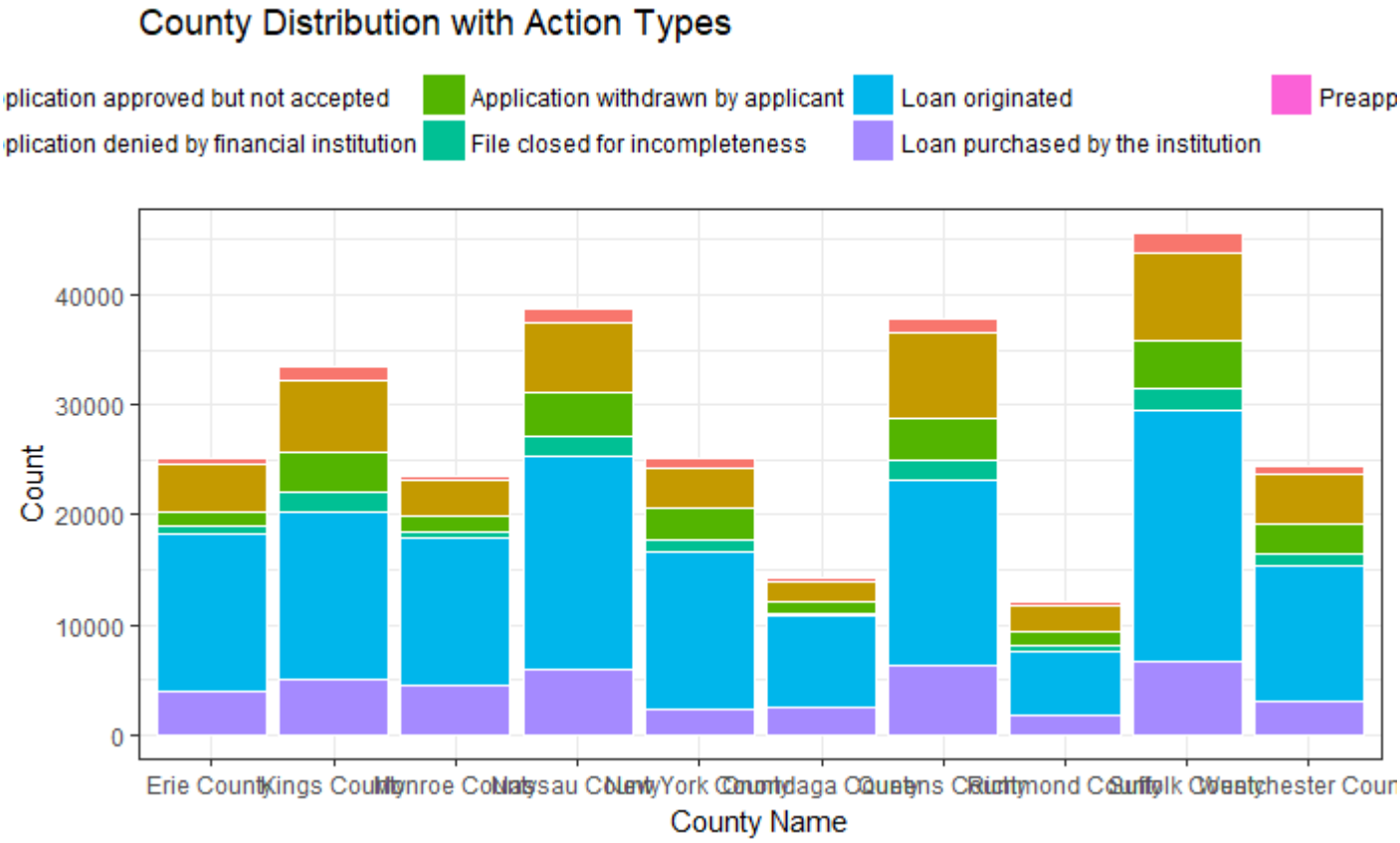
## Counties and Loan distribution

We display the Counties and the Loans Type distribution.

Hide

```
Top10Counties = hmda %>%
  filter(!is.na(county_name)) %>%
  group_by(county_name) %>%
  summarise(CountLoanPurpose = n() ) %>%
  mutate(percentage = ( CountLoanPurpose/sum(CountLoanPurpose) ) *100 ) %>%
  mutate(county_name = reorder(county_name, percentage)) %>%
  arrange(desc(percentage)) %>%
  head(10)
hmda %>%
  filter(!is.na(county_name)) %>%
  filter(county_name %in% Top10Counties$county_name) %>%
  group_by(county_name,action_taken_name) %>%
  summarise(CountLoanPurpose = n() ) %>%


  ggplot(aes(x = county_name,y = CountLoanPurpose,fill = action_taken_name)) +
  geom_bar(stat='identity',colour="white") +
  labs(x = 'County Name', y = 'Count', title = 'County Distribution with Action Types') +
  theme_bw() + theme(legend.position="top")
```



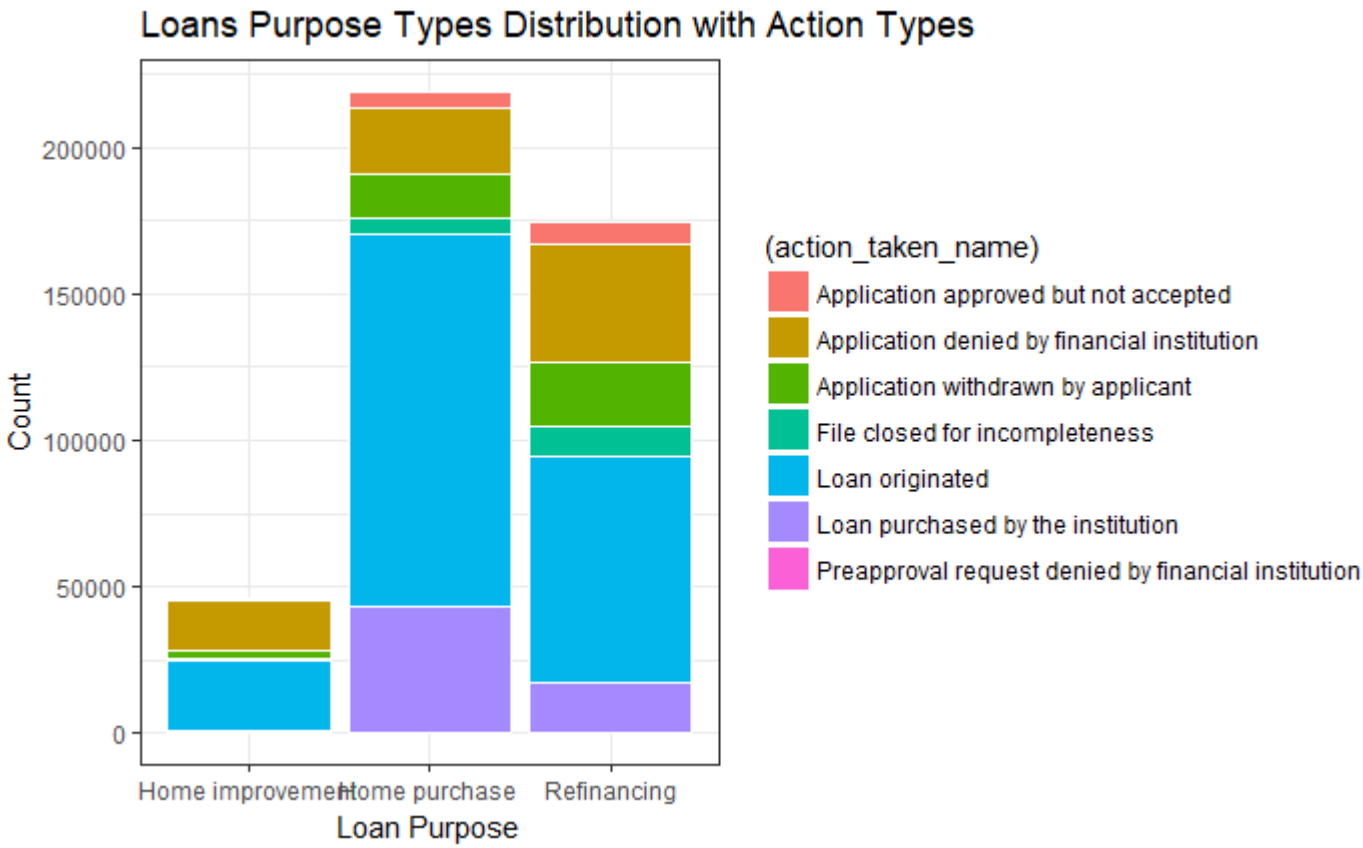

## Loan purpose types and their actions

The following bar graph shows the Loan Purpose Types along with the different actions.

Hide

```
hmda %>%
  filter(!is.na(loan_purpose_name)) %>%
  group_by(loan_purpose_name,action_taken_name) %>%
  summarise(CountLoanPurpose = n() ) %>%

  ggplot(aes(x = loan_purpose_name,y = CountLoanPurpose,fill =(action_taken_name))) +
  geom_bar(stat='identity',colour="white") +
  labs(x = 'Loan Purpose', y = 'Count', title = 'Loans Purpose Types Distribution with Action Ty
pes') +
  theme_bw()
```




## Modelling using Classification and Regression Trees

We predict whether the status of the Loan would be Loan originated or not. The following tree shows the conditions which would be used to determine whether the would be Loan originated or not.

Select Columns for modelling Here we select the columns which would be required for modelling.We make the columns as factors so that they can be used for the CART model.

Hide

```
selectedCols = c("action_taken","applicant_ethnicity",
"applicant_income_000s","applicant_race_1","co_applicant_ethnicity",
"co_applicant_sex", "county_code","hoepa_status","lien_status",
"loan_purpose","loan_type","msamd",
"owner_occupancy","preapproval",
"property_type","purchaser_type","loan_amount_000s")
homeMortgage_selectedCols = hmda %>% select(selectedCols) %>%
  mutate(isLoanOriginated = FALSE)  %>%
  mutate(isLoanOriginated = replace(isLoanOriginated, action_taken == 1, TRUE)) %>%
  select(-action_taken)
homeMortgage_selectedCols$applicant_ethnicity = as.factor(homeMortgage_selectedCols$applicant_et
hnicity)
homeMortgage_selectedCols$applicant_race_1 = as.factor(homeMortgage_selectedCols$applicant_ethni
city)
homeMortgage_selectedCols$co_applicant_ethnicity = as.factor(homeMortgage_selectedCols$co_applic
ant_ethnicity)
homeMortgage_selectedCols$co_applicant_sex = as.factor(homeMortgage_selectedCols$co_applicant_se
x)
homeMortgage_selectedCols$county_code = as.factor(homeMortgage_selectedCols$county_code)
homeMortgage_selectedCols$hoepa_status =  as.factor(homeMortgage_selectedCols$hoepa_status)
homeMortgage_selectedCols$lien_status =  as.factor(homeMortgage_selectedCols$lien_status)
homeMortgage_selectedCols$loan_purpose =  as.factor(homeMortgage_selectedCols$loan_purpose)
homeMortgage_selectedCols$loan_type =  as.factor(homeMortgage_selectedCols$loan_type)
homeMortgage_selectedCols$owner_occupancy =  as.factor(homeMortgage_selectedCols$owner_occupanc
y)
homeMortgage_selectedCols$preapproval =  as.factor(homeMortgage_selectedCols$preapproval)
homeMortgage_selectedCols$property_type =  as.factor(homeMortgage_selectedCols$property_type)
homeMortgage_selectedCols$purchaser_type =  as.factor(homeMortgage_selectedCols$purchaser_type)
```

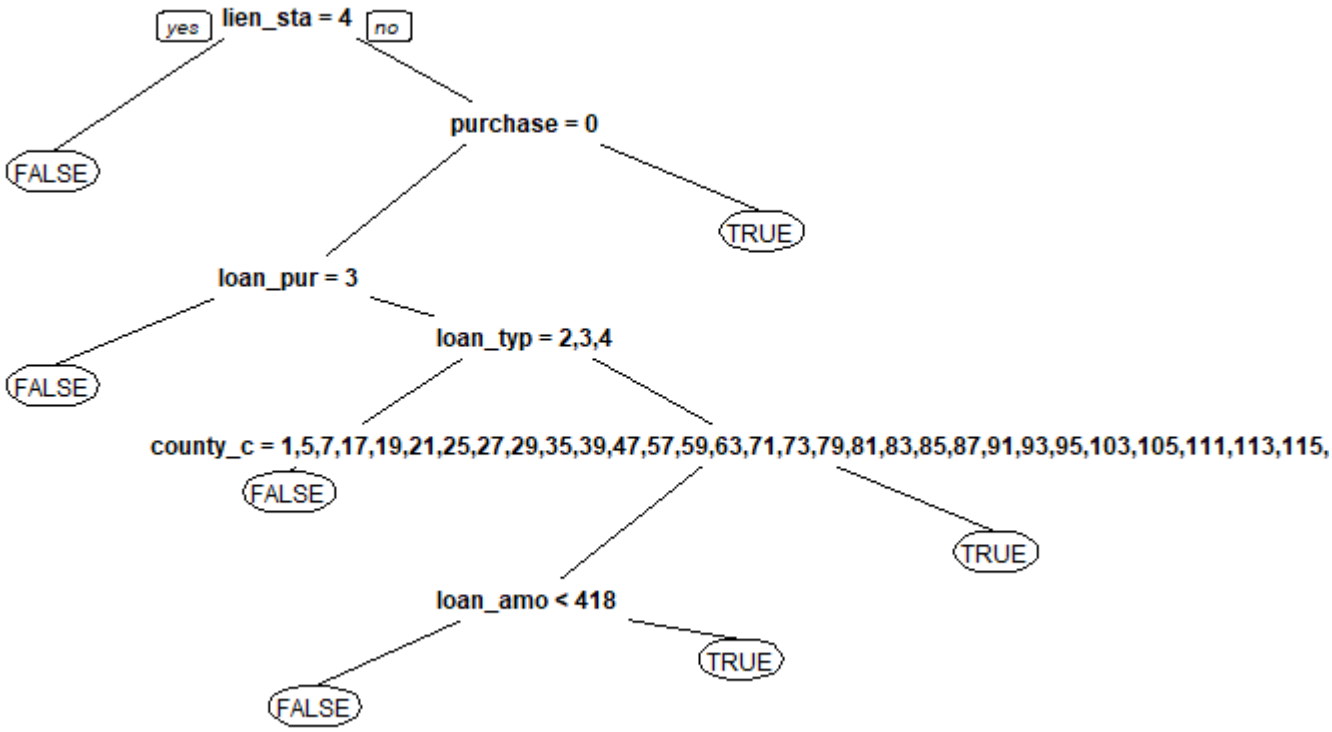
## Build and Visualize the CART model

We build and visualize the CART model. Through this model, we can examine the most important features which impact the decision for Loan Origination.

Hide

```
set.seed(3000)
split = sample.split(homeMortgage_selectedCols$isLoanOriginated, SplitRatio = 0.8)
Train = subset(homeMortgage_selectedCols, split==TRUE)
Test = subset(homeMortgage_selectedCols, split==FALSE)

# CART model
homeMortgageTree = rpart(isLoanOriginated ~., method="class", data = Train, control=rpart.contro
l(minbucket=5))
prp(homeMortgageTree)
```




Performance of the model:

Hide

```
library(ROCR)
```

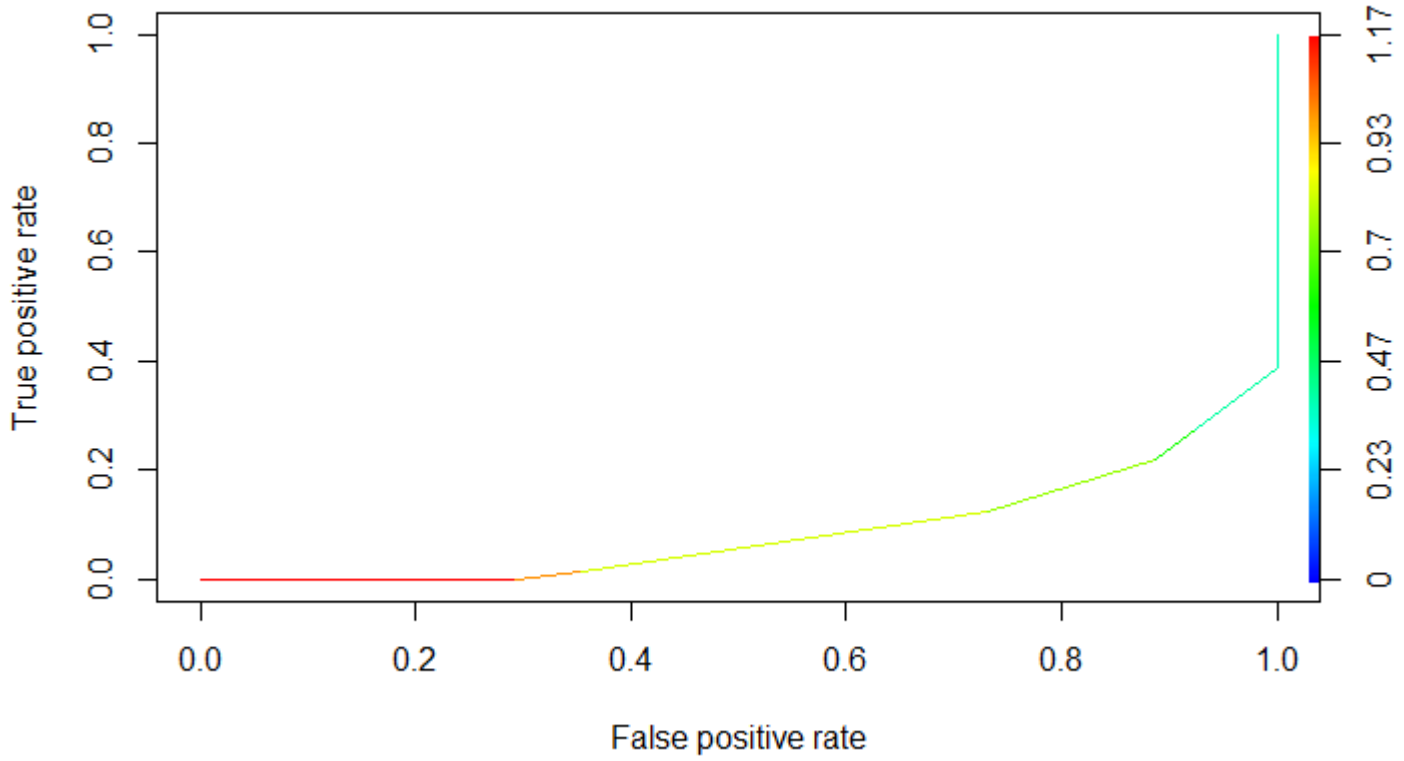
```
package <U+393C><U+3E31>ROCR<U+393C><U+3E32> was built under R version 3.4.3Loading required pac
kage: gplots
package <U+393C><U+3E31>gplots<U+393C><U+3E32> was built under R version 3.4.3
Attaching package: <U+393C><U+3E31>gplots<U+393C><U+3E32>

The following object is masked from <U+393C><U+3E31>package:stats<U+393C><U+3E32>:

    lowess
```

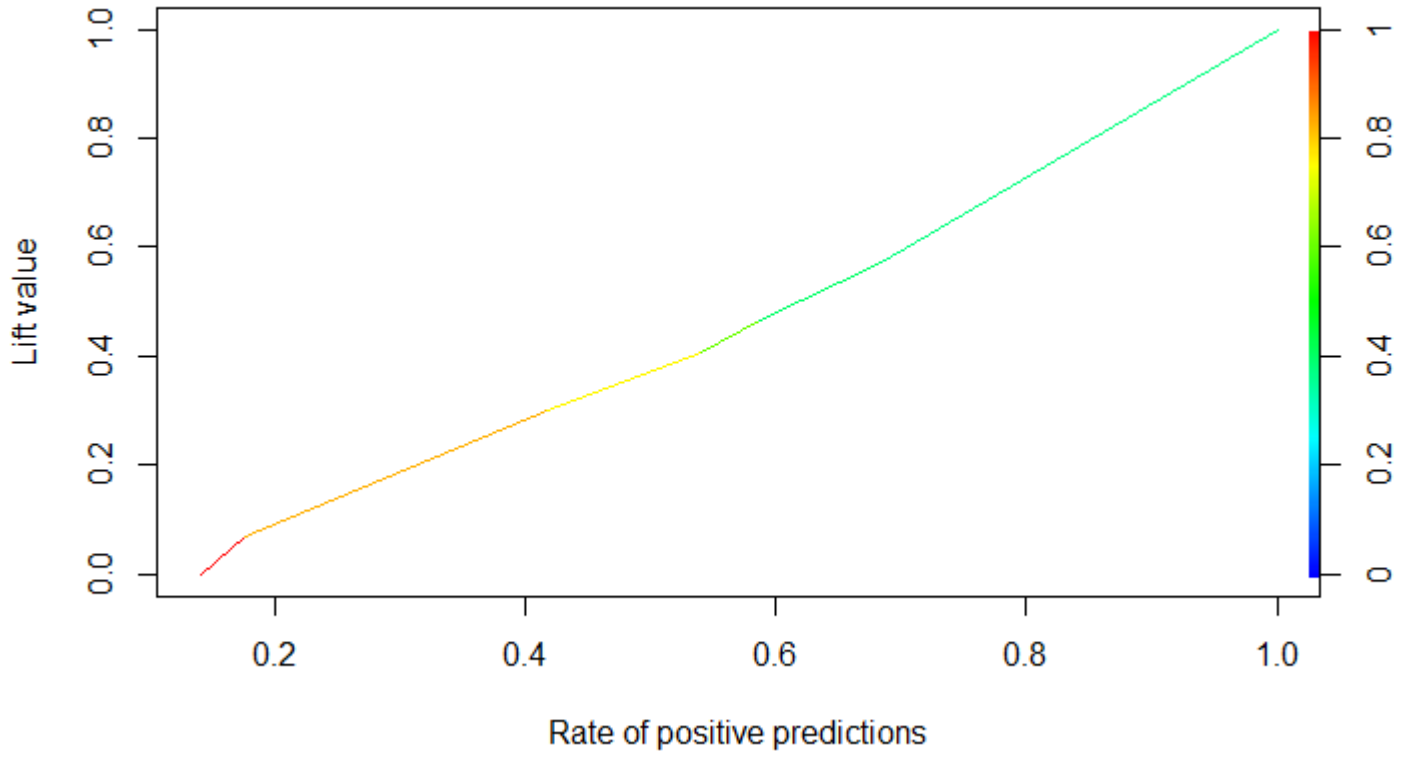
Hide

```
roc_pred <- prediction(pred[,1], Test$isLoanOriginated)
plot(performance(roc_pred, measure="tpr", x.measure="fpr"), colorize=TRUE)
```




Hide

```
plot(performance(roc_pred, measure="lift", x.measure="rpp"), colorize=TRUE)
```

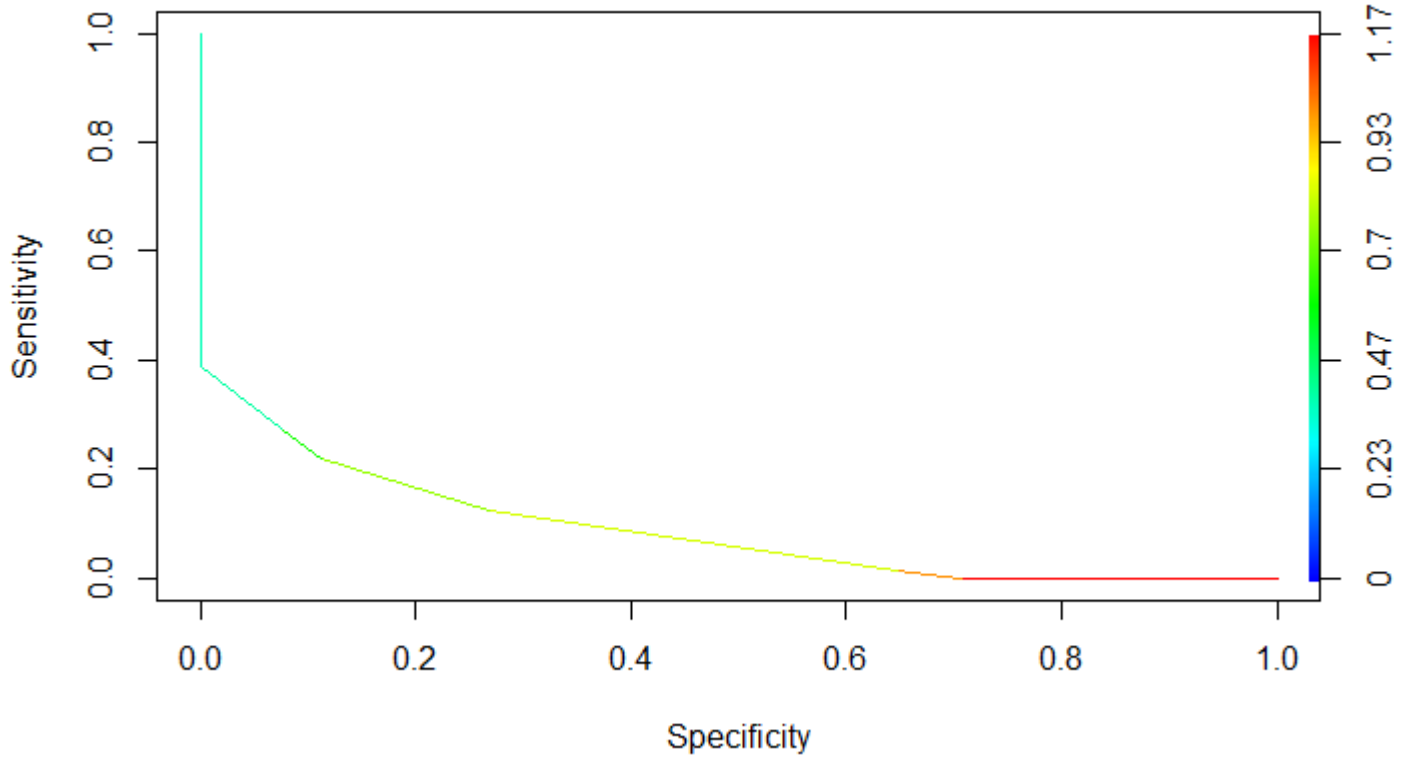


Here we can see that the model is not doing very well. The tighter the ROC curve hugs towards the left the better is the model.

Sensitivity/specificity curve and precision/recall curve:

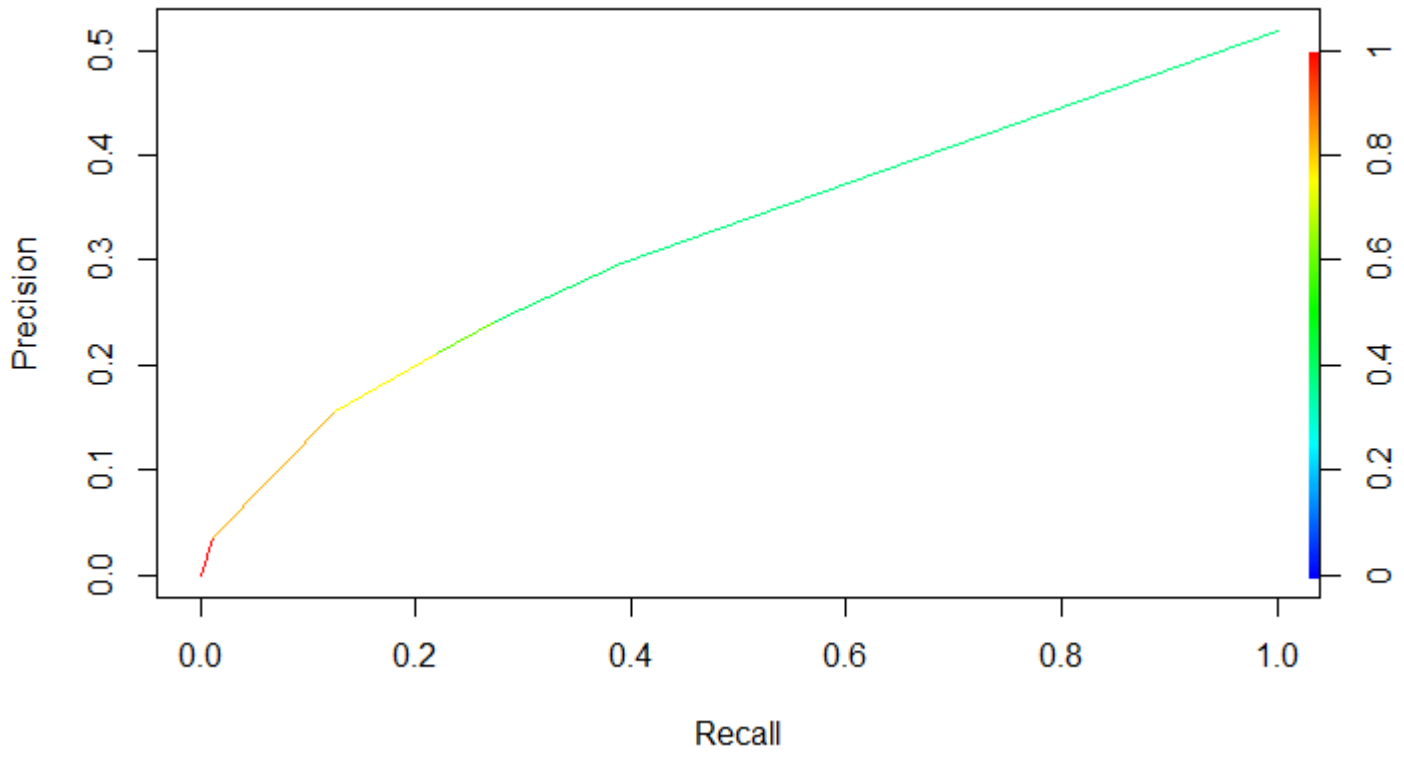
Hide

```
plot(performance(roc_pred, measure="sens", x.measure="spec"), colorize=TRUE)
```




Hide

```
plot(performance(roc_pred, measure="prec", x.measure="rec"), colorize=TRUE)
```




## Conclusion:

Lien_stat, Purchase, Loan_Pur, Loan_type, County_c and Loan_amount are the most important variables to decide whether a mortgage application will be accepted or not.