HMDA Final Report

Kankan Yang, Rita Li, Shiqi Lin, Yixuan Wang

April 10, 2018

The URL for our Team GitHub repository is https://github.com/ritali517/Marketing-Analytics_Shiqi_Rita_Yixuna_Kankan.git
The URL of kernel website is https://www.kaggle.com/jboysen/ny-home-mortgage/kernels
The URL of Kaggle dataset is https://www.kaggle.com/jboysen/ny-home-mortgage/data

```
setwd("~/ConsumerDB")
NYHMDA = read.csv("HMDA.csv")
```

1. Managerial objective

The managerial objective is to specify the key determinants of whether an applicant can obtain a mortgage from financial institutions. The determinants may include applicant characteristics, property types, loan purpose and location. Therefore, we should figure out the relationships between loan action taken and relevant variables through regression or other statistical methods.

2. Measurement type of each variable

Nominal (33 in total): action_taken_name, agency_name, agency_abbr, applicant_ethnicity_name, applicant_race_name_1, applicant_sex_name, county_name, hoepa_status_name, lien_status_name, loan_purpose_name, loan_type_name, msamd_name, owner_occupancy_name, preapproval_name, property_type_name, purchaser_type_name, action_taken, agency_code, applicant_ethnicity, applicant_race_1, applicant_sex, county_code, hoepa_status, lien_status, loan_purpose, loan_type, msamd, owner_occupancy, preapproval, property_type, purchaser_type, respondent_id, sequence_number

Ratio (8 in total): applicant_income_000s, hud_median_family_income, laon_amount_000s, number_of_1_to_4_family_units, number_of_owner_occupied_units, minority_population, population, tract_to_msamd_income

3.1. Delete unnecessary variables

First, we delete all variables that we deem unnecessary, such as variables concerning HUD, which are community level data.

```
NYHMDA_clear<-subset(NYHMDA,select=-
c(10,11,12,13,15,16,17,18,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,40,41,42,43,44,45,46,47,60,61,66,67,68,
69,70,77,71,73,74,75,76,78,56,57))
```

3.2. Chi square tests

3.2.1. Gender & Loan action taken

```
NYHMDA_clear$Gender_ismissing="0"
NYHMDA_clear[NYHMDA_clear$applicant_sex=="3"|NYHMDA_clear$applicant_sex=="4",]$Gender_ismissing="1"
library(MASS)
Gender_loan_action<-table(NYHMDA_clear$action_taken_name,NYHMDA_clear$Gender_ismissing)
chisq.test(Gender_loan_action)#
##  Pearson's Chi-squared test
##
## data:  Gender_loan_action
## X-squared = 136800, df = 6, p-value < 2.2e-16
```

According to the Chi-square test, p-value is less than 0.01. Therefore, observations missing under applicant's sex are not missing at random and we will use the mod applicant sex to impute missing observations and create another variable to note whether the observation was missing at first.

Mode function

```
Mode=function(x){
  ux=sort(unique(x))
```

```
  tabx=table(x)
  maxf=ux[which(tabx==max(tabx))]
  return(maxf)
}

Mode(NYHMDA_clear$applicant_sex)

## [1] 1

NYHMDA_clear[NYHMDA_clear$applicant_sex=="3"|NYHMDA_clear$applicant_sex=="4",]$applicant_sex="1"
NYHMDA_clear[NYHMDA_clear$applicant_sex_name=="Information not provided by applicant in mail,Internet,or
telephone application"|NYHMDA_clear$applicant_sex_name=="Not applicable",]$applicant_sex_name="Male"
```

3.2.2. Race & Loan action taken

```
NYHMDA_clear$Race_ismissing="0"
NYHMDA_clear[NYHMDA_clear$applicant_race_1=="6"|NYHMDA_clear$applicant_race_1=="7",]$Race_ismissing="1"
library(MASS)
Race_loan_action<-table(NYHMDA_clear$action_taken_name,NYHMDA_clear$Race_ismissing)
chisq.test(Race_loan_action)
##  Pearson's Chi-squared test
##
## data:  Race_loan_action
## X-squared = 102040, df = 6, p-value < 2.2e-16
```

According to the Chi-square test, p-value is less than 0.01. Therefore, again we will use the mod value to impute missing observations and create another variable to note whether the observation was missing at first.

```
Mode(NYHMDA_clear$applicant_race_1)

## [1] 5

NYHMDA_clear[NYHMDA_clear$applicant_race_1=="6"|NYHMDA_clear$applicant_race_1=="7",]$applicant_race_1="5"
NYHMDA_clear[NYHMDA_clear$applicant_race_name_1=="Information not provided by applicant in mail,Internet,or
telephone application"|NYHMDA_clear$applicant_race_name_1=="Not applicable",]$applicant_race_name_1="White"
```

3.2.3. Ethnicity & Loan action taken

```
NYHMDA_clear$Ethnicity_ismissing="0"
NYHMDA_clear[NYHMDA_clear$applicant_ethnicity=="3"|NYHMDA_clear$applicant_ethnicity=="4",]$Ethnicity_ismissing
="1"
library(MASS)
Ethnicity_loan_action<-table(NYHMDA_clear$action_taken_name,NYHMDA_clear$Ethnicity_ismissing)
chisq.test(Ethnicity_loan_action)
##  Pearson's Chi-squared test
##
## data:  Ethnicity_loan_action
## X-squared = 106050, df = 6, p-value < 2.2e-16
```

According to the Chi-square test, p-value is less than 0.01. Therefore, again we will use the mod value to impute missing observations and create another variable to note whether the observation was missing at first.

```
Mode(NYHMDA_clear$applicant_ethnicity)

## [1] 2

NYHMDA_clear[NYHMDA_clear$applicant_ethnicity=="3"|NYHMDA_clear$applicant_ethnicity=="4",]$applicant_ethnicity
="2"
NYHMDA_clear[NYHMDA_clear$applicant_ethnicity_name=="Information not provided by applicant in mail,Internet,or
telephone application"|NYHMDA_clear$applicant_ethnicity_name=="Not applicable",]$applicant_ethnicity_name="Not
Hispanic or Latino"
```

3.2.4. Income & Loan action taken

```
NYHMDA_clear$Income_ismissing="0"
NYHMDA_clear[is.na(NYHMDA_clear$applicant_income_000s),]$Income_ismissing="1"
library(MASS)
Income_loan_action<-table(NYHMDA_clear$action_taken_name,NYHMDA_clear$Income_ismissing)
chisq.test(Income_loan_action)
##  Pearson's Chi-squared test
##
## data:  Income_loan_action
## X-squared = 75171, df = 6, p-value < 2.2e-16
```

According to the Chi-square test, p-value is less than 0.01. Therefore, we will use the median income level to impute missing observations and create another variable to note whether the observation was missing at first.

```
NYHMDA_clear$applicant_income_000s[is.na(NYHMDA_clear$applicant_income_000s)]=median(as.numeric(NYHMDA_clear$applicant_income_000s),na.rm = TRUE)
```

3.2.5. Lien status & Loan action taken

```
NYHMDA_clear$Lien_ismissing="0"
NYHMDA_clear[NYHMDA_clear$lien_status=="4",]$Lien_ismissing="1"
library(MASS)
Lien_loan_action<-table(NYHMDA_clear$action_taken_name,NYHMDA_clear$Lien_ismissing)
chisq.test(Lien_loan_action)
##  Pearson's Chi-squared test
##
## data:  Lien_loan_action
## X-squared = 439650, df = 6, p-value < 2.2e-16
```

According to the Chi-square test, p-value is less than 0.01. Therefore, again we will use the mod value to impute missing observations and create another variable to note whether the observation was missing at first.

```
Mode(NYHMDA_clear$lien_status)

## [1] 1

NYHMDA_clear[NYHMDA_clear$lien_status=="4",]$lien_status="1"
NYHMDA_clear[NYHMDA_clear$lien_status_name=="Not applicable",]$lien_status_name="Secured by a first lien"
```

3.2.6. Owner occupancy & Loan action taken

```
NYHMDA_clear$Occupancy_ismissing="0"
NYHMDA_clear[NYHMDA_clear$owner_occupancy=="3",]$Occupancy_ismissing="1"
library(MASS)
Occupancy_loan_action<-table(NYHMDA_clear$action_taken_name,NYHMDA_clear$Occupancy_ismissing)
chisq.test(Occupancy_loan_action)
##  Pearson's Chi-squared test
##
## data:  Occupancy_loan_action
## X-squared = 489.98, df = 6, p-value < 2.2e-16
```

According to the Chi-square test, p-value is less than 0.01. Therefore, again we will use the mod value to impute missing observations and create another variable to note whether the observation was missing at first.

```
Mode(NYHMDA_clear$owner_occupancy)

## [1] 1

NYHMDA_clear[NYHMDA_clear$owner_occupancy=="3",]$owner_occupancy="1"
NYHMDA_clear[NYHMDA_clear$owner_occupancy_name=="Not applicable",]$owner_occupancy_name="Owner-occupied as a principal dwelling"
```

3.3. Summary and rename dataset

```
NYHMDA_new<-NYHMDA_clear
```

The Data Frame now contains 35 variables, down from 79 initially and the number of observations are 439654, all missing observations are imputed with either mode or median value and we also created missing variable identifiers.

4. Table summarizing range/variation of key variables

4.1. County name

```
county_table <- table(NYHMDA_new$county_name)
frame <- as.data.frame(county_table)
frame[order(frame[,2],decreasing=TRUE),][1:10,]

##                Var1  Freq
## 53     Suffolk County 45525
## 31      Nassau County 38797
## 42      Queens County 37866
## 25       Kings County 33540
## 16        Erie County 25109
## 32    New York County 25046
## 61 Westchester County 24371
## 29      Monroe County 23536
## 35    Onondaga County 14277
## 44    Richmond County 12153
```

Loan applicants in 2015 were from 63 different counties in New York States. Here we only show 10 counties that had most applicants. For example, there were 45,525 applicants in Suffolk in 2015, which took the most part of total applicants.

4.2. Agency name

```
table(NYHMDA_new$agency_abbr)

##
##   CFPB   FDIC    FRS    HUD   NCUA    OCC
## 177762  15555  10211 150441  50944  34741
```

There are 6 agencies to apply for loans.From the table above, we find that most loans were originated in CFPB(Consumer Financial Protection Bureau) and HUD(Department of Housing and Urban Development).

4.3. HOEPA status

```
table(NYHMDA_new$hoepa_status_name)

##
##      HOEPA loan Not a HOEPA loan
##              60          439594
```

HOEPA loan is the closed-end equity loan bearing rates or fees above a specific percentage or amount under the Home Ownership and Equity Protection Act.Among all loan applications, only 60 were HOEPA loan, while most were not HOEPA loans.

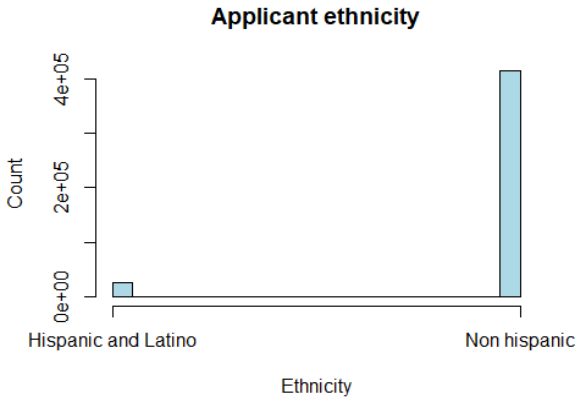
5. Analysis for key variables using histogram/density plot

Below we provide histogram/density plot for some variables that we want to have a more visual understanding of their distributions.

5.1. Actions in loans

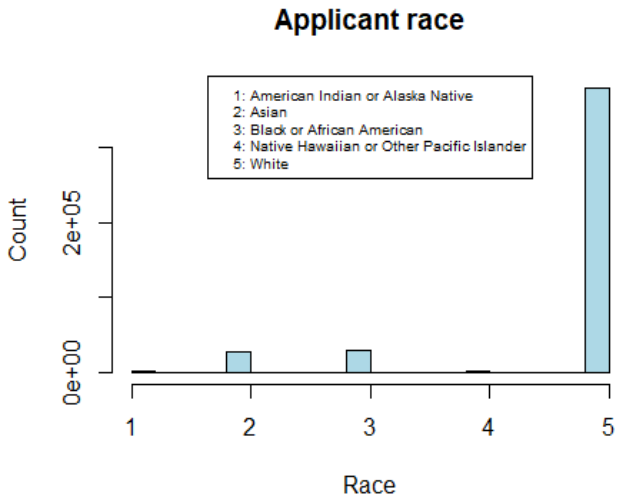
According to the census, 228054 over 439654 loans were originated. Rest of loans not originated can be classified into 6 categories: loans approved but not accepted (14180), loans denied (79697), which is the second to top loan action taken, loans withdrawn by applicant (39496), loans closed for incompleteness (16733), loans purchased by institution (61490) and pre-approval denied (4).

5.2. Ethnicity



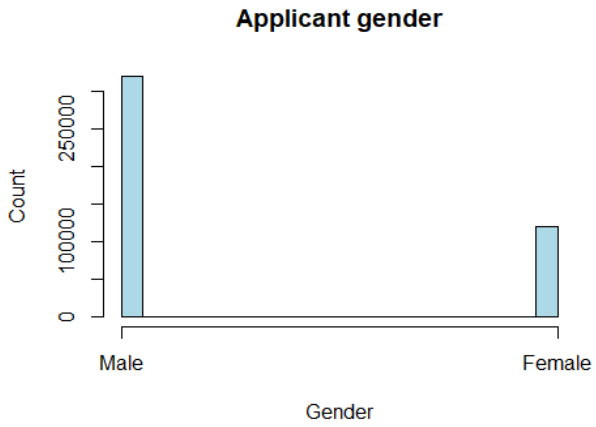
Among all loan applicants, over 94% of loans (414581/439654) are applied by non-Hispanic or Latino individuals. According to the law, agencies shouldn't discriminate against individuals in different ethnicities. Therefore, it's interesting to look into whether loan approval depends on applicant's ethnicity or not.

5.3. Race



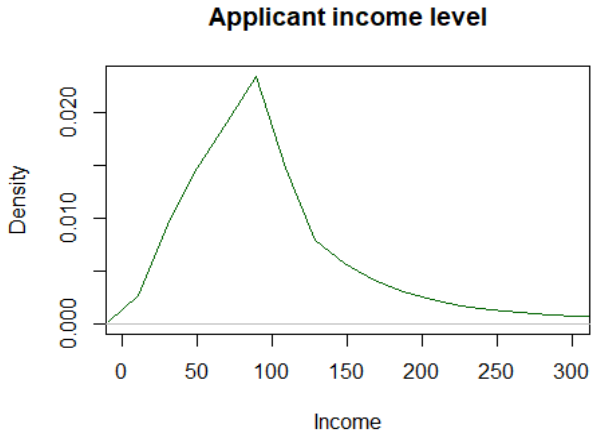
We present this analysis to learn about demographics of the applicants. Among all applicants, 75.9% are white, 6.34% are distributed almost equally between African American and Asian and the rest belongs to other races.

5.4. Gender



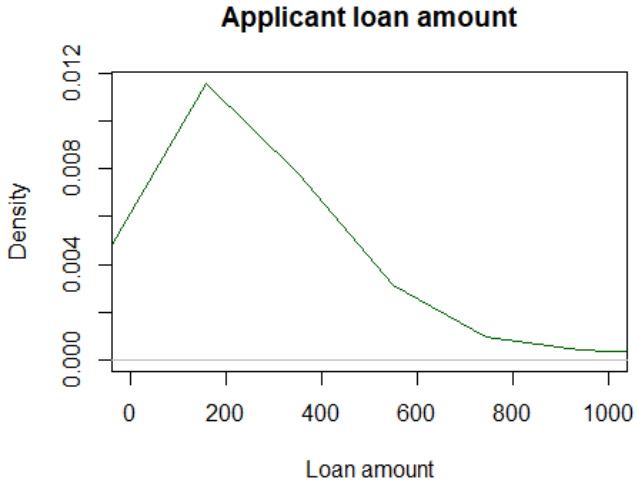
Among all applicants, the number of male applicants is 319777 over 439654, over twice the number of female applicants. The difference can be that female are less likely to be in the labor force compared to their male counterparts, thus not able to provide complete documents (stable income) for loan application, which decreases the likelihood of loan approval.

5.5. Applicant's income



Among all applicants, most people have annual income of around 90,000 dollars. Another thing worth mentioning is individuals with zero income (less than 20,000 people). Since we use the median value for those applicants whose income is not applicable, there might be some bias here and we create another variable to indicate the missing observations.

5.6. Loan amount

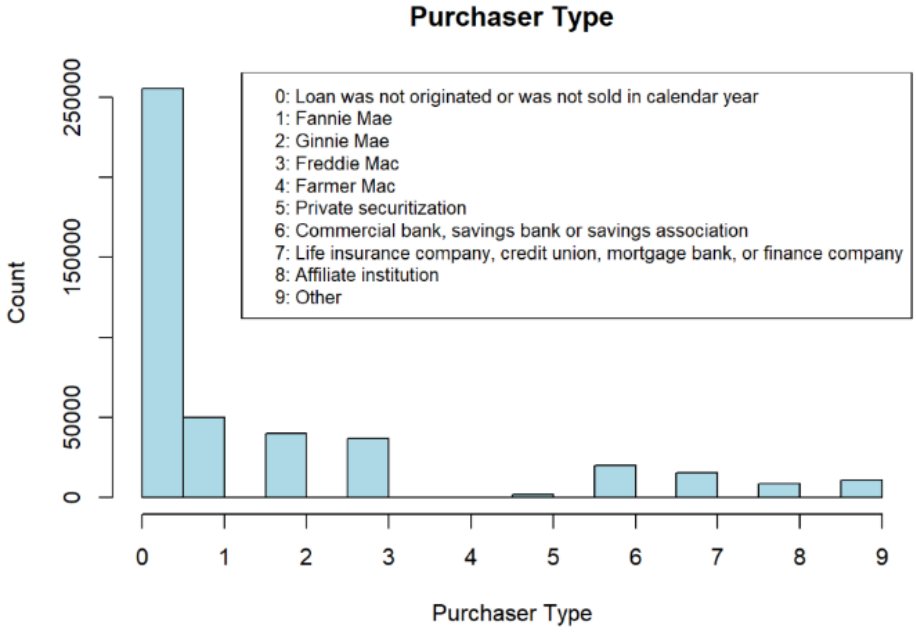



```
summary(NYHMDA_new$loan_amount_000s)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   102.0   208.0   333.3   366.0 99999.0
```

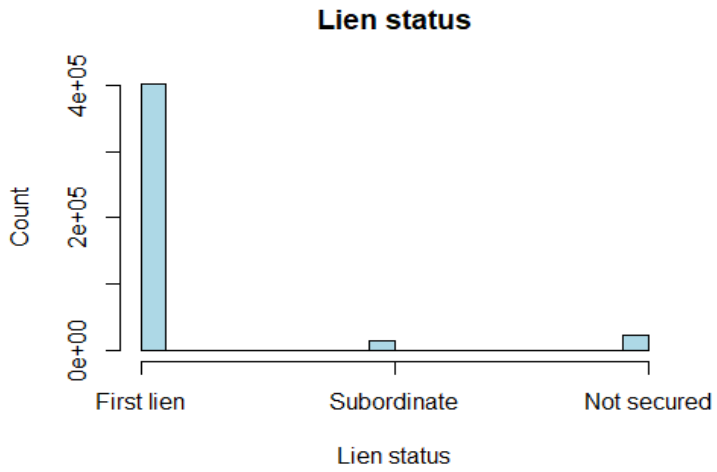
The average loan amounts is around 333,300 dollars. This variables has an IQR of 264,000 dollars, which means that most loans amount from $102,000 to $366,000.

5.7 Purchaser type



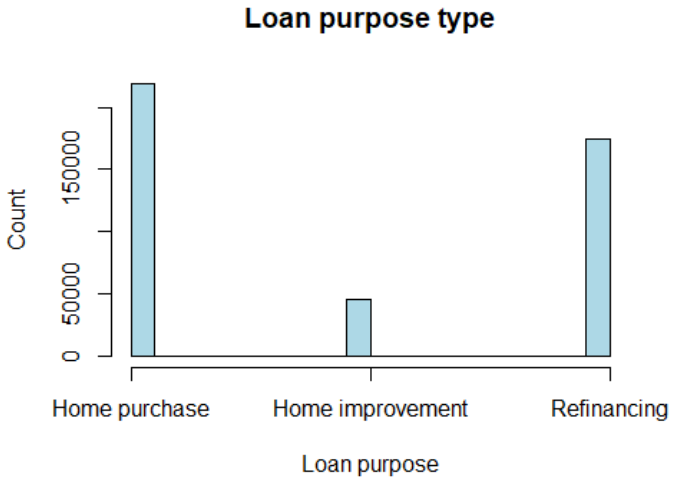
From the chart of purchaser type, we know that most of home mortgage in NY State is either not sold or not originated. For home mortgage that is sold, main purchasers are Fannie Mae, Ginnie Mae, Freddie Mae and banks. The information shows that secondary home mortgage market in NY State is relatively safe and cautious because most of purchasers are government-sponsored enterprise.

5.8 Lien Status



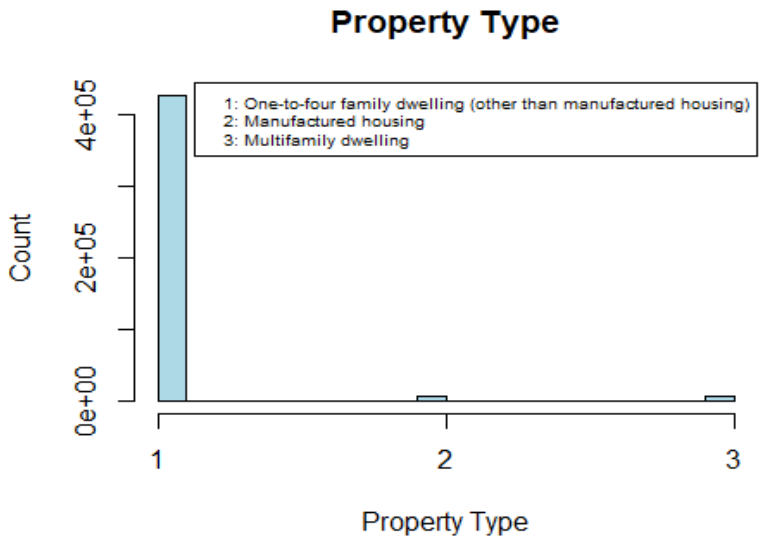
From the chart of lien status, we can see that Most financial institutions are first lien debt holders of the mortgages, which means that if a home mortgage borrower defaults, the home mortgage lender is the first one to get the lien. This is a good phenomenon because borrowers will be less likely to default because of the lien. Also, lenders are protected from losing lots of money.

5.9 Loan Purpose



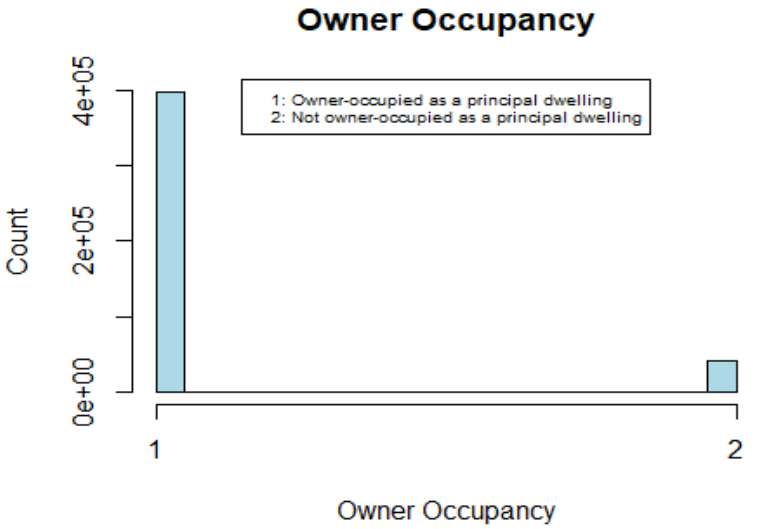
In terms of loan purpose type, home purchase and refinancing are the most prevalent purposes. After all, in general home purchase needs more money than home improvement does. Therefore, information in this chart makes sense.

5.10 Property Type



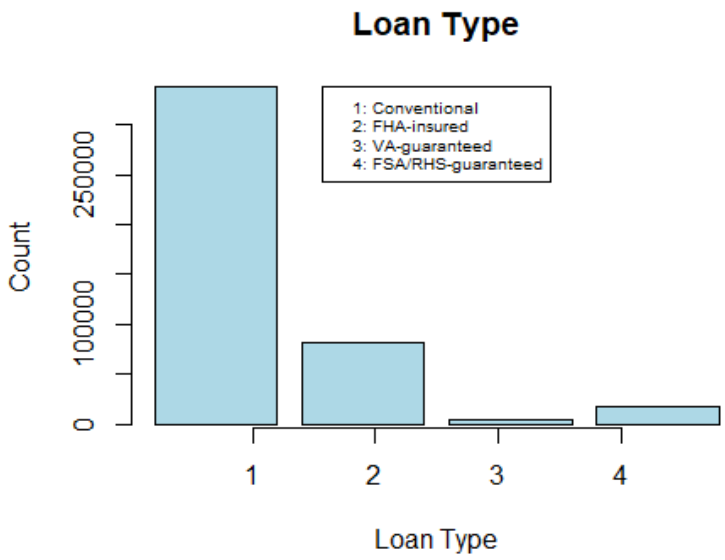
For property type, almost all the property is One-to-four family dwelling (other than manufactured housing). It means that home mortgage application of family dwelling is more than manufactured housing.

5.11 Owner Occupancy



From the chart of owner occupancy, we can see that most of the situation is that the property to which the loan application relates will be the borrower's principal dwelling. This is a good signal because in this way, borrowers will pay more attention to repaying the debt.

5.12 Loan Type

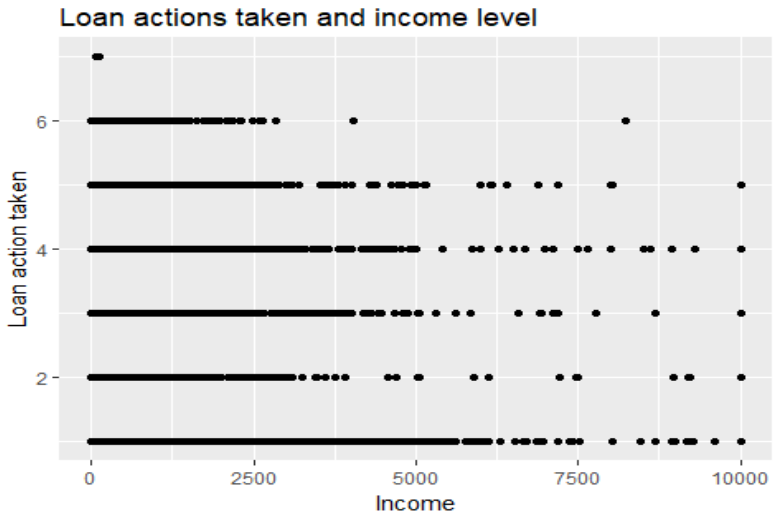


For loan type, most of the loans are conventional loans, but not FHA, VA, FSA, or RHS loans. A FHA loan is a loan insured by the Federal Housing Administration (FHA). A VA loan is a loan guaranteed by the Veterans Administration (VA). A FSA/RHS-guaranteed loan is guaranteed by U.S. department of agriculture. Unlike the other three loan types, conventional loans are not insured or guaranteed by federal government. This result makes sense because only a small number of people are qualified applicants for insured or guaranteed loans. Conventional loans are riskier for mortgage lenders.

6. Relationships between variables especially to loan action taken

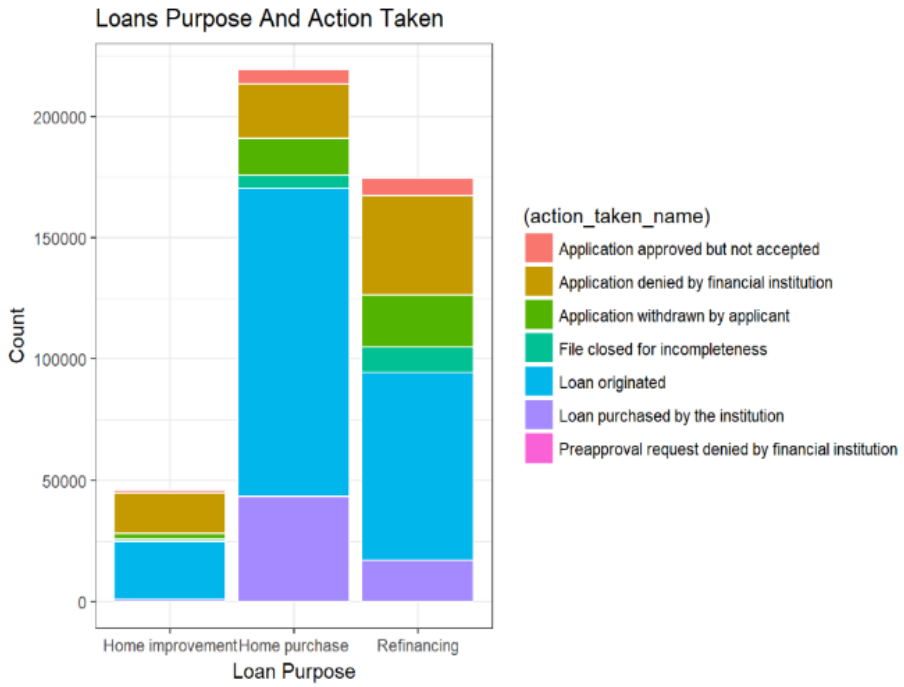
In order to better present the context of pictures, we use the variable “action_taken_name” instead of “action_taken” to represent loan action taken. The corresponding relationship between these two variables are as follows: 1 = Loan originated 2 = Application approved but not accepted 3 = Application denied by financial institution 4 = Application withdrawn by applicant 5 = File closed for incompleteness 6 = Loan purchased by institution 7 = Pre-approval request denied by financial institution

6.1. Income level to loan action taken



In terms of loan actions taken, we expect applicants with higher annual income have higher chance of obtaining a mortgage and the pattern do shows that applicants who have obtained a loan have higher annual income than those applicants whose loans were denied. Income level for Applicants whose loan applications were originated are more likely to lie between 2,500,000 and 7,500,000. While those whose applications were denied usually had income level lower than 5,000,000. In particular, all applicants whose preapproval were denied had relatively lower level of income.

6.2. Loan purpose and loan actions taken



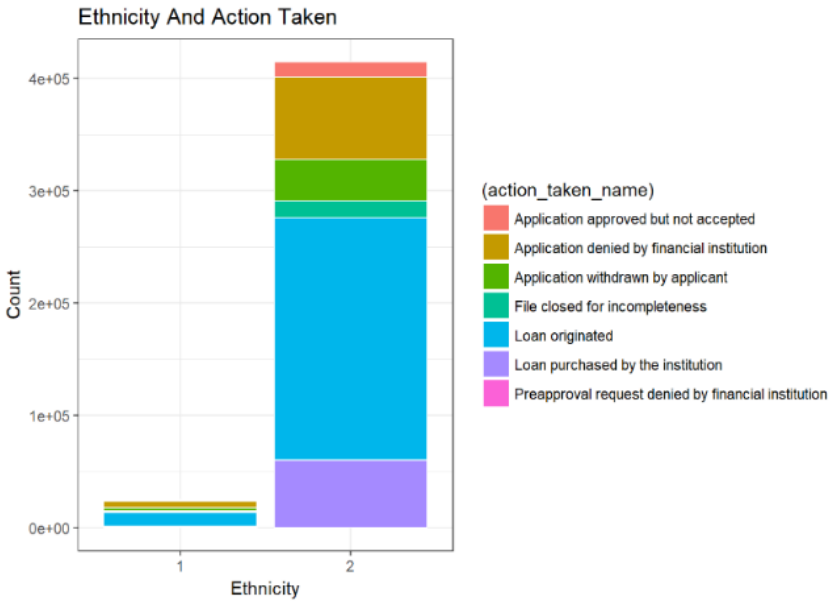
For the plots with two categorical variables, we referred to the plots in the kernel[1]. As for loan purpose, the probability of getting a loan is higher when the purpose of loan is home purchase rather than home improvement and refinancing. Loan application for home improvement is likely to secure a loan as usually it is hard to loan applicants to justify the amount of money they need to borrow for renovation that can generate sufficient cash flows for repayment in the future.

6.3. Gender and loan actions taken

On the x-axis, 1 represents male; 2 represents female. According to this picture, male applicants are more likely to get the loan originated. However, we want to simply conclude that females are discriminated in loan applications. Compared to male, female may have relatively lower income and income level is usually a key determinant of loan origination.

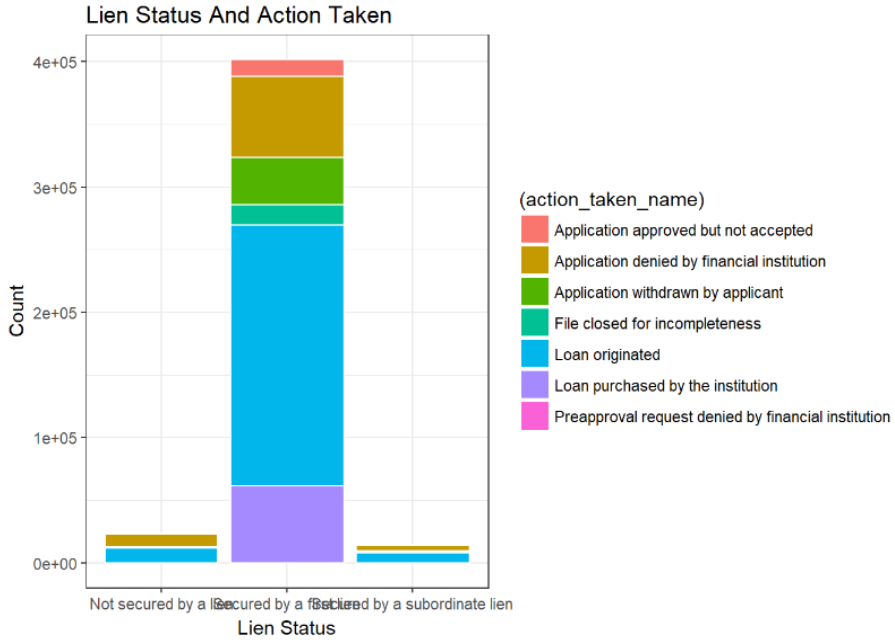
6.4. Ethnicity and loan actions taken

[1] https://www.kaggle.com/jboysen/ny-home-mortgage/kernels



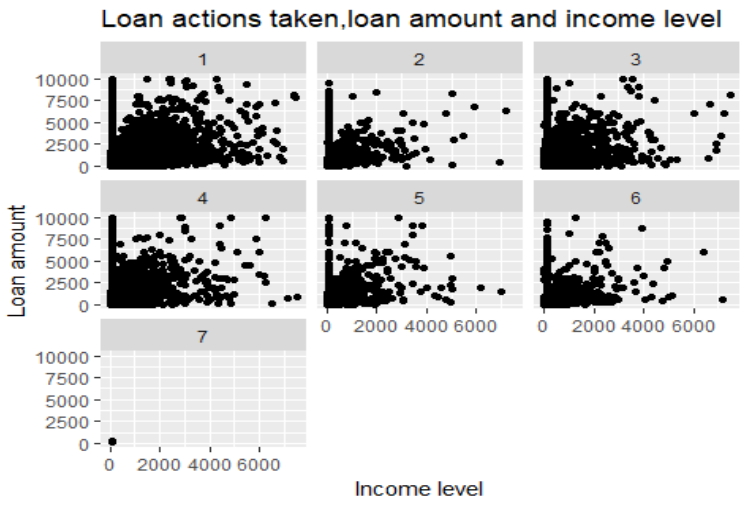
On the x-axis, 1 represents Hispanic or Latino; 2 represents not Hispanic or Latino. The result of applicant ethnicity is similar to gender. Applicants who are not Hispanic or Latino are more likely to get the loan. The reason can be there are certain preferential policies for Hispanic or Latinos in terms of applying for home mortgages. However, the result might not be very robust and we need further regression analysis as most observations are not Hispanic or Latino.

6.5. Lien status and loan actions taken



Every type of lien status has some loans originated. However, loans secured by a first lien have the biggest probability to be originated. For loans secured by a subordinate lien or even not secured, rejection rates are very high. Therefore, compared to some bivariate plots shown above, we believe lien status is a key determinant for loan status.

6.6. Loan amount and income level



From the 7 pictures above, we can see that applicants who have higher income tend to apply for greater amount of loan, and these applications are more likely to be approved, as shown by the positive slopes. In addition, income levels of most applications denied by financial institution were relatively low. Therefore, income level is an important factor for applicants to decide the amount of loan they want to apply and for financial institutions to decide whether to approve loan applications. In the 7 pictures above, we can see income level clustered around 90,000 dollars as a result of mean imputation for missing observations.

7. Define new variables

7.1 Define “Loan_Approval”

The first variable we need in our model is the dependent variable. The purpose is to predicate whether the loan will be approved or denied by the financial institutions. From the data, we can see that action taken can be divided into three categories: loan approved, loan not approved and process incompleted. Therefore, we create dummy variable Loan_Approval and completeness to represent these three categories. We regard 1 (Loan originated),2 (Application approved but not accepted) and 6 (Loan purchased by institution) as loan approved. We regard 3 (Application denied by financial institution), 5 (File closed for incompleteness), 7 (Preapproval request denied by financial institution) as loan not approved. We regard 4 (Application withdrawn by applicant) as process incompleteness.

7.2. Define “Income_level”

We also need to define income level because it may have significant influence on loan approval. We divide applicant income into three levels according to the first quartile and the third quartile. When applicant’s income is higher than $130000, it’s regarded as 1 (high income level). When applicant’s income is lower than $63000, it’s regarded as 3 (low income level). Others can be seen as 2 (middle income level).

8. T-tests between groups

8.1. Difference in loan amounts between loan approved and denied

```
t.test(NYHMDA_new[NYHMDA_new$Loan_Approval==1,]$loan_amount_000s,NYHMDA_new[NYHMDA_new$Loan_Approval==0,]$loan_amount_000s)
```

Here we want to discover how loan amount can differ between approved loan and denied loan. In the t-test, our Null Hypothesis is: the difference in the mean of loan amount is equal to 0 between approved loan and denied loan. And the result (in appendix) shows a very small p-value (0.0001346), which means we should reject the null hypothesis. Thus, loan amount difference between these two groups (approved loan & denied loan) is significant, and the mean loan amount for approved loans is lower than that for denied loans.

8.2. Difference in loan approval rate between high and low income applicants

```
t.test(NYHMDA_new[NYHMDA_new$Income_Level==1,]$Loan_Approval,NYHMDA_new[NYHMDA_new$Income_Level==3,]$Loan_Approval,alternative = "greater")
```

When we analyze the relationship between income level and loan action taken, we only know the general correlation between these two variables. However, we want to further study the quantitative relationship. The Null Hypothesis in this t-test is: the difference in the mean of loan approval rate is less than 0 between high income group and low income group. Since p-value is 2.2e-16, we should reject the null hypothesis and accept the alternative. The result (in appendix) tells us that high income group has higher loan approval rate (69.23%) than low income group does (62.56%). This result makes sense because high income people usually have higher credit than low income people, so financial institutions are more willing to lend money to high income people.

9. Regression analysis

According to all the analysis above, we pick several most important variables including loan amount, applicant income, applicant sex, applicant race and income available as independent variables and loan approval as the dependent variable. We build the initial model to get the first impression and later, we need to do some tests and add other variables into our final model.

```
initial_model <- lm(NYHMDA_new$Loan_Approval ~
NYHMDA_new$loan_amount_000s+NYHMDA_new$applicant_income_000s+factor(NYHMDA_new$applicant_sex)+factor(NYHMDA_new$applicant_race_1))
```

This reduced version of regression shows an adjusted R-square of 1.115%, suggesting that all variables included only account for 1.115% of variations in loan approval rate. All variables are significant at the 99% confidence level. According the the result attached in the appendix, higher loan amount and income level correspond to lower loan approval rate, which is a bit hard to believe as higher income is often considered more promising in loan application. However, since we didn't yet account for omitted variable bias, the coefficient of some variables may be biased.

10. Partitioning data

Here we separate the data into 30% of testing data and 70% of training data. Then, we will run sets of regressions using training data and compare predictability using the testing data.

```
set.seed(2333)
train.index<-sample(c(1:dim(NYHMDA_new)[1]),dim(NYHMDA_new)[1]*0.7)
NYHMDA.train<-NYHMDA_new[train.index,]
NYHMDA.test<-NYHMDA_new[-train.index,]
```

11. Control for omitted variables

The initial model only includes most basic variables, including income, loan amount, sex race. Therefore, for control variables, we include in lien status, ethnicity, loan purpose, property type and missing variable identifiers for sex, income, race and ethnicity.

According to our histograms in the fifth part and knowledge for loan application, lien status and loan purpose are two very important variables. Below we created 4 vectors to generate 54 different regression models and we use two criteria, AIC/BIC and mse (with training and testing data) to compare different models.

```
LienStatus = c('','+lien_status','+lien_status+lien_status*loan_amount_000s')
Propertytype = c('','+property_type','+property_type+property_type*loan_amount_000s')
LoanPurpose= c('','+loan_purpose','+loan_purpose+loan_purpose*loan_amount_000s')
Ethnicity=c('','+applicant_ethnicity')
formulaSet =
paste('Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1
+Gender_ismissing+Race_ismissing',apply(expand.grid(LienStatus,Propertytype,LoanPurpose,Ethnicity),1,paste,collapse=''))
model<-lm(as.formula(formulaSet[1]),data=NYHMDA_new)
```

12. Compare models with MSE and AIC and BIC MSE of 50 models

```
mse<-rep(0,54)
for (i in 1:54) {
Modelset<-lm(as.formula(formulaSet[i]),data=NYHMDA.train)
mse[i] <- mean((Modelset$fitted.values-NYHMDA.test$Loan_Approval)^2)}

matrix <- data.frame(mse=mse,model=formulaSet,order(mse))
getwd()

## [1] "C:/Users/rital/OneDrive/Desktop"

write.csv(matrix,file="C:/Users/rital/OneDrive/Documents/Desktop/mse.HMDA.csv")
```

According to the MSE result attached in the appendix, the first model, which include all elements in the base model and missing variable identifiers, is the best as mean square error is the smallest among all models. More complex models, in our case, such as models include loan purpose and lien status fail the MSE test as these models might fit the training dataset too well and unfortunately overfitting.

Below we will then use another criteria, AIC and BIC, to investigate which model best explain the dataset.

```
modelStat = data.frame(formulaSet=rep(NA,54),AIC = rep(NA,54), BIC = rep(NA,54))
rowNum = 1
for(i in 1:54){
   currentFit = lm(as.formula(formulaSet[i]),data=NYHMDA_new)
   modelStat[rowNum,] = c(i,AIC(currentFit), BIC(currentFit))
   rowNum = rowNum + 1
}
print(modelStat)
```

According to the result attached for AIC and BIC, even though extra variables are penalized, the most complex model, which contains interaction terms and all variables in reduced models is the best. The result confirm our previous analysis that complex models overfitting the dataset as the last model, in reality, yields the largest MSE. In conclusion, the first model is the best as we are using this dataset for prediction purpose and have to be cautious about overfitting the dataset.

13. Final Model

```
model<-lm(as.formula(formulaSet[1]),data=NYHMDA_new)
```

Our final model includes loan amount, income, gender, race, and whether income, race, and gender are missing in the original dataset. Except for income and sex, other variables are all highly significant. From the regression result of this model (in the appendix), we know that the higher the loan amount, the lower the loan approval possibility. As for race, loan approval rates for all other races are higher than that for American Indian or Alaska Native. And male applicants are easier to get loans than female applicants, but the difference is not significant. When we look at the income item, we find that applicants who didn't disclose their income have higher loan approval rate. The reason maybe that certain types of loans don't attach importance to applicants' income level. Instead, they may appreciate applicant's ability to repay. For example, some people buy house for leasing, then the financial institute should look at the potential future cash flow from leasing, rather than applicant's income level.

14. Customer segments and average loan approval rate for each customer segmentation

In our final model, we included in two segment variables which are gender and race. And race is the only significant demographic variables. Therefore, we have 10 customer segments and their corresponding average loan approval rates. The results for average loan approval rate is attached in the appendix.

For each segment, we can calculate average loan approval based on intercept, loan amount coefficient, applicant income coefficient, applicant sex coefficient, and race coefficient. From the results we can see that the No. 10 segment White Male has the highest approval rate, 0.688; No. 1 segment American Indian Female has the lowest approval rate, 0.44.

15. Suggestions and recommendations

From our analysis, we find that the loan approval rate is mainly influenced by loan amount, race, and gender. Lower loan amount, White, or Male leads to higher approval rate. From mortgage borrowers' perspective, they should not require very high loan amount if they want to improve their loan approval rate. Especially, they should calculate how much they can afford monthly.

From lenders' perspective, this analysis can be used to set threshold for candidates with different situations. For example, given that loan amount is an important influencer, mortgage lenders can set appropriate loan amount and refuse too high loan amount requirements. Besides, with this analysis, mortgage lenders can educate people how to improve their loan approval rate. This is a win-win strategy because both borrowers and lenders will operate in a healthier financial environment.

Appendix

R code and regression analysis

5.1 Loan action taken

```
hist(NYHMDA_new$action_taken,col="lightblue",main="Loan action taken",ylab="Count",xlab="Actions taken")
legend('topright',c("1: Loan originated","2: Application approved but not accepted","3: Denied","4: Withdrawed" ,"5: File closed for incompleteness","6: Loan purchased by institution","7: Preapproval denied"),cex=0.6)
```

5.2 Ethnicity

```
hist(as.numeric(NYHMDA_new$applicant_ethnicity),col="lightblue",main="Applicant ethnicity",ylab="Count",xlab="Ethnicity",xaxt="n")
axis(side=1,at=c(1,2),labels=c("Hispanic and Latino","Non hispanic"))
```

5.3 Race

```
hist(as.numeric(NYHMDA_new$applicant_race_1),col="lightblue",main="Applicant race",ylab="Count",xlab="Race")
legend('top',c("1: American Indian or Alaska Native","2: Asian","3: Black or African American","4: Native Hawaiian or Other Pacific Islander" ,"5: White"),cex=0.6)
```

5.4 Gender

```
hist(as.numeric(NYHMDA  new$applicant  sex),col="lightblue",main="Applicant gender",ylab="Count",xlab="Gender",xaxt="n")
axis(side=1,at=c(1,2),labels=c("Male","Female"))
```

5.5 Applicant income

```
NYHMDA_new$applicant_income_000s <- as.numeric(NYHMDA_new$applicant_income_000s)
plot(density(NYHMDA_new$applicant_income_000s),xlab="Income",ylab="Density",main="Applicant income level",xlim=c(1,300),col="darkgreen")
```

5.6 Loan amount

```
plot(density(as.numeric(NYHMDA  new$loan_amount_000s)),xlab="Loan amount",ylab="Density",main="Applicant loan amount",xlim=c(1,1000),col="darkgreen")
```

5.7 Purchaser type

```
hist(NYHMDA_new$purchaser_type,col="lightblue", main="Purchaser Type", ylab="Count", xlab="Purchaser Type",xaxt="n")
axis(side=1,at=c(0,1,2,3,4,5,6,7,8,9),labels=c("0","1","2","3","4","5","6","7","8","9"))
legend('topright',c("0: Loan was not originated or was not sold in calendar year","1: Fannie Mae","2: Ginnie Mae","3: Freddie Mac","4: Farmer Mac"
                    ,"5: Private securitization","6: Commercial bank, savings bank or savings association",
                    "7: Life insurance company, credit union, mortgage bank, or finance company",
                    "8: Affiliate institution", "9: Other"),cex=0.6)
```

5.8 Lien Status

```
hist(as.numeric(NYHMDA_new$lien_status),col="lightblue",
    main="Lien status",ylab="Count",xlab="Lien status",xaxt="n")
axis(side=1,at=c(1,2,3),labels=c("First lien","Subordinate","Not secured"))
```

5.9 Loan purpose

```
hist(NYHMDA_new$loan_purpose,col="lightblue",
    main="Loan purpose type",ylab="Count",xlab="Loan purpose",xaxt="n")
axis(side=1,at=c(1,2,3),labels=c("Home purchase","Home improvement","Refinancing"))
```

5.10 Property type

```
hist(NYHMDA_new$property_type,col="lightblue",
    main="Property Type",
    ylab="Count",
    xlab="Property Type",xaxt="n")
axis(side=1,at=c(1,2,3),labels=c("1","2","3"))
legend('topright',c("1: One-to-four family dwelling (other than manufactured housing)",
                    "2: Manufactured housing",
                    "3: Multifamily dwelling"),cex=0.6)
```

5.11 Owner occupancy

```
hist(as.numeric(NYHMDA_new$owner_occupancy),col="lightblue",
    main="Owner Occupancy",
    ylab="Count",
    xlab="Owner Occupancy",xaxt="n")
axis(side=1,at=c(1,2),labels=c("1","2"))
legend('top',c("1: Owner-occupied as a principal dwelling",
           "2: Not owner-occupied as a principal dwelling"),cex=0.6)
```

5.12 Loan type

```
plot(NYHMDA_new$loan_type_name,col="lightblue",
    main="Loan Type",
    ylab="Count",
    xlab="Loan Type",xaxt="n")
axis(side=1,at=c(1,2,3,4),labels=c("1","2","3","4"))
legend('top',c("1: Conventional",
           "2: FHA-insured","3: VA-guaranteed","4: FSA/RHS-guaranteed"),cex=0.6)
```

6.1 Income level and loan actions taken

```
library(ggplot2)
ggplot(NYHMDA_new,aes(x=NYHMDA_new$applicant_income_000s,y=NYHMDA_new$action_taken,group=factor(1)))+
  geom_point()+
  labs(x="Income",y="Loan action taken",title="Loan actions taken and income level")
```

6.2 Loan purpose and loan actions taken

```
library(dplyr)

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(tidyr)
library(ggplot2)
NYHMDA_new %>%
  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 And Action Taken') +
  theme_bw()
```

6.3. Gender and loan actions taken

```
NYHMDA_new %>%
  group_by(applicant_sex_name,action_taken_name) %>%
  summarise(CountGender = n() ) %>%

  ggplot(aes(x = applicant_sex_name,y = CountGender,fill =(action_taken_name))) +
  geom_bar(stat='identity',colour="white") +
  labs(x = 'Gender', y = 'Count', title = 'Gender And Action Taken') +
  theme_bw()
```

6.4. Ethnicity and loan actions taken

```
NYHMDA_new %>%
  group_by(applicant_ethnicity_name,action_taken_name) %>%
  summarise(CountEthnicity = n() ) %>%

  ggplot(aes(x = applicant_ethnicity_name,y = CountEthnicity,fill =(action_taken_name))) +
  geom_bar(stat='identity',colour="white") +
  labs(x = 'Ethnicity', y = 'Count', title = 'Ethnicity And Action Taken') +
  theme_bw()
```

6.5. Lien status and loan actions taken

```
NYHMDA_new %>%
  group_by(lien_status_name,action_taken_name) %>%
  summarise(CountLienStatus = n() ) %>%

  ggplot(aes(x = lien_status_name,y = CountLienStatus,fill =(action_taken_name))) +
  geom_bar(stat='identity',colour="white") +
  labs(x = 'Lien Status', y = 'Count', title = 'Lien Status And Action Taken') +
  theme_bw()
```

6.6. Loan amount and income level

```
ip<-ggplot(data=NYHMDA_new,aes(x=NYHMDA_new$applicant_income_000s,y=NYHMDA_new$loan_amount_000s))+
  geom_point()+
  facet_wrap(~NYHMDA_new$action_taken)+
  labs(x="Income level",y="Loan amount",title="Loan actions taken,loan amount and income level")
ip+xlim(0,7500)+ylim(0,10000)
```

7.1 Define “Loan_Approval”

```
NYHMDA_new$Loan_Approval = 0
NYHMDA_new[NYHMDA_new$action_taken==2|NYHMDA_new$action_taken==1|NYHMDA_new$action_taken==6,]$Loan_Approval
<-1
NYHMDA_new[NYHMDA_new$action_taken==3|NYHMDA_new$action_taken==5|NYHMDA_new$action_taken==7,]$Loan_Approval
<-0
NYHMDA  new$completeness = "0"
NYHMDA_new[NYHMDA_new$action_taken==4,]$completeness<-"1"
```

7.2 Define “Income_level”

```
summary(NYHMDA_new$applicant_income_000s)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    63.0    90.0   133.2   130.0  9999.0

NYHMDA_new$Income_Level = "0"
#Middle income individuals
NYHMDA_new[NYHMDA_new$applicant_income_000s<130&NYHMDA_new$applicant_income_000s>=63,]$Income_Level<-"2"
#High income individuals
NYHMDA_new[NYHMDA_new$applicant_income_000s>=130,]$Income_Level<-"1"
#Low income individuals
NYHMDA_new[NYHMDA_new$applicant_income_000s<63,]$Income_Level<-"3"
```

8.1. Difference in loan amounts between loan approved and denied

```
t.test(NYHMDA_new[NYHMDA_new$Loan_Approval==1,]$loan_amount_000s,NYHMDA_new[NYHMDA_new$Loan_Approval==0,]$lo
an_amount_000s)

##
##  Welch Two Sample t-test
##
## data:  NYHMDA_new[NYHMDA_new$Loan_Approval == 1, ]$loan_amount_000s and
NYHMDA_new[NYHMDA_new$Loan_Approval == 0, ]$loan_amount_000s
## t = 3.8179, df = 291680, p-value = 0.0001346
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.801141 21.150754
## sample estimates:
## mean of x mean of y
##  337.6453  323.6694
```

8.2. Difference in loan approval rate between high and low income applicants

```
t.test(NYHMDA  new[NYHMDA_new$Income_Level==1,]$Loan_Approval,NYHMDA_new[NYHMDA_new$Income_Level==3,]$Loan_Approval,alternative = "greater")

##
##  Welch Two Sample t-test
##
## data:  NYHMDA_new[NYHMDA_new$Income_Level == 1, ]$Loan_Approval and NYHMDA_new[NYHMDA_new$Income_Level == 3, ]$Loan_Approval
## t = 33.696, df = 218180, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.06473505        Inf
## sample estimates:
## mean of x mean of y
## 0.6936430 0.6255857

table(NYHMDA_new$Loan_Approval,NYHMDA_new$Income_Level)

##
##         1      2      3
##   0  33850  61407  40673
##   1  76642 159124  67958
```

9. Regression results for the initial model

```
summary(initial_model)
##
## Call:
## lm(formula = NYHMDA_new$Loan_Approval ~ NYHMDA_new$loan_amount_000s +
##     NYHMDA_new$applicant_income_000s + factor(NYHMDA_new$applicant_sex) +
##     factor(NYHMDA_new$applicant_race_1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8821 -0.6784  0.2833  0.2874  0.6008
##
## Coefficients:
##                                      Estimate Std. Error t value
## (Intercept)                          4.663e-01  1.112e-02  41.915
## NYHMDA_new$loan_amount_000s          1.659e-06  6.021e-07   2.755
## NYHMDA_new$applicant_income_000s    -1.755e-05  2.833e-06  -6.197
## factor(NYHMDA_new$applicant_sex)2   -3.521e-02  1.575e-03 -22.351
## factor(NYHMDA_new$applicant_race_1)2  1.969e-01  1.144e-02  17.213
## factor(NYHMDA_new$applicant_race_1)3  8.745e-02  1.143e-02   7.649
## factor(NYHMDA_new$applicant_race_1)4  8.418e-02  1.706e-02   4.936
## factor(NYHMDA_new$applicant_race_1)5  2.516e-01  1.113e-02  22.598
##                                     Pr(>|t|)
## (Intercept)                          < 2e-16 ***
## NYHMDA_new$loan_amount_000s          0.00587 **
## NYHMDA_new$applicant_income_000s    5.77e-10 ***
## factor(NYHMDA_new$applicant_sex)2    < 2e-16 ***
## factor(NYHMDA_new$applicant_race_1)2  < 2e-16 ***
## factor(NYHMDA_new$applicant_race_1)3 2.04e-14 ***
## factor(NYHMDA_new$applicant_race_1)4 7.99e-07 ***
## factor(NYHMDA_new$applicant_race_1)5  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4595 on 439646 degrees of freedom
## Multiple R-squared:  0.01151,    Adjusted R-squared:  0.0115
## F-statistic: 731.5 on 7 and 439646 DF,  p-value: < 2.2e-16
```

12. Compare models with mse and AIC and BIC MSE of 50 models
Regression results with mse as single criteria

	mse	model	order.mse.
1	0.20292649 1	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing	1
2	0.20450402 8	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status	4
3	0.20455103 3	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+lien_status*loan_amount_000s	7
4	0.20313814 2	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +property_type	28
5	0.20466397 3	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+property_type	31
6	0.20470198 7	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+lien_status*loan_amount_000s+property_type	34
7	0.20315865 6	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +property_type+property_type*loan_amount_000s	2
8	0.20473980 2	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+property_type+property_type*loan_amount_000s	3
9	0.20485233 7	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+lien_status*loan_amount_000s+property_type+property_type*loan_amount_000s	5
10	0.21380431 7	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +loan_purpose	6
11	0.21508467 6	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+loan_purpose	8
12	0.21512700 8	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+lien_status*loan_amount_000s+loan_purpose	9
13	0.21404803 8	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +property_type+loan_purpose	29
14	0.21527961	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing	30

	4	+lien_status+property_type+loan_purpose	
15	0.215312625	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+lien_status*loan_amount_000s+property_type+loan_purpose	32
16	0.21411166 1	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +property_type+property_type*loan_amount_000s+loan_purpose	33
17	0.21541454 1	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+property_type+property_type*loan_amount_000s+loan_purpose	35
18	0.21554673 1	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+lien_status*loan_amount_000s+property_type+property_type*loan_amount_000s+loan_purpose	36
19	0.21380959 1	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +loan_purpose+loan_purpose*loan_amount_000s	10
20	0.21508787 1	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+loan_purpose+loan_purpose*loan_amount_000s	19
21	0.21513773 9	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+lien_status*loan_amount_000s+loan_purpose+loan_purpose*loan_amount_000s	13
22	0.21405816	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +property_type+loan_purpose+loan_purpose*loan_amount_000s	22
23	0.21528635 4	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+property_type+loan_purpose+loan_purpose*loan_amount_000s	16
24	0.21532766 6	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+lien_status*loan_amount_000s+property_type+loan_purpose+loan_purpose*loan_amount_000s	25
25	0.21411504 1	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +property_type+property_type*loan_amount_000s+loan_purpose+loan_purpose*loan_amount_000s	37
26	0.21541488	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+property_type+property_type*loan_amount_000s+loan_purpose+loan_purpose*loan_amount_000s	46
27	0.21555541	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+lien_status*loan_amount_000s+property_type+property_type*loan_amount_000	40

		s+loan_purpose+loan_purpose*loan_amount_000s	
28	0.20334126 8	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +applicant_ethnicity	49
29	0.20493338 8	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+applicant_ethnicity	43
30	0.20497966 8	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+lien_status*loan_amount_000s+applicant_ethnicity	52
31	0.20356475 9	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +property_type+applicant_ethnicity	11
32	0.20510351 3	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+property_type+applicant_ethnicity	20
33	0.20514057 2	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+lien_status*loan_amount_000s+property_type+applicant_ethnicity	12
34	0.20358341 9	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +property_type+property_type*loan_amount_000s+applicant_ethnicity	21
35	0.20517609 2	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+property_type+property_type*loan_amount_000s+applicant_ethnicity	14
36	0.20528539 8	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+lien_status*loan_amount_000s+property_type+property_type*loan_amount_000s+applicant_ethnicity	23
37	0.21426022 6	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +loan_purpose+applicant_ethnicity	15
38	0.21555249 6	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+loan_purpose+applicant_ethnicity	24
39	0.21559408 8	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+lien_status*loan_amount_000s+loan_purpose+applicant_ethnicity	17
40	0.21451733 3	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +property_type+loan_purpose+applicant_ethnicity	26
41	0.21575923	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+property_type+loan_purpose+applicant_ethnicity	18

42	0.21579128 7	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+lien_status*loan_amount_000s+property_type+loan_purpose+applicant_ethnicity	38
43	0.21457767 7	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +property_type+property_type*loan_amount_000s+loan_purpose+applicant_ethnicity	27
44	0.21588976 6	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+property_type+property_type*loan_amount_000s+loan_purpose+applicant_ethnicity	47
45	0.21601835 8	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+lien_status*loan_amount_000s+property_type+property_type*loan_amount_000s+loan_purpose+applicant_ethnicity	39
46	0.21426565 1	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +loan_purpose+loan_purpose*loan_amount_000s+applicant_ethnicity	48
47	0.21555579 3	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+loan_purpose+loan_purpose*loan_amount_000s+applicant_ethnicity	41
48	0.21560498 1	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+lien_status*loan_amount_000s+loan_purpose+loan_purpose*loan_amount_000s+applicant_ethnicity	50
49	0.21452782 1	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +property_type+loan_purpose+loan_purpose*loan_amount_000s+applicant_ethnicity	42
50	0.21576624 5	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+property_type+loan_purpose+loan_purpose*loan_amount_000s+applicant_ethnicity	51
51	0.21580666 7	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+lien_status*loan_amount_000s+property_type+loan_purpose+loan_purpose*loan_amount_000s+applicant_ethnicity	44
52	0.21458140 9	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +property_type+property_type*loan_amount_000s+loan_purpose+loan_purpose*loan_amount_000s+applicant_ethnicity	53
53	0.21589021 3	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+property_type+property_type*loan_amount_000s+loan_purpose+loan_purpose*loan_amount_000s+applicant_ethnicity	45
54	0.21602737 9	Loan_Approval~loan_amount_000s+applicant_income_000s+Income_ismissing+applicant_sex+applicant_race_1+Gender_ismissing+Race_ismissing +lien_status+lien_status*loan_amount_000s+property_type+property_type*loan_amount_000	54

		s+loan_purpose+loan_purpose*loan_amount_000s+applicant_ethnicity	

Regression analysis with AIC and BIC as single criteria

```
##    formulaSet      AIC      BIC
## 1           1 555741.6 555873.5
## 2           2 552356.4 552510.3
## 3           3 552278.4 552454.3
## 4           4 555332.2 555475.1
## 5           5 552048.0 552212.9
## 6           6 551982.1 552169.0
## 7           7 555263.2 555417.1
## 8           8 551835.8 552011.7
## 9           9 551657.7 551855.6
## 10         10 532232.4 532375.3
## 11         11 529299.9 529464.8
## 12         12 529225.3 529412.2
## 13         13 531736.7 531890.6
## 14         14 528904.7 529080.6
## 15         15 528843.1 529041.0
## 16         16 531550.2 531715.1
## 17         17 528531.8 528718.7
## 18         18 528314.4 528523.3
## 19         19 532216.5 532370.5
## 20         20 529290.5 529466.4
## 21         21 529207.2 529405.1
## 22         22 531706.8 531871.7
## 23         23 528884.9 529071.8
## 24         24 528813.6 529022.5
## 25         25 531536.9 531712.8
## 26         26 528529.3 528727.2
## 27         27 528297.3 528517.2
## 28         28 554921.4 555064.3
## 29         29 551498.4 551663.3
## 30         30 551421.8 551608.7
## 31         31 554487.6 554641.5
## 32         32 551169.1 551345.0
## 33         33 551104.8 551302.7
## 34         34 554423.5 554588.4
## 35         35 550964.7 551151.6
## 36         36 550791.6 551000.5
## 37         37 531281.5 531435.4
## 38         38 528314.7 528490.6
## 39         39 528241.6 528439.5
## 40         40 530756.6 530921.5
## 41         41 527893.8 528080.7
## 42         42 527834.0 528042.9
## 43         43 530578.4 530754.3
## 44         44 527531.5 527729.4
## 45         45 527319.9 527539.8
## 46         46 531265.0 531430.0
## 47         47 528304.9 528491.8
## 48         48 528223.0 528431.8
## 49         49 530725.4 530901.3
## 50         50 527873.0 528070.9
## 51         51 527803.4 528023.3
## 52         52 530563.9 530750.8
## 53         53 527528.3 527737.2
## 54         54 527301.9 527532.7
```

13. Regression results of the final model

```
summary(model)
```

```
##
## Call:
## lm(formula = as.formula(formulaSet[1]), data = NYHMDA_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0093 -0.6448  0.3112  0.3136  0.5652
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.436e-01  1.103e-02  40.230  < 2e-16 ***
## loan_amount_000s     -3.087e-06  6.011e-07  -5.135 2.82e-07 ***
## applicant_income_000s -3.189e-06  2.821e-06  -1.130    0.258
## Income_ismissing1     4.739e-02  2.296e-03  20.639  < 2e-16 ***
## applicant_sex2       -2.618e-03  1.619e-03  -1.616    0.106
## applicant_race_12     2.053e-01  1.134e-02  18.106  < 2e-16 ***
## applicant_race_13     8.817e-02  1.133e-02   7.784 7.06e-15 ***
## applicant_race_14     8.861e-02  1.690e-02   5.244 1.57e-07 ***
## applicant_race_15     2.456e-01  1.104e-02  22.250  < 2e-16 ***
## Gender_ismissing1     2.731e-01  3.722e-03  73.383  < 2e-16 ***
## Race_ismissing1      -1.668e-01  3.297e-03 -50.593  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4552 on 439643 degrees of freedom
## Multiple R-squared:  0.02972,    Adjusted R-squared:  0.0297
## F-statistic:  1347 on 10 and 439643 DF,  p-value: < 2.2e-16
```

14. Customer segments and average loan approval rate for each customer segmentation

Gender	Race	Group	Intercept	Loan amount coefficient	Median loan amount	Applicant income coefficient	Median applicant income	Applicant sex coefficient	Race coefficient	Average loan approval
Female	**American Indian**	1	0.4436	-3.087E-06	208	-0.000003188	90	-0.00262	0	0.440053
Female	**Asian**	2	0.4436	-3.087E-06	208	-0.000003188	90	-0.00262	0.2053	0.645353
Female	**Black or African American**	3	0.4436	-3.087E-06	208	-0.000003188	90	-0.00262	0.08817	0.528223
Female	**Native Hawaiian**	4	0.4436	-3.087E-06	208	-0.000003188	90	-0.00262	0.08861	0.528663
Female	**White**	5	0.4436	-3.087E-06	208	-0.000003188	90	-0.00262	0.2456	0.685653
Male	**American Indian**	6	0.4436	-3.087E-06	208	-0.000003188	90	0	0	0.442671
Male	**Asian**	7	0.4436	-3.087E-06	208	-0.000003188	90	0	0.2053	0.647971

Male	**Black or African American**	8	0.4436	-3.087E-06	208	-0.000003188	90	0	0.08817	0.530841
Male	**Native Hawaiian**	9	0.4436	-3.087E-06	208	-0.000003188	90	0	0.08861	0.531281
Male	**White**	10	0.4436	-3.087E-06	208	-0.000003188	90	0	0.2456	0.688271