Interim Report

Shine Yao,Alex Nguyen,Sharon Bao,Charlie Chen

3/14/2018

## R Markdown

\*The URL for our Team GitHub repository is:<https://github.com/shineloveyc/Marketing-Analytics-Project>

\*Color Scheme Reference:<http://colorbrewer2.org/#type=sequential&scheme=PuBuGn&n=3>

\*Small parts of the code refer to :<https://www.kaggle.com/jboysen/ny-home-mortgage/kernels>

## Load the package

library(tidyverse)  
library(ggplot2)  
library(rpart)  
library(rpart.plot)  
library(caTools)  
library(caret)  
library(DT)  
library(knitr)  
library(dplyr)  
library(rcompanion)

## 1. What is the data problem? What is the final managerial objective?

The data set organized under The Home Mortgage Disclosure Act (HMDA) covers all mortgage decision made in 2015 for the state of New York. Under this Act, financial institutions with offices in mmetropolitan areas are required to disclose reports and documents about mortgages data; for example the disposition of each application for mortgage credit, the characteristics of mortgages initiated during the year, or personal demographic information about loans applicants. In this data set, these mortgages information are revealed to help answer questions of: whether lenders are serving housing needs of their communities, whether public officials can rely on these data to make decisions or issue policies, and whether there are certain lending patterns. The data set is pulled by the Consumer Finance Protection Board. Thus, their goal would be to figure out where in New York mortgages are most likely to get approved. Another managerial objective is also to analyze which factors (such as gender, ethnicity, income) play a major role in mortgage decisions. From here, they can make necessary prediction about what probability of getting a loan approved based on the criteria ,and also reduce the workload by focusing on examing those key preidctors instead of all variables.

## 2. Describe the measurement types of each variable

*Interval Variable: as\_of\_year* Ratio Variables: applicant\_income\_000s,hud\_median\_family\_income, loan\_amount\_000s, number\_of\_1\_to\_4\_family\_units, number\_of\_owner\_occupied\_units, minority\_population, population, rate\_spread, tract\_to\_msamd\_income \*Nominal Variables: all other variables (eg: action\_taken, agency\_name, country\_name …)

## 3. How do you handle missing data in this dataset?

After we know what the variables represent, we need to check if there exist missing data in the dataset and handle the problem in order to do further analysis.

homeMortgage = read.csv("ny\_hmda\_2015.csv")  
View(homeMortgage)

## Warning: running command ''/usr/bin/otool' -L '/Library/Frameworks/  
## R.framework/Resources/modules/R\_de.so'' had status 1

sapply(homeMortgage, function(x) sum(is.na(x)))

## action\_taken action\_taken\_name   
## 0 0   
## agency\_code agency\_abbr   
## 0 0   
## agency\_name applicant\_ethnicity   
## 0 0   
## applicant\_ethnicity\_name applicant\_income\_000s   
## 0 61003   
## applicant\_race\_1 applicant\_race\_2   
## 0 438044   
## applicant\_race\_3 applicant\_race\_4   
## 439567 439639   
## applicant\_race\_5 applicant\_race\_name\_1   
## 439645 0   
## applicant\_race\_name\_2 applicant\_race\_name\_3   
## 0 0   
## applicant\_race\_name\_4 applicant\_race\_name\_5   
## 0 0   
## applicant\_sex applicant\_sex\_name   
## 0 0   
## application\_date\_indicator as\_of\_year   
## 0 0   
## census\_tract\_number co\_applicant\_ethnicity   
## 1667 0   
## co\_applicant\_ethnicity\_name co\_applicant\_race\_1   
## 0 0   
## co\_applicant\_race\_2 co\_applicant\_race\_3   
## 439166 439622   
## co\_applicant\_race\_4 co\_applicant\_race\_5   
## 439650 439652   
## co\_applicant\_race\_name\_1 co\_applicant\_race\_name\_2   
## 0 0   
## co\_applicant\_race\_name\_3 co\_applicant\_race\_name\_4   
## 0 0   
## co\_applicant\_race\_name\_5 co\_applicant\_sex   
## 0 0   
## co\_applicant\_sex\_name county\_code   
## 0 1192   
## county\_name denial\_reason\_1   
## 0 381644   
## denial\_reason\_2 denial\_reason\_3   
## 427782 437280   
## denial\_reason\_name\_1 denial\_reason\_name\_2   
## 0 0   
## denial\_reason\_name\_3 edit\_status   
## 0 355634   
## edit\_status\_name hoepa\_status   
## 0 0   
## hoepa\_status\_name lien\_status   
## 0 0   
## lien\_status\_name loan\_purpose   
## 0 0   
## loan\_purpose\_name loan\_type   
## 0 0   
## loan\_type\_name msamd   
## 0 36375   
## msamd\_name owner\_occupancy   
## 0 0   
## owner\_occupancy\_name preapproval   
## 0 0   
## preapproval\_name property\_type   
## 0 0   
## property\_type\_name purchaser\_type   
## 0 0   
## purchaser\_type\_name respondent\_id   
## 0 0   
## sequence\_number state\_code   
## 0 0   
## state\_abbr state\_name   
## 0 0   
## hud\_median\_family\_income loan\_amount\_000s   
## 1667 0   
## number\_of\_1\_to\_4\_family\_units number\_of\_owner\_occupied\_units   
## 2236 2049   
## minority\_population population   
## 1696 1696   
## rate\_spread tract\_to\_msamd\_income   
## 430914 1771

After checking, we find out that missing data exist for these nominal variables: “applicant\_race\_2”, “applicant\_race\_3”, “applicant\_race\_4”, “applicant\_race\_5”, “census\_tract\_number”, “co\_applicant\_race\_2”, “co\_applicant\_race\_3”, “co\_applicant\_race\_4”, “co\_applicant\_race\_5”, “county\_code”, “denial\_reason\_1”, “denial\_reason\_2”, “denial\_reason\_3”, “edit status”, “msamd”. Moreover, missing data exist for these ratio variables: “applicant\_income\_000s”, “hud\_median\_family\_income”, “number\_of\_1\_to\_4\_family\_units”, “number\_of\_owner\_occupied\_units”, “minority\_population”, “population”, “rate\_spread”, “tract\_to\_msamd\_income”.

In order to drop the missing data in the dataset, we first categorize the missing data into three categories: MCAR, MAR, and MNAR. We find out that all missing data for nominal variables are NMAR, as they are all missing due to applicant left blank for description variables, or the option is not appliable to them. Therefore, we will fill these data with “NA” or word explainations to indicate that these data are not being collected or not apppliable to the applicants.

#replace missing data with word explainations  
homeMortgage$applicant\_race\_2 [is.na(homeMortgage$applicant\_race\_2)] <- "No second race"  
homeMortgage$applicant\_race\_3 [is.na(homeMortgage$applicant\_race\_3)] <- "No third race"  
homeMortgage$applicant\_race\_4 [is.na(homeMortgage$applicant\_race\_4)] <- "No fourth race"  
homeMortgage$applicant\_race\_5 [is.na(homeMortgage$applicant\_race\_5)] <- "No fifth race"  
homeMortgage$co\_applicant\_race\_2 [is.na(homeMortgage$co\_applicant\_race\_2)] <- "No second race"  
homeMortgage$co\_applicant\_race\_3 [is.na(homeMortgage$co\_applicant\_race\_3)] <- "No thrid race"  
homeMortgage$co\_applicant\_race\_4 [is.na(homeMortgage$co\_applicant\_race\_4)] <- "No fourth race"  
homeMortgage$co\_applicant\_race\_5 [is.na(homeMortgage$co\_applicant\_race\_5)] <- "No fifth race"  
homeMortgage$county\_code [is.na(homeMortgage$county\_code)] <- "Information missing"  
homeMortgage$denial\_reason\_1 [is.na(homeMortgage$denial\_reason\_1)] <- "No denial reason"  
homeMortgage$denial\_reason\_2 [is.na(homeMortgage$denial\_reason\_2)] <- "No second denial reason"  
homeMortgage$denial\_reason\_3 [is.na(homeMortgage$denial\_reason\_3)] <- "No third denial reason"  
homeMortgage$edit\_status [is.na(homeMortgage$edit\_status)] <- "No edit status"  
homeMortgage$msamd[is.na(homeMortgage$msamd)] <- "Information missing"  
homeMortgage$census\_tract\_number [is.na(homeMortgage$census\_tract\_number)] <- "Information missing"

All missing data for ratio variables are MAR, therefore we will use the mean to replace the missing data in the 11 ratio variable columns.

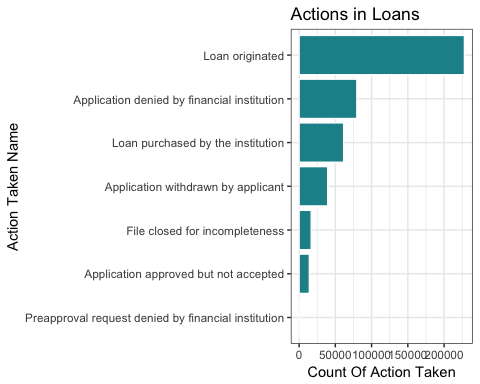
#replace missing data with the mean of each variable  
homeMortgage$applicant\_income\_000s[is.na(homeMortgage$applicant\_income\_000s)] <- round( mean(homeMortgage$applicant\_income\_000s, na.rm = TRUE))  
homeMortgage[is.na(homeMortgage$hud\_median\_family\_income), ]$hud\_median\_family\_income <- round( mean(homeMortgage$hud\_median\_family\_income, na.rm = TRUE))  
homeMortgage[is.na(homeMortgage$number\_of\_1\_to\_4\_family\_units), ]$number\_of\_1\_to\_4\_family\_units <- round( mean(homeMortgage$number\_of\_1\_to\_4\_family\_units, na.rm = TRUE))  
homeMortgage[is.na(homeMortgage$number\_of\_owner\_occupied\_units), ]$number\_of\_owner\_occupied\_units <- round( mean(homeMortgage$number\_of\_owner\_occupied\_units, na.rm = TRUE))  
homeMortgage[is.na(homeMortgage$minority\_population), ]$minority\_population <- round( mean(homeMortgage$minority\_population, na.rm = TRUE))  
homeMortgage[is.na(homeMortgage$population), ]$population <- round( mean(homeMortgage$population, na.rm = TRUE))  
homeMortgage[is.na(homeMortgage$rate\_spread), ]$rate\_spread <- round( mean(homeMortgage$rate\_spread, na.rm = TRUE))  
homeMortgage[is.na(homeMortgage$tract\_to\_msamd\_income), ]$tract\_to\_msamd\_income <- round( mean(homeMortgage$tract\_to\_msamd\_income, na.rm = TRUE))  
#recheck if all missing data have been handled  
sapply(homeMortgage, function(x) sum(is.na(x)))

## action\_taken action\_taken\_name   
## 0 0   
## agency\_code agency\_abbr   
## 0 0   
## agency\_name applicant\_ethnicity   
## 0 0   
## applicant\_ethnicity\_name applicant\_income\_000s   
## 0 0   
## applicant\_race\_1 applicant\_race\_2   
## 0 0   
## applicant\_race\_3 applicant\_race\_4   
## 0 0   
## applicant\_race\_5 applicant\_race\_name\_1   
## 0 0   
## applicant\_race\_name\_2 applicant\_race\_name\_3   
## 0 0   
## applicant\_race\_name\_4 applicant\_race\_name\_5   
## 0 0   
## applicant\_sex applicant\_sex\_name   
## 0 0   
## application\_date\_indicator as\_of\_year   
## 0 0   
## census\_tract\_number co\_applicant\_ethnicity   
## 0 0   
## co\_applicant\_ethnicity\_name co\_applicant\_race\_1   
## 0 0   
## co\_applicant\_race\_2 co\_applicant\_race\_3   
## 0 0   
## co\_applicant\_race\_4 co\_applicant\_race\_5   
## 0 0   
## co\_applicant\_race\_name\_1 co\_applicant\_race\_name\_2   
## 0 0   
## co\_applicant\_race\_name\_3 co\_applicant\_race\_name\_4   
## 0 0   
## co\_applicant\_race\_name\_5 co\_applicant\_sex   
## 0 0   
## co\_applicant\_sex\_name county\_code   
## 0 0   
## county\_name denial\_reason\_1   
## 0 0   
## denial\_reason\_2 denial\_reason\_3   
## 0 0   
## denial\_reason\_name\_1 denial\_reason\_name\_2   
## 0 0   
## denial\_reason\_name\_3 edit\_status   
## 0 0   
## edit\_status\_name hoepa\_status   
## 0 0   
## hoepa\_status\_name lien\_status   
## 0 0   
## lien\_status\_name loan\_purpose   
## 0 0   
## loan\_purpose\_name loan\_type   
## 0 0   
## loan\_type\_name msamd   
## 0 0   
## msamd\_name owner\_occupancy   
## 0 0   
## owner\_occupancy\_name preapproval   
## 0 0   
## preapproval\_name property\_type   
## 0 0   
## property\_type\_name purchaser\_type   
## 0 0   
## purchaser\_type\_name respondent\_id   
## 0 0   
## sequence\_number state\_code   
## 0 0   
## state\_abbr state\_name   
## 0 0   
## hud\_median\_family\_income loan\_amount\_000s   
## 0 0   
## number\_of\_1\_to\_4\_family\_units number\_of\_owner\_occupied\_units   
## 0 0   
## minority\_population population   
## 0 0   
## rate\_spread tract\_to\_msamd\_income   
## 0 0

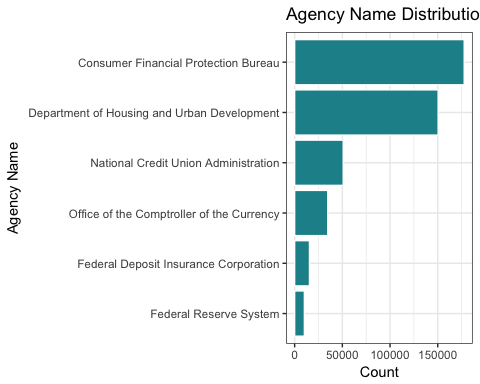
## 4. Create a table summarizing the range or variation in each variable. Add statistics (mean,median, standard deviation, etc.) as you deem necessary.

In this dataset, we have 11 continous variables and 67 discrete variables. For the 67 discrete variables, many of them are indicators for other variables. For example, “applicant\_race\_1” is the indicator variable for “applicant\_race\_name\_1”. Therefore, we will only present frequency tables for selected discrete variables that we deem important for analysis. Hence, the frequency tables for “action\_taken\_name”, “agency\_name”, “applicant\_ethnicity\_name”, “applicant\_sex\_name”, “county\_name”,“hoepa\_status\_name”,“lien\_status\_name”, “loan\_purpose\_name”,“loan\_type\_name”, “property\_type\_name”, “purchaser\_type\_name”, and “state\_name” are presented below. Please check the Appendix for table of modes for all discrete variables.

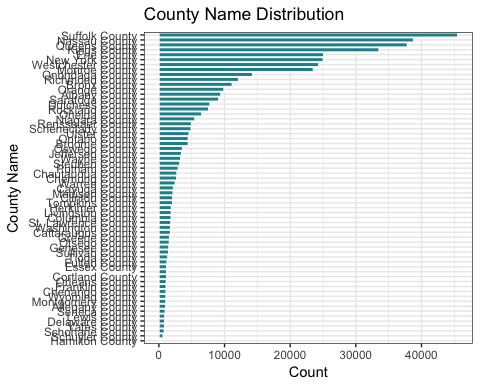
###for reference, not add in offical report  
# frequency distribution tables for selected discrete variables  
# "purchaser\_type\_name", and "state\_name"  
  
#Histogram for action\_taken\_name  
fillColor = "#1c9099"  
homeMortgageStatus <- homeMortgage %>%  
 group\_by(action\_taken\_name) %>%  
 summarize(count = n()) %>%  
 arrange(desc(count))  
  
homeMortgageStatus %>% ggplot(aes(x = reorder(action\_taken\_name,count), y = count)) +   
 geom\_bar(stat = 'identity',colour = "white", fill = fillColor) +  
 labs(x = 'Action Taken Name', y = 'Count Of Action Taken', title = 'Actions in Loans') +  
 coord\_flip() +   
 theme\_bw()



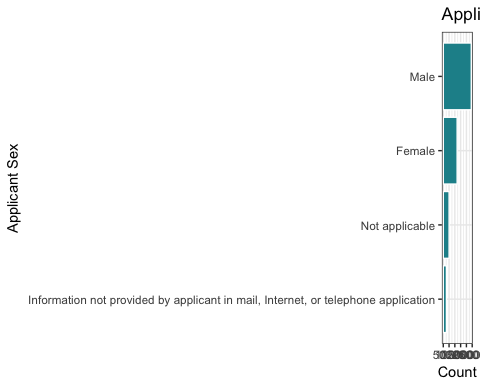
#Histogram for agency\_name  
fillColor = "#1c9099"  
agency\_name <- homeMortgage %>%  
 group\_by(agency\_name) %>%   
 summarize(count = n()) %>%  
 arrange(desc(count))   
  
agency\_name %>% ggplot(aes(x = reorder(agency\_name, count), y = count)) +   
 geom\_bar(stat = 'identity',colour="white", fill =fillColor) +  
 labs(x = 'Agency Name', y = 'Count', title = 'Agency Name Distribution') +  
 coord\_flip() +   
 theme\_bw()



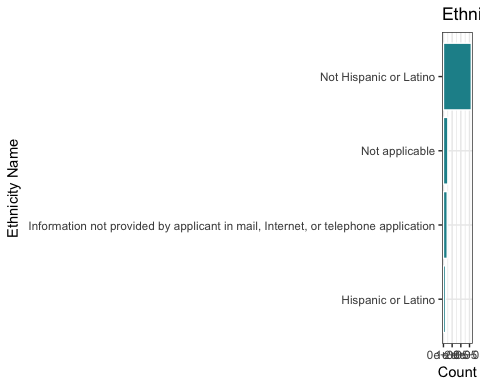
#Histogram for county\_name  
 fillColor = "#1c9099"  
 county\_name <- homeMortgage %>%  
 group\_by(county\_name) %>%   
 summarize(count = n()) %>%  
 arrange(desc(count))   
   
 county\_name %>% ggplot(aes(x = reorder(county\_name, count), y = count)) +   
 geom\_bar(stat = 'identity',colour="white", fill =fillColor) +  
 labs(x = 'County Name', y = 'Count', title = 'County Name Distribution') +  
 coord\_flip() +   
 theme\_bw()



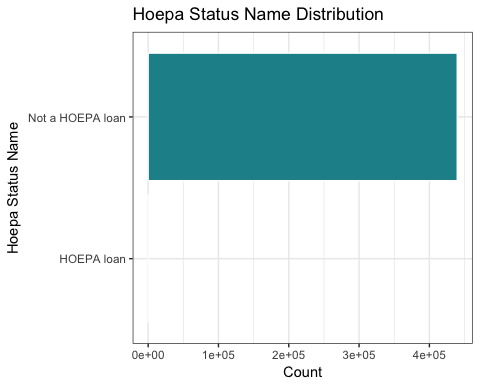
#Histogram for applicant\_sex\_name  
fillColor = "#1c9099"  
  
applicant\_sex\_name <- homeMortgage %>%  
 group\_by(applicant\_sex\_name) %>%   
 summarize(count = n()) %>%  
 arrange(desc(count))   
  
applicant\_sex\_name %>% ggplot(aes(x = reorder(applicant\_sex\_name, count), y = count)) +   
 geom\_bar(stat = 'identity',colour="white", fill =fillColor) +  
 labs(x = 'Applicant Sex', y = 'Count', title = 'Applicant Sex Distribution') +  
 coord\_flip() +   
 theme\_bw()



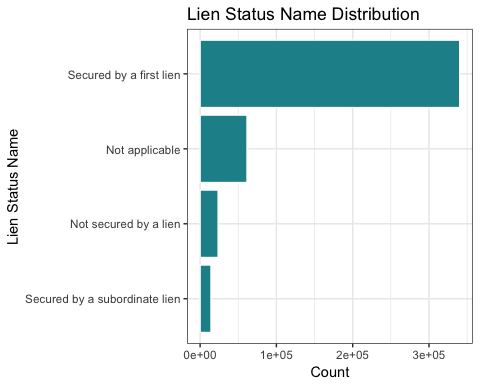
#Histogram for applicant\_ethnicity\_name  
ethnicity\_name <- homeMortgage %>%  
 group\_by(applicant\_ethnicity\_name) %>%   
 summarize(count = n()) %>%  
 arrange(desc(count))   
  
ethnicity\_name %>% ggplot(aes(x = reorder(applicant\_ethnicity\_name, count), y = count)) +   
 geom\_bar(stat = 'identity', colour="white", fill =fillColor) +  
 labs(x = 'Ethnicity Name', y = 'Count', title = 'Ethnicity Name Distribution') +  
 coord\_flip() +   
 theme\_bw()



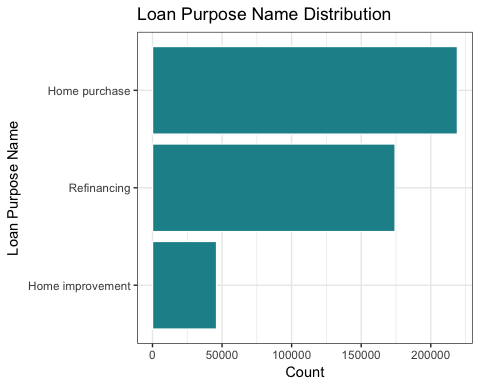
#Histogram for hoepa\_status\_name  
fillColor = "#1c9099"  
  
hoepa\_status\_name <- homeMortgage %>%  
 group\_by(hoepa\_status\_name) %>%   
 summarize(count = n()) %>%  
 arrange(desc(count))   
  
hoepa\_status\_name %>% ggplot(aes(x = reorder(hoepa\_status\_name, count), y = count)) +   
 geom\_bar(stat = 'identity',colour="white", fill =fillColor) +  
 labs(x = 'Hoepa Status Name', y = 'Count', title = 'Hoepa Status Name Distribution') +  
 coord\_flip() +   
 theme\_bw()



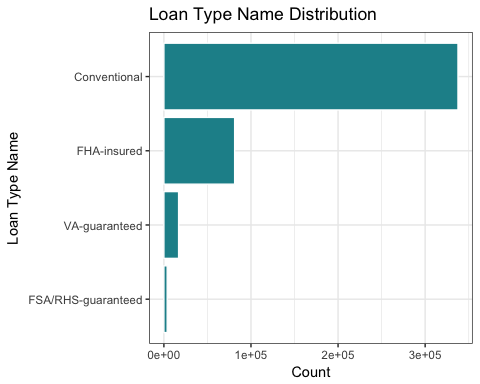
#Histogram for lien\_status\_name  
fillColor = "#1c9099"  
  
lien\_name <- homeMortgage %>%  
 group\_by(lien\_status\_name) %>%   
 summarize(count = n()) %>%  
 arrange(desc(count))   
  
lien\_name %>% ggplot(aes(x = reorder(lien\_status\_name, count), y = count)) +   
 geom\_bar(stat = 'identity',colour="white", fill =fillColor) +  
 labs(x = 'Lien Status Name', y = 'Count', title = 'Lien Status Name Distribution') +  
 coord\_flip() +   
 theme\_bw()



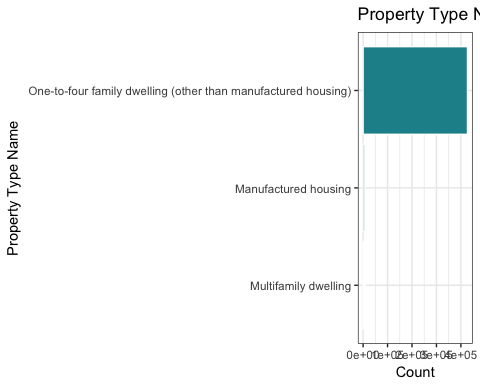
#Histogram for loan\_purpose\_name  
fillColor = "#1c9099"  
  
loan\_purpose\_name <- homeMortgage %>%  
 group\_by(loan\_purpose\_name) %>%   
 summarize(count = n()) %>%  
 arrange(desc(count))   
  
loan\_purpose\_name %>% ggplot(aes(x = reorder(loan\_purpose\_name, count), y = count)) +   
 geom\_bar(stat = 'identity',colour="white", fill =fillColor) +  
 labs(x = 'Loan Purpose Name', y = 'Count', title = 'Loan Purpose Name Distribution') +  
 coord\_flip() +   
 theme\_bw()



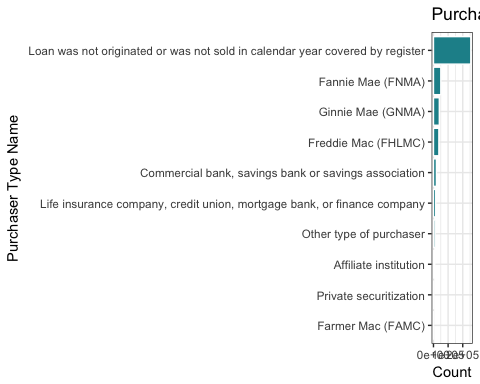
#Histogram for loan\_type\_name  
fillColor = "#1c9099"  
  
loan\_type\_name <- homeMortgage %>%  
 group\_by(loan\_type\_name) %>%   
 summarize(count = n()) %>%  
 arrange(desc(count))   
  
loan\_type\_name %>% ggplot(aes(x = reorder(loan\_type\_name, count), y = count)) +   
 geom\_bar(stat = 'identity',colour="white", fill =fillColor) +  
 labs(x = 'Loan Type Name', y = 'Count', title = 'Loan Type Name Distribution') +  
 coord\_flip() +   
 theme\_bw()



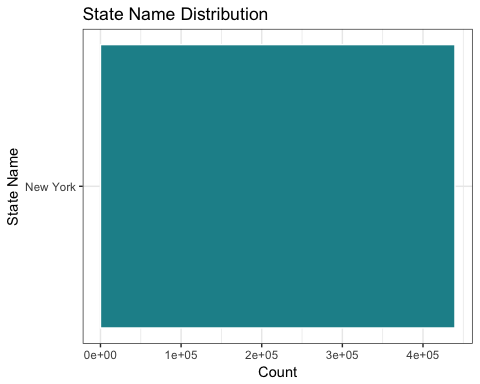
#Histogram for property\_type\_name  
fillColor = "#1c9099"  
property\_type\_name <- homeMortgage %>%  
 group\_by(property\_type\_name) %>%   
 summarize(count = n()) %>%  
 arrange(desc(count))   
  
property\_type\_name %>% ggplot(aes(x = reorder(property\_type\_name, count), y = count)) +   
 geom\_bar(stat = 'identity',colour="white", fill =fillColor) +  
 labs(x = 'Property Type Name', y = 'Count', title = 'Property Type Name Distribution') +  
 coord\_flip() +   
 theme\_bw()



#Histogram for purchaser\_type\_name  
fillColor = "#1c9099"  
purchaser\_type\_name <- homeMortgage %>%  
 group\_by(purchaser\_type\_name) %>%   
 summarize(count = n()) %>%  
 arrange(desc(count))   
  
purchaser\_type\_name %>% ggplot(aes(x = reorder(purchaser\_type\_name, count), y = count)) +   
 geom\_bar(stat = 'identity',colour="white", fill =fillColor) +  
 labs(x = 'Purchaser Type Name', y = 'Count', title = 'Purchaser Type Name Distribution') +  
 coord\_flip() +   
 theme\_bw()



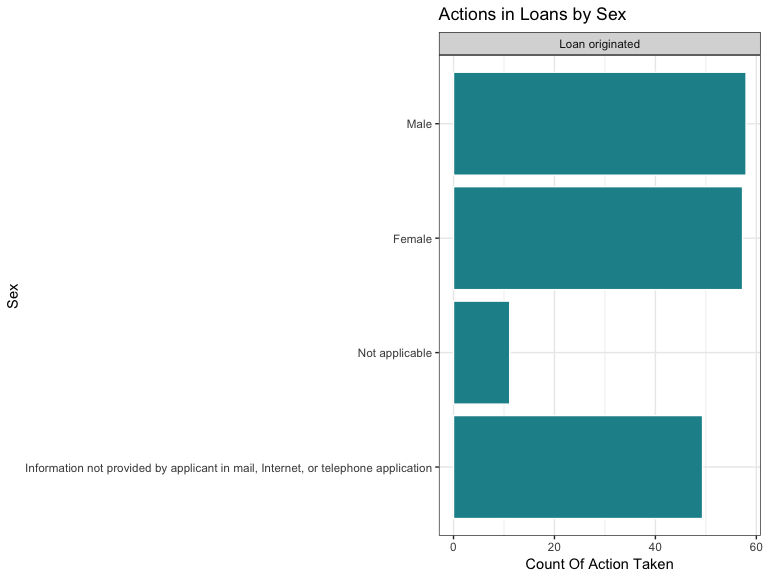
#Histogram for state\_name  
fillColor = "#1c9099"  
state\_name <- homeMortgage %>%  
 group\_by(state\_name) %>%   
 summarize(count = n()) %>%  
 arrange(desc(count))   
  
state\_name %>% ggplot(aes(x = reorder(state\_name, count), y = count)) +   
 geom\_bar(stat = 'identity',colour="white", fill =fillColor) +  
 labs(x = 'State Name', y = 'Count', title = 'State Name Distribution') +  
 coord\_flip() +   
 theme\_bw()



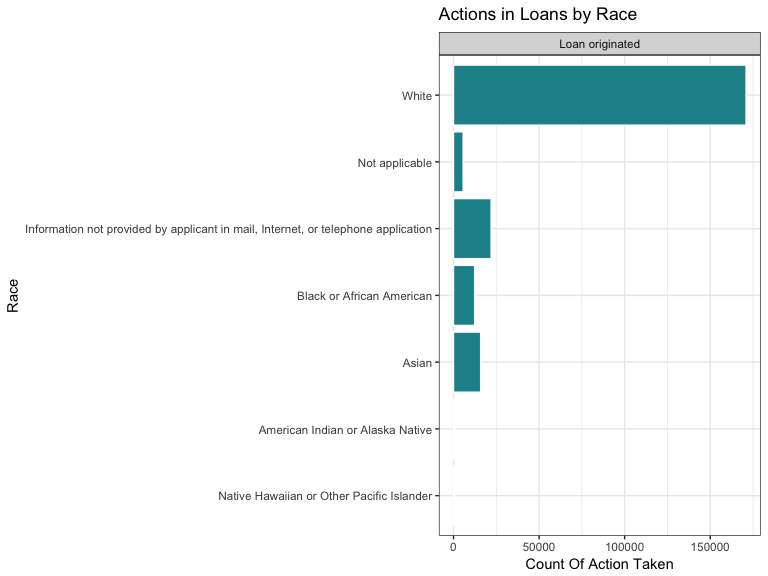
#Create a data summary table for continuous variables  
# From Part 2, we know that only "applicant\_income\_000s", "application\_date\_indicator", "as\_of\_year", "hud\_median\_family\_income", "loan\_amount\_000s", "number\_of\_1\_to\_4\_family\_units", "number\_of\_owner\_occupied\_units", "minority\_population", "population", "rate\_spread", "tract\_to\_msamd\_income" are continous variables. Therefore, there are only 11 continuous variables exist in the dataset. We will provide statistics including median, mode, mean, range, and standard deviation.  
  
homeMortgageContinuous.df <- data.frame(matrix(NA, nrow = 11, ncol = 5))  
names(homeMortgageContinuous.df) <- c("Mean", "Median", "Mode","Range", "Standard Deviation")  
rownames(homeMortgageContinuous.df) <- c("applicant\_income\_000s", "application\_date\_indicator", "as\_of\_year", "hud\_median\_family\_income", "loan\_amount\_000s", "number\_of\_1\_to\_4\_family\_units", "number\_of\_owner\_occupied\_units", "minority\_population", "population", "rate\_spread", "tract\_to\_msamd\_income")  
  
#Define the Mode function  
Mode = function(x){  
ux = sort(unique(x))  
tabx = table(x)  
maxf = ux[which(tabx == max(tabx))]  
return(maxf)}  
  
#disable scientific notion   
options(scipen = 999)  
  
# Create seperate dataset "homeMortgageContinuous"

## 5. Create bivariate frequency distributions (tables or plots) for key variables

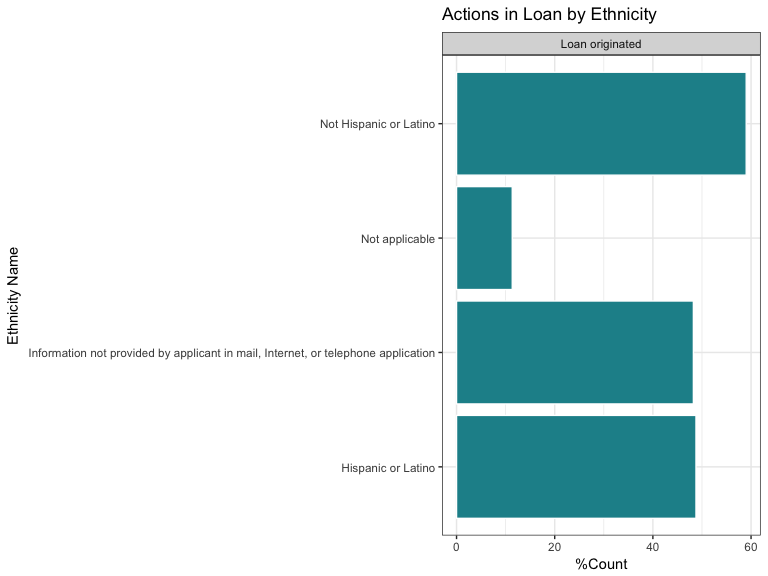
# Sex and Loan originated  
fillColor = "#1c9099"  
library(dplyr)  
applicant\_sex\_name <- homeMortgage %>%  
 filter(!is.na(applicant\_sex\_name)) %>%  
 group\_by(applicant\_sex\_name) %>%   
 summarize(CountOfSex = n()) %>%  
 arrange(desc(CountOfSex))   
  
homeMortgage\_applicant\_sex = homeMortgage %>% filter(action\_taken\_name == "Loan originated" )%>%group\_by(action\_taken\_name,applicant\_sex\_name) %>%  
 summarise(CountOfActionTaken = n()) %>%  
 arrange(desc(CountOfActionTaken))  
  
homeMortgageStatus\_sex = inner\_join(applicant\_sex\_name,homeMortgage\_applicant\_sex)%>%mutate(percentage = (CountOfActionTaken/CountOfSex) \* 100 )   
  
homeMortgageStatus\_sex%>%  
ggplot(aes(x = reorder(applicant\_sex\_name, CountOfSex),   
 y = percentage)) +  
 geom\_bar(stat='identity',colour="white", fill =fillColor) +  
 facet\_wrap(~ action\_taken\_name) +  
 labs(x = 'Sex', y = 'Count Of Action Taken', title = 'Actions in Loans by Sex') +  
 coord\_flip() +   
 theme\_bw()



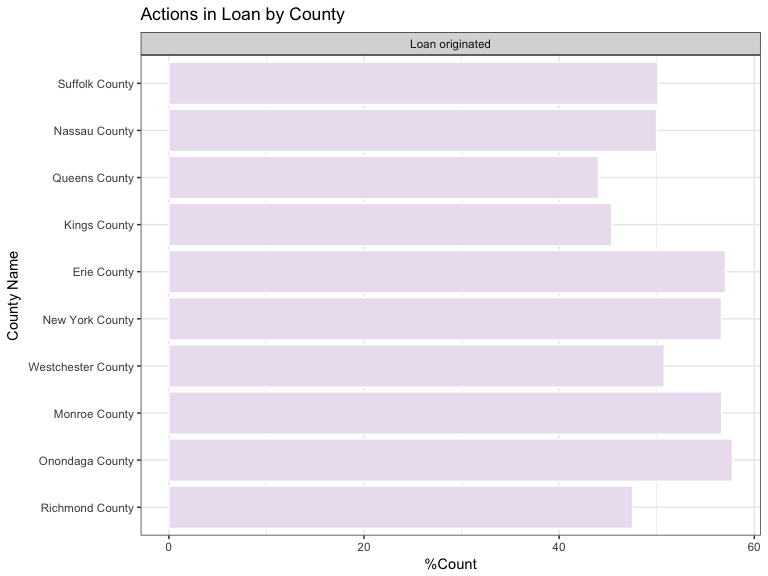
# race and Loan Originated  
fillColor = "#1c9099"  
library(dplyr)  
fillColor = "#1c9099"  
race\_name <- homeMortgage %>%  
 filter(!is.na(applicant\_race\_name\_1)) %>%  
 group\_by(applicant\_race\_name\_1) %>%   
 summarize(CountOfRace = n()) %>%  
 arrange(desc(CountOfRace))   
  
homeMortgage\_applicant\_race= homeMortgage %>% filter(action\_taken\_name == "Loan originated" )%>%group\_by(action\_taken\_name,applicant\_race\_name\_1) %>%   
 summarise(CountOfActionTaken = n()) %>%  
 arrange(desc(CountOfActionTaken))  
  
homeMortgageStatus\_race = inner\_join(race\_name,homeMortgage\_applicant\_race)   
View(homeMortgageStatus\_race)  
  
homeMortgageStatus\_race%>%  
ggplot(aes(x = reorder(applicant\_race\_name\_1, CountOfRace),   
 y = CountOfActionTaken)) +  
 geom\_bar(stat='identity',colour="white", fill =fillColor) +  
 facet\_wrap(~ action\_taken\_name) +  
 labs(x = 'Race', y = 'Count Of Action Taken', title = 'Actions in Loans by Race') +  
 coord\_flip() +   
 theme\_bw()



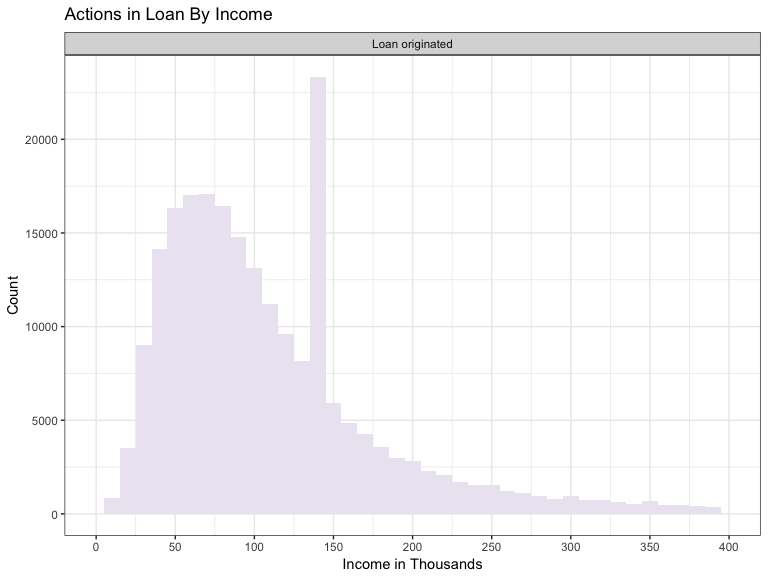
# Ethnicity and Loan Originated  
ethnicity\_name <- homeMortgage %>%  
 filter(!is.na(applicant\_ethnicity\_name)) %>%  
 group\_by(applicant\_ethnicity\_name) %>%   
 summarize(CountOfEthinity = n()) %>%  
 arrange(desc(CountOfEthinity))   
  
homeMortgage\_applicant\_ethnicity= homeMortgage %>% filter(action\_taken\_name == "Loan originated" )%>%group\_by(action\_taken\_name,applicant\_ethnicity\_name) %>%  
 summarise(CountOfActionTaken = n()) %>%  
 arrange(desc(CountOfActionTaken))  
  
homeMortgageStatus\_ethnicity = inner\_join(ethnicity\_name,homeMortgage\_applicant\_ethnicity) %>%mutate(percentage = (CountOfActionTaken/CountOfEthinity) \* 100 )   
  
homeMortgageStatus\_ethnicity %>% ggplot(aes(x = reorder(applicant\_ethnicity\_name, CountOfEthinity), y = percentage)) +   
 geom\_bar(stat = 'identity', colour="white", fill =fillColor) +  
 facet\_wrap(~ action\_taken\_name) +  
 labs(x = 'Ethnicity Name', y = '%Count', title = 'Actions in Loan by Ethnicity') +  
 coord\_flip() +   
 theme\_bw()



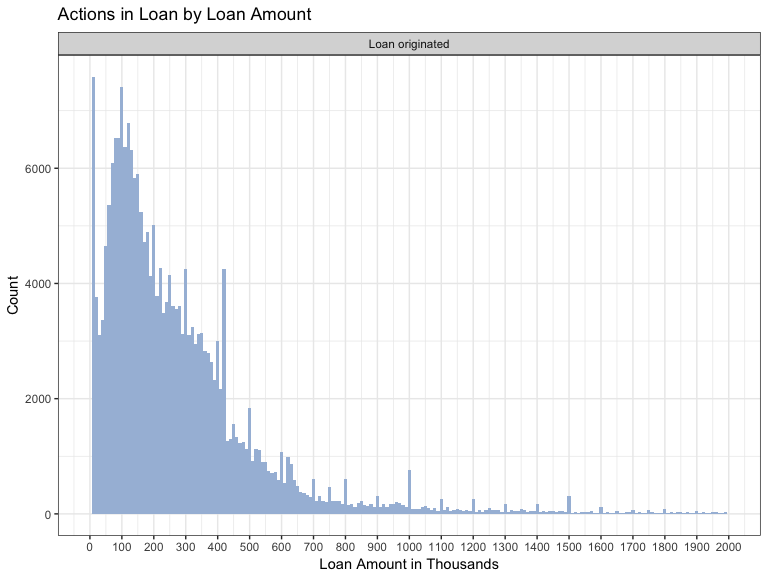
# county and Loan Originated  
fillColor = "#ece2f0"  
county\_Top10 <- homeMortgage %>%  
 filter(!is.na(county\_name)) %>%  
 group\_by(county\_name) %>%   
 summarize(CountOfCounty = n()) %>%  
 arrange(desc(CountOfCounty)) %>%  
 head(10)   
  
homeMortgage\_applicant\_county= homeMortgage %>% filter(action\_taken\_name == "Loan originated" )%>%group\_by(action\_taken\_name,county\_name) %>%  
 summarise(CountOfActionTaken = n()) %>%  
 arrange(desc(CountOfActionTaken))  
  
homeMortgageStatus\_county = inner\_join(county\_Top10,homeMortgage\_applicant\_county)%>%mutate(percentage = (CountOfActionTaken/CountOfCounty) \* 100 )   
  
 homeMortgageStatus\_county %>% ggplot(aes(x= reorder(county\_name, CountOfCounty),y=percentage)) +   
 geom\_col(colour="white", fill =fillColor) +  
 facet\_wrap(~ action\_taken\_name) +  
 labs(x = 'County Name', y = '%Count', title = 'Actions in Loan by County') +  
   
 coord\_flip() +   
 theme\_bw()



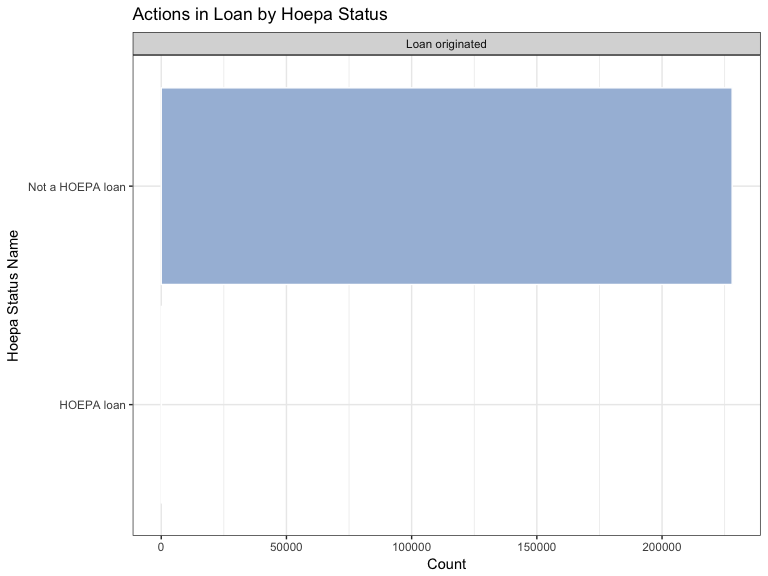
# Income and Loan Originated Distribution  
breaks = seq(0,400,50)  
fillColor = "#ece7f2"  
  
homeMortgage %>%  
 filter(action\_taken\_name == "Loan originated" )%>%group\_by(action\_taken\_name) %>%  
ggplot(aes(applicant\_income\_000s)) +  
 scale\_x\_continuous(limits = c(0, 400),breaks=breaks ) +  
 facet\_wrap(~action\_taken\_name)+  
 geom\_histogram(binwidth = 10,fill = fillColor) +  
 labs(x = 'Income in Thousands', y = 'Count', title = 'Actions in Loan By Income') + theme\_bw()



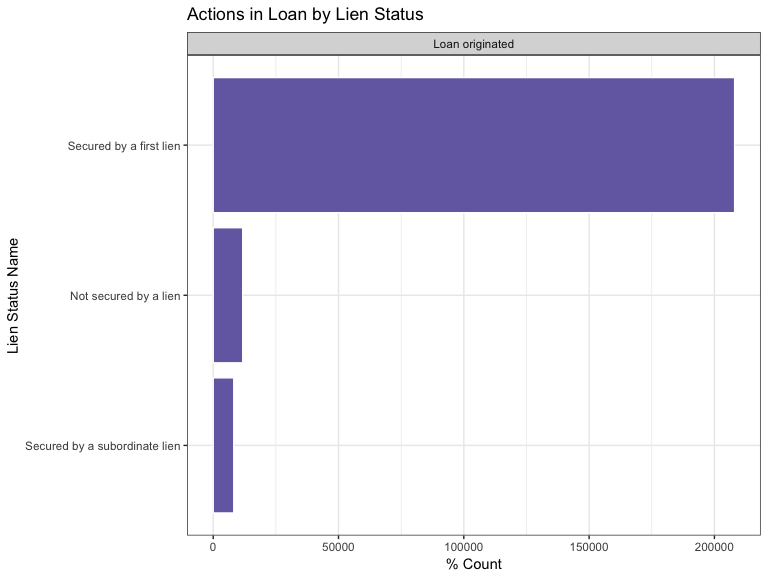
# Loan amount and Loan Originated Distribution   
breaks = seq(0,2000,100)  
fillColor = "#a6bddb"  
  
homeMortgage %>%   
 filter(action\_taken\_name == "Loan originated" )%>%group\_by(action\_taken\_name) %>%  
 ggplot(aes(loan\_amount\_000s)) +   
 scale\_x\_continuous(limits = c(0, 2000),breaks=breaks ) +  
 facet\_wrap(~action\_taken\_name)+  
 geom\_histogram(binwidth = 10,fill = fillColor) +  
 labs(x = 'Loan Amount in Thousands', y = 'Count', title = 'Actions in Loan by Loan Amount') + theme\_bw()



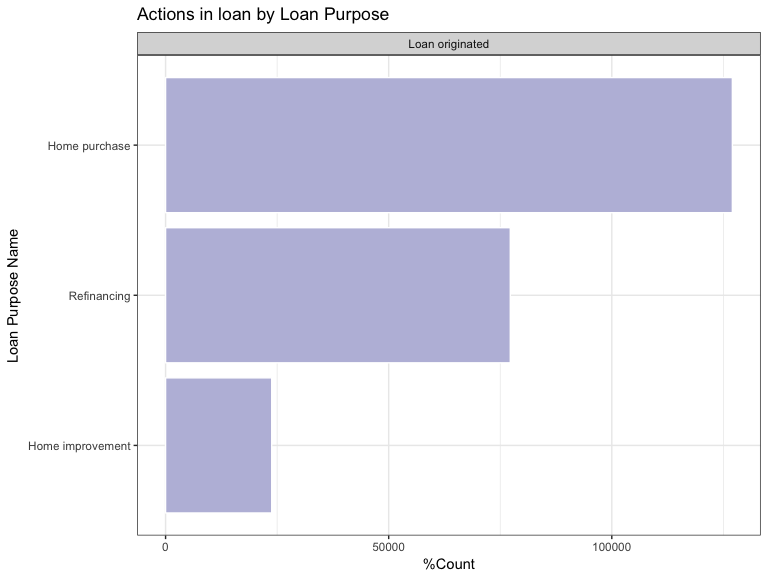
#hoepa status and Loan Originated Distribution  
fillColor = "#a6bddb"  
  
hoepa\_status\_name <- homeMortgage %>%  
 filter(!is.na(hoepa\_status\_name)) %>%  
 group\_by(hoepa\_status\_name) %>%   
 summarize(CountOfHoepa = n()) %>%  
 arrange(desc(CountOfHoepa))   
  
homeMortgage\_applicant\_hoepa= homeMortgage %>% filter(action\_taken\_name == "Loan originated" )%>%group\_by(action\_taken\_name,hoepa\_status\_name) %>%  
 summarise(CountOfActionTaken = n()) %>%  
 arrange(desc(CountOfActionTaken))  
  
homeMortgageStatus\_hoepa = inner\_join(hoepa\_status\_name,homeMortgage\_applicant\_hoepa)   
  
homeMortgageStatus\_hoepa %>% ggplot(aes(x= reorder(hoepa\_status\_name, CountOfHoepa),y=CountOfActionTaken)) +   
 geom\_col(colour="white", fill =fillColor) +  
 facet\_wrap(~ action\_taken\_name) +  
 labs(x = 'Hoepa Status Name', y = 'Count', title = 'Actions in Loan by Hoepa Status') +  
   
 coord\_flip() +   
 theme\_bw()



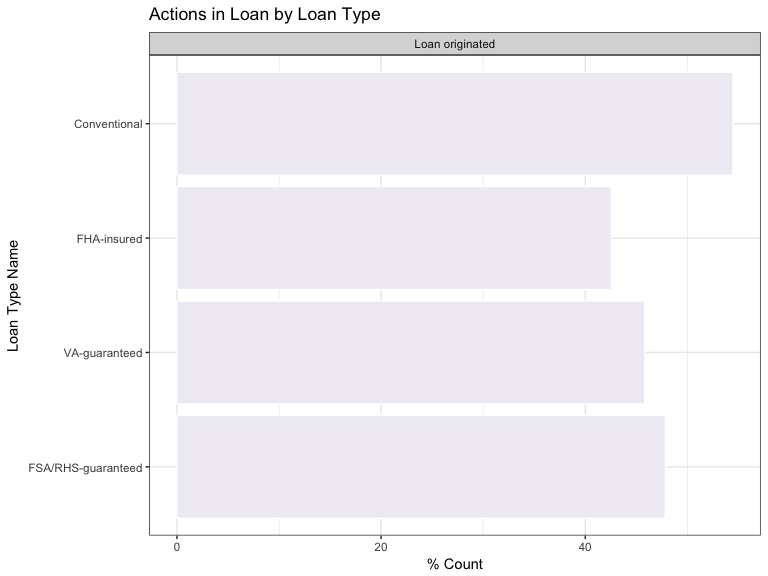
#lien status and Loan Action  
fillColor = "#756bb1"  
  
lien\_name <- homeMortgage %>%  
 filter(!is.na(lien\_status\_name)) %>%  
 group\_by(lien\_status\_name) %>%   
 summarize(CountOfLien = n()) %>%  
 arrange(desc(CountOfLien))   
  
homeMortgage\_applicant\_lien= homeMortgage %>% filter(action\_taken\_name == "Loan originated" )%>%group\_by(action\_taken\_name,lien\_status\_name) %>%  
 summarise(CountOfActionTaken = n()) %>%  
 arrange(desc(CountOfActionTaken))  
  
homeMortgageStatus\_lien = inner\_join(lien\_name,homeMortgage\_applicant\_lien)  
  
 homeMortgageStatus\_lien %>% ggplot(aes(x = reorder(lien\_status\_name, CountOfLien), y = CountOfActionTaken)) +   
 geom\_bar(stat = 'identity',colour="white", fill =fillColor) +  
 facet\_wrap(~ action\_taken\_name) +  
 labs(x = 'Lien Status Name', y = '% Count', title = 'Actions in Loan by Lien Status') +  
 coord\_flip() +   
 theme\_bw()



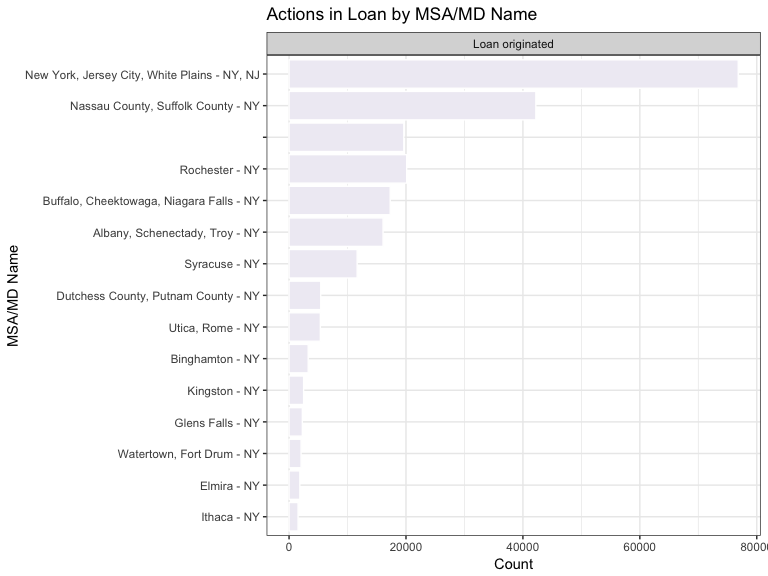
# loan purpose and mortgage action  
fillColor = "#bcbddc"  
  
loan\_purpose\_name <- homeMortgage %>%  
 filter(!is.na(loan\_purpose\_name)) %>%  
 group\_by(loan\_purpose\_name) %>%   
 summarize(count = n()) %>%  
 arrange(desc(count))   
  
homeMortgage\_applicant\_purpose= homeMortgage %>% filter(action\_taken\_name == "Loan originated" )%>%group\_by(action\_taken\_name,loan\_purpose\_name) %>%  
 summarise(CountOfActionTaken = n()) %>%  
 arrange(desc(CountOfActionTaken))  
  
homeMortgageStatus\_purpose = inner\_join(loan\_purpose\_name,homeMortgage\_applicant\_purpose)%>%mutate(percentage = (CountOfActionTaken/count) \* 100 )  
  
homeMortgageStatus\_purpose %>% ggplot(aes(x = reorder(loan\_purpose\_name, count), y = CountOfActionTaken)) +   
 geom\_bar(stat = 'identity',colour="white", fill =fillColor) +  
 facet\_wrap(~ action\_taken\_name) +  
 labs(x = 'Loan Purpose Name', y = '%Count', title = 'Actions in loan by Loan Purpose') +  
 coord\_flip() +   
 theme\_bw()



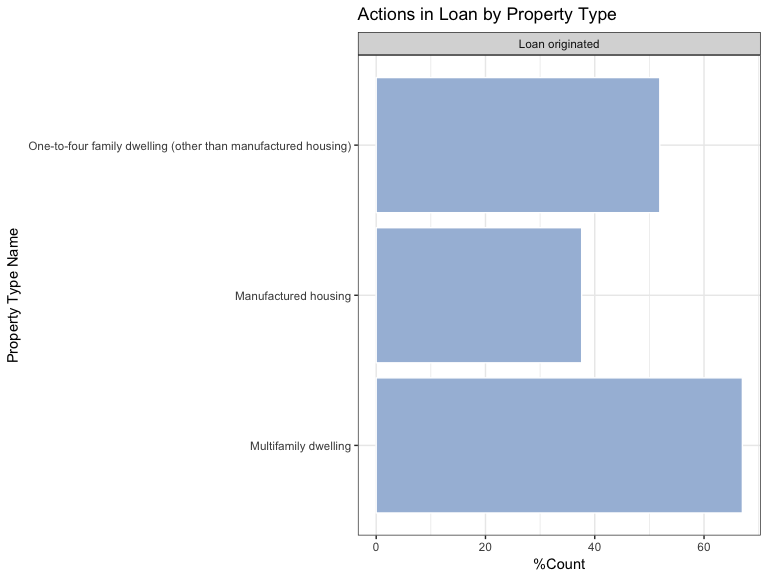
#loan type and mortgage action  
fillColor = "#efedf5"  
  
loan\_type\_name <- homeMortgage %>%  
 filter(!is.na(loan\_type\_name)) %>%  
 group\_by(loan\_type\_name) %>%   
 summarize(count = n()) %>%  
 arrange(desc(count))   
  
homeMortgage\_applicant\_type= homeMortgage %>% filter(action\_taken\_name == "Loan originated" )%>%group\_by(action\_taken\_name,loan\_type\_name) %>%  
 summarise(CountOfActionTaken = n()) %>%  
 arrange(desc(CountOfActionTaken))  
  
homeMortgageStatus\_type = inner\_join(loan\_type\_name,homeMortgage\_applicant\_type)%>%mutate(percentage = (CountOfActionTaken/count) \* 100 )  
  
homeMortgageStatus\_type%>% ggplot(aes(x = reorder(loan\_type\_name, count), y = percentage)) +   
 geom\_bar(stat = 'identity',colour="white", fill =fillColor) +  
 facet\_wrap(~ action\_taken\_name) +  
 labs(x = 'Loan Type Name', y = '% Count', title = 'Actions in Loan by Loan Type') +  
 coord\_flip() +   
 theme\_bw()



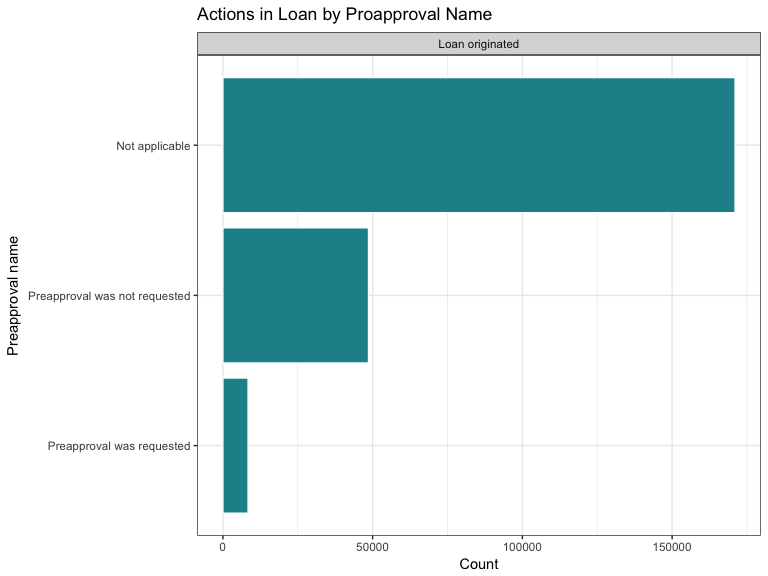
# msamd\_name and loan action  
fillColor = "#efedf5"  
  
msamd\_name <- homeMortgage %>%  
 filter(!is.na(msamd\_name)) %>%  
 group\_by(msamd\_name) %>%   
 summarize(count = n()) %>%  
 arrange(desc(count))   
  
homeMortgage\_applicant\_msamd= homeMortgage %>% filter(action\_taken\_name == "Loan originated" )%>%group\_by(msamd\_name,action\_taken\_name) %>%  
 summarise(CountOfActionTaken = n()) %>%  
 arrange(desc(CountOfActionTaken))  
  
homeMortgageStatus\_msamd = inner\_join(msamd\_name,homeMortgage\_applicant\_msamd)   
  
homeMortgageStatus\_msamd %>% ggplot(aes(x= reorder(msamd\_name, count),y=CountOfActionTaken)) +   
 geom\_bar(stat = 'identity',colour="white", fill =fillColor) +  
 facet\_wrap(~action\_taken\_name)+  
 labs(x = 'MSA/MD Name', y = 'Count', title = 'Actions in Loan by MSA/MD Name') +  
 coord\_flip() +   
 theme\_bw()



#property\_type and loan action  
fillColor = "#a6bddb"  
  
property\_type\_name <- homeMortgage %>%  
 filter(!is.na(property\_type\_name)) %>%  
 group\_by(property\_type\_name) %>%   
 summarize(count = n()) %>%  
 arrange(desc(count))   
  
homeMortgage\_applicant\_type= homeMortgage %>% filter(action\_taken\_name == "Loan originated" )%>%group\_by(property\_type\_name,action\_taken\_name) %>%  
 summarise(CountOfActionTaken = n()) %>%  
 arrange(desc(CountOfActionTaken))  
  
homeMortgageStatus\_type = inner\_join(property\_type\_name,homeMortgage\_applicant\_type)%>%mutate(percentage = (CountOfActionTaken/count) \* 100 )   
  
homeMortgageStatus\_type %>% ggplot(aes(x = reorder(property\_type\_name, count), y = percentage)) +   
 geom\_bar(stat = 'identity',colour="white", fill =fillColor) +  
 facet\_wrap(~action\_taken\_name)+  
 labs(x = 'Property Type Name', y = '%Count', title = 'Actions in Loan by Property Type') +  
 coord\_flip() +   
 theme\_bw()



#Histogram for preapproval\_name  
fillColor = "#1c9099"  
  
preapproval\_name <- homeMortgage %>%  
 filter(!is.na(preapproval\_name)) %>%  
 group\_by(preapproval\_name) %>%   
 summarize(count = n()) %>%  
 arrange(desc(count))   
  
homeMortgage\_applicant\_preapproval= homeMortgage %>% filter(action\_taken\_name == "Loan originated" )%>%group\_by(preapproval\_name,action\_taken\_name) %>%  
 summarise(CountOfActionTaken = n()) %>%  
 arrange(desc(CountOfActionTaken))  
  
homeMortgageStatus\_preapproval = inner\_join(preapproval\_name,homeMortgage\_applicant\_preapproval)   
  
homeMortgageStatus\_preapproval %>% ggplot(aes(x= reorder(preapproval\_name, count),y=CountOfActionTaken)) +   
 geom\_bar(stat = 'identity',colour="white", fill =fillColor) +  
 facet\_wrap(~action\_taken\_name)+  
 labs(x = 'Preapproval name', y = 'Count', title = 'Actions in Loan by Proapproval Name ') +  
 coord\_flip() +   
 theme\_bw()



#purchaser type and loan action  
fillColor = "#1c9099"  
  
purchaser\_type\_name <- homeMortgage %>%  
 filter(!is.na(purchaser\_type\_name)) %>%  
 group\_by(purchaser\_type\_name) %>%   
 summarize(count = n()) %>%  
 arrange(desc(count))   
  
homeMortgage\_purchaser\_type= homeMortgage %>% filter(action\_taken\_name == "Loan originated" )%>%group\_by(purchaser\_type\_name,action\_taken\_name) %>%  
 summarise(CountOfActionTaken = n()) %>%  
 arrange(desc(CountOfActionTaken))  
  
homeMortgageStatus\_purchaser = inner\_join(purchaser\_type\_name,homeMortgage\_purchaser\_type)%>%mutate(percentage = (CountOfActionTaken/count) \* 100 )   
  
homeMortgageStatus\_purchaser %>% ggplot(aes(x= reorder(purchaser\_type\_name, count),y= percentage)) +   
 geom\_bar(stat = 'identity',colour="white", fill =fillColor) +  
 facet\_wrap(~action\_taken\_name)+  
 labs(x = 'Purchaser type', y = '%Count', title = 'Actions in Loan by Purchaser Type') +  
 coord\_flip() +   
 theme\_bw()



## 6. Discuss what the data patterns indicate, and what this could mean for the probability of getting a loan approved.

homeMortgage = homeMortgage %>%  
 mutate(Loan\_Approval = ifelse(action\_taken\_name == "Loan originated", "1", "0"))  
  
#loan mean comparison  
ln = homeMortgage %>% filter(Loan\_Approval == "1")  
mean1=mean(ln$loan\_amount\_000s,na.rm=TRUE)  
print(mean1)

## [1] 352.9921

ln1 = homeMortgage %>% filter(Loan\_Approval == "0")  
mean2= mean(ln1$loan\_amount\_000s,na.rm = TRUE)  
print(mean2)

## [1] 312.1271

# t test  
t.test(ln$loan\_amount\_000s,ln1$loan\_amount\_000s)

##   
## Welch Two Sample t-test  
##   
## data: ln$loan\_amount\_000s and ln1$loan\_amount\_000s  
## t = 11.698, df = 405460, p-value < 0.00000000000000022  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 34.01836 47.71155  
## sample estimates:  
## mean of x mean of y   
## 352.9921 312.1271

# T-test type: Independent-Samples (or Unpaired Samples) t-test  
# Null hypothesis: true difference in means of loan amounts between approval or not is equal to 0  
#p < 0.05, significantly reject the null hypothesis, which shows that the mean of loan amout between approved or not approved is significantly different from each other.  
  
  
  
#High and low income threshold  
homeMortgage = homeMortgage %>%  
 mutate(Income\_Level = ifelse(applicant\_income\_000s>=68.7, "1", "0"))  
# t test  
inc = homeMortgage %>% filter(Loan\_Approval == "1",Income\_Level=="1")  
inc\_1 = homeMortgage %>% filter(Income\_Level=="1")  
inc\_1\_0 = homeMortgage %>% filter(Loan\_Approval == "0",Income\_Level=="1")  
count\_1\_1 = nrow(inc)  
count\_1 = nrow(inc\_1)  
count\_1\_0 = nrow(inc\_1\_0)  
Rate\_high = count\_1\_1/count\_1  
print(Rate\_high)

## [1] 0.5179412

inc1 = homeMortgage %>% filter(Loan\_Approval == "1",Income\_Level=="0")  
inc\_0 = homeMortgage %>% filter(Income\_Level=="0")  
inc\_0\_0 = homeMortgage %>% filter(Loan\_Approval == "0",Income\_Level=="0")  
count1\_0 = nrow(inc1)  
count\_0 = nrow(inc\_0)  
count\_0\_0 = nrow(inc\_0\_0)  
Rate\_low = count1\_0/count\_0  
print(Rate\_low)

## [1] 0.5206241

# chi-square test  
chisq.data = matrix(c(count\_1,count\_1\_0,count1\_0,count\_0\_0), nrow = 2, ncol = 2)  
rownames(chisq.data) = c("High Income","Low Income")  
colnames(chisq.data) = c("Loan Approval","Loan not Approved")  
  
chisq.test(chisq.data)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: chisq.data  
## X-squared = 10262, df = 1, p-value < 0.00000000000000022

#The chi-square test tests whether the row and column variables are independent (Null Hypothesis) or are associated (Alternate Hypothesis). Here, we are measuring, whether high and low income would influence the possibility of getting loan approved.  
#NUll hypothesis, The income level and loan approval condition are independent (not significantly influence each other)  
#p<0.05, null hypothesis is rejected, which shows that there is a signigicant difference of loan approval rate between high and low income.

From the bivariate analysis, we could find there are certain pattern in out data. a. We find that most undisclosured information (not applicable) leads to loans purchased by the institution. *Race and loan originated: White is the race which is the most likely get the loan originated also the race who apply the loan most.* Ethnicity:Not hispanic has higher possibility of getting loan originated than hispanic/latino. *County: county seems make difference on loan actions, but we are eager to believe such differenc is not caused by the geographic location itself, but the characteristics of applicants live in such counties matter more.* Income: we find out the income distribution has similar patterns in different loan actions.Most applicants of home mortgage have the income range from 25000 to 150000, and the most people has the income of 75000. *Loan amount: most applicants apply for loan amount range from 50000 to 500000, and most likely, when loan amount falls into 100 to 300 is most likely to get the application approved.* Hoepa status: whether holding a hoepa loan or not will influence the loan application result is hard to be measured from the current data. Most people who apply home mortgage dont hold a hoepa loan. About half of the non-hoepa-loan holders will get their home mortgage loan originated. *Lien status: Most people who apply for homeMortgage are secured by a first lien and they are more likely to get the loan originated.* Loan purpose: most people who apply for home mortage have the purpose of home purchase than refinancing, least people apply for home improvement. The possibility of getting loan originated is slightly higher in home purchase than refinancing and home improvement. *Loan type: loan type would impact the loan origination.Conventional loan are obviously more likely to get originated compared the rest. But, the insured applications, especially FHA-insured ansd FAS/RHS-guaranteed are more likely to get loan purchased by institution.* Property type: almost all applications are for one to four family dwelling property, and around half of the application would be originated, followed by the possibility of denied by financial institution or purchased by other institution, followed by smaller number of applications withdrawed by applicants. *Preapproval: a large number of applicants dont have the applicable information about preapproval requirement, whether preapproval is required does not significantly affect the loan origination, as it shows that most applicant will get loan originated with or without preapproval.* Purchaser type: does influence the loan origination, it has significant differences among different types of purchasers. Mostly, the loan purchaser is insurance firm or commercial bank, which we called traditional financial institute is mostly like to let the loan originated

## 7. Discuss the determinants of loan approval with respect to loan amounts, household income, applicant’s gender and race. Clearly state the dependent and independent variables used in your regression analysis. Interpret the findings of this analysis, and note the R-squared.

\*need to change the code based on previous impute data!

#define independent variables  
logit\_variables = c("action\_taken",   
"applicant\_income\_000s","applicant\_race\_1","applicant\_sex","loan\_amount\_000s")  
  
#subet the data  
homeMortgage\_logit = homeMortgage %>% select(logit\_variables) %>%  
 mutate(Loan\_Approval = FALSE) %>%  
 mutate(Loan\_Approval = replace(Loan\_Approval, action\_taken == 1, TRUE)) %>%  
 select(-action\_taken)  
  
is.factor(homeMortgage\_logit$Loan\_Approval)

## [1] FALSE

#check missing value  
sapply(homeMortgage\_logit,function(x) sum(is.na(x)))

## applicant\_income\_000s applicant\_race\_1 applicant\_sex   
## 0 0 0   
## loan\_amount\_000s Loan\_Approval   
## 0 0

#Impute the missing data by replace it with col mean  
homeMortgage\_logit$applicant\_income\_000s[is.na(homeMortgage\_logit$applicant\_income\_000s)] <- mean(homeMortgage\_logit$applicant\_income\_000s,na.rm=T)  
  
#re-check missing value  
sapply(homeMortgage\_logit,function(x) sum(is.na(x)))

## applicant\_income\_000s applicant\_race\_1 applicant\_sex   
## 0 0 0   
## loan\_amount\_000s Loan\_Approval   
## 0 0

#convert to factors  
homeMortgage\_logit$applicant\_race\_1 <- as.factor(homeMortgage\_logit$applicant\_race\_1)  
homeMortgage\_logit$applicant\_sex <- as.factor(homeMortgage\_logit$applicant\_sex)  
#homeMortgage\_logit$Loan\_Approval <- as.factor(homeMortgage\_logit$Loan\_Approval)  
  
#build the basic binomial logistic regression  
model\_1 <- glm(Loan\_Approval~., family=binomial(link='logit'), data = homeMortgage\_logit)  
  
options(scipen = 999)  
summary(model\_1)

##   
## Call:  
## glm(formula = Loan\_Approval ~ ., family = binomial(link = "logit"),   
## data = homeMortgage\_logit)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -6.498 -1.333 0.965 1.017 2.188   
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) -0.494289783 0.049652447 -9.955  
## applicant\_income\_000s -0.000074744 0.000014058 -5.317  
## applicant\_race\_12 0.680017311 0.051018458 13.329  
## applicant\_race\_13 0.147170917 0.050972563 2.887  
## applicant\_race\_14 0.305440739 0.075186576 4.062  
## applicant\_race\_15 0.853774241 0.049700915 17.178  
## applicant\_race\_16 0.233236820 0.051449974 4.533  
## applicant\_race\_17 0.399395928 0.228162175 1.750  
## applicant\_sex2 0.027013119 0.007258438 3.722  
## applicant\_sex3 0.150472518 0.018579752 8.099  
## applicant\_sex4 -2.165700402 0.223123999 -9.706  
## loan\_amount\_000s 0.000233821 0.000006311 37.049  
## Pr(>|z|)   
## (Intercept) < 0.0000000000000002 \*\*\*  
## applicant\_income\_000s 0.000000105594632116 \*\*\*  
## applicant\_race\_12 < 0.0000000000000002 \*\*\*  
## applicant\_race\_13 0.003886 \*\*   
## applicant\_race\_14 0.000048563031493556 \*\*\*  
## applicant\_race\_15 < 0.0000000000000002 \*\*\*  
## applicant\_race\_16 0.000005807644339130 \*\*\*  
## applicant\_race\_17 0.080034 .   
## applicant\_sex2 0.000198 \*\*\*  
## applicant\_sex3 0.000000000000000555 \*\*\*  
## applicant\_sex4 < 0.0000000000000002 \*\*\*  
## loan\_amount\_000s < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 608874 on 439653 degrees of freedom  
## Residual deviance: 559042 on 439642 degrees of freedom  
## AIC: 559066  
##   
## Number of Fisher Scoring iterations: 5

#produce the formula  
as.formula(  
 paste0("Loan\_Approval~", round(coefficients(model\_1)[1],3), "",   
 paste(sprintf(" %+.2f\*%s ",   
 coefficients(model\_1)[-1],   
 names(coefficients(model\_1)[-1])),   
 collapse="")  
 )  
)

## Loan\_Approval ~ -0.494 - 0 \* applicant\_income\_000s + 0.68 \* applicant\_race\_12 +   
## 0.15 \* applicant\_race\_13 + 0.31 \* applicant\_race\_14 + 0.85 \*   
## applicant\_race\_15 + 0.23 \* applicant\_race\_16 + 0.4 \* applicant\_race\_17 +   
## 0.03 \* applicant\_sex2 + 0.15 \* applicant\_sex3 - 2.17 \* applicant\_sex4 +   
## 0 \* loan\_amount\_000s

#compute Pseudo-R-squared  
nagelkerke(model\_1)

## $Models  
##   
## Model: "glm, Loan\_Approval ~ ., binomial(link = \"logit\"), homeMortgage\_logit"  
## Null: "glm, Loan\_Approval ~ 1, binomial(link = \"logit\"), homeMortgage\_logit"  
##   
## $Pseudo.R.squared.for.model.vs.null  
## Pseudo.R.squared  
## McFadden 0.0818421  
## Cox and Snell (ML) 0.1071550  
## Nagelkerke (Cragg and Uhler) 0.1429410  
##   
## $Likelihood.ratio.test  
## Df.diff LogLik.diff Chisq p.value  
## -11 -24916 49832 0  
##   
## $Number.of.observations  
##   
## Model: 439654  
## Null: 439654  
##   
## $Messages  
## [1] "Note: For models fit with REML, these statistics are based on refitting with ML"  
##   
## $Warnings  
## [1] "None"

*The dependent variablei is Loan\_Approval, and the independent variables are loan amounts, household income, appllicant’s gender and race.* The logit model is(Pseudo.R.squared=0.0818422) : Loan\_Approval ~ -0.494 - 0 \* applicant\_income\_000s + 0.68 \* applicant\_race\_12 + 0.15 \* applicant\_race\_13 + 0.31 \* applicant\_race\_14 + 0.85 \* applicant\_race\_15 + 0.23 \* applicant\_race\_16 + 0.4 \* applicant\_race\_17 + 0.03 \* applicant\_sex2 + 0.15 \* applicant\_sex3 - 2.17 \* applicant\_sex4 + 0 \* loan\_amount\_000s \*Note: applicant\_race\_11:American Indian or Alaska Native applicant\_race\_12:Asia applicant\_race\_13: Balck or African American applicant\_race\_14:Native Hawaiian or Other Pacific Islander applicant\_race\_15:White applicant\_race\_16:Information not provided by applicant in mail, Internet, or telephone application applicant\_race\_17:Not applicable applicant\_sex1:Male applicant\_sex2:Female applicant\_sex3:Information not provided by applicant in mail, Internet, or telephone application applicant\_sex4:No applicatable

\*Model interpretation: All independent variables are statistically significant(p<0.05) in our model.Given all other variables being equal, Asia applicant will increase the log odds(of being originated) by 0.68 compared with American Indian or Alaska Native applicant;applicant with race Balck or African American will increase the log odds(of being originated) by 0.15 compared with American Indian or Alaska Native applican ; applicant with race Native Hawaiian or Other Pacific Islander will increase the log odds(of being originated) by 0.16 compared with American Indian or Alaska Native, White applicant will increase the log odds(of being originated) by 0.85 compared with American Indian or Alaska Native, applicant with race information abse nce will increase the log odds(of being originated) by 0.23 compared with American Indian or Alaska Native, applicant with with race information not applicatable will increase the log odds(of being originated) by 0.4 compared with American Indian or Alaska Native applicant; while control all other variables equal, female applicant with sex will increase the log odds(of being originated) by 0.03 compared to male applicant, applicant with sex information absence will increase the log odds(of being originated) by 0.15 compared to male applicant, applicant with sex information not applicatable will decrease the log odds(of being originated) by 2.17 compared to male applicant.

## 8. You may extend this analysis by further controlling for omitted variable biases. Mention which variables you are using as controls, and re-do the regression from part (6) after adding these controls. Interpret your findings and note the change in R-squared. [Hint: For this part, you may split your data into “train” and “test” datasets, and comment on which regression model provides the best fit. Alternatively, you may use any of likelihood or Information Criterion techniques discussed in class.]

#adding controls to original regression data set: county, Loan type, Purchaser type,Lien status  
homeMortgage\_logit2 <- cbind(homeMortgage\_logit,homeMortgage[,c('loan\_type', 'purchaser\_type','lien\_status')])  
sapply(homeMortgage\_logit2,function(x) sum(is.na(x)))

## applicant\_income\_000s applicant\_race\_1 applicant\_sex   
## 0 0 0   
## loan\_amount\_000s Loan\_Approval loan\_type   
## 0 0 0   
## purchaser\_type lien\_status   
## 0 0

#A/B Test with/without controls, based on t-test results  
# Set random seed.  
set.seed(111)  
  
# Shuffle the dataset; build train and validation at 8:2  
homeMortgage\_logit2\_train\_index<- sample(row.names(homeMortgage\_logit2), 0.8\*dim(homeMortgage\_logit2)[1])  
homeMortgage\_logit2\_valid\_index<- setdiff(row.names(homeMortgage\_logit2), homeMortgage\_logit2\_train\_index)  
homeMortgage\_logit2\_train <- homeMortgage\_logit2[homeMortgage\_logit2\_train\_index, ]  
homeMortgage\_logit2\_valid <- homeMortgage\_logit2[homeMortgage\_logit2\_valid\_index, ]  
  
#covert controls to factor  
  
homeMortgage\_logit2$loan\_type <- as.factor(homeMortgage\_logit2$loan\_type)  
homeMortgage\_logit2$purchaser\_type <- as.factor(homeMortgage\_logit2$purchaser\_type)  
homeMortgage\_logit2$lien\_status <- as.factor(homeMortgage\_logit2$lien\_status)  
  
#build the basic binomial logistic regression  
model\_test <- glm(Loan\_Approval~., family=binomial(link='logit'), data = homeMortgage\_logit2\_train)  
model\_control <- glm(Loan\_Approval~., family=binomial(link='logit'), data = homeMortgage\_logit2\_train[,-c(6:8)])  
  
options(scipen = 999)  
summary(model\_test)

##   
## Call:  
## glm(formula = Loan\_Approval ~ ., family = binomial(link = "logit"),   
## data = homeMortgage\_logit2\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.4698 -0.9786 0.0377 0.7911 3.0566   
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) 1.086266550 0.066273038 16.391  
## applicant\_income\_000s 0.000053978 0.000016017 3.370  
## applicant\_race\_12 0.287142208 0.066652911 4.308  
## applicant\_race\_13 -0.092382810 0.066896030 -1.381  
## applicant\_race\_14 0.159653846 0.101055776 1.580  
## applicant\_race\_15 0.681296027 0.064988728 10.483  
## applicant\_race\_16 0.026621583 0.067192009 0.396  
## applicant\_race\_17 0.443922122 0.303398412 1.463  
## applicant\_sex2 0.057113618 0.009592709 5.954  
## applicant\_sex3 0.188537307 0.023836780 7.910  
## applicant\_sex4 -1.706782935 0.297343118 -5.740  
## loan\_amount\_000s 0.000117976 0.000006233 18.929  
## loan\_type -0.637081024 0.008508399 -74.877  
## purchaser\_type 0.860356733 0.004664340 184.454  
## lien\_status -1.001884232 0.006352989 -157.703  
## Pr(>|z|)   
## (Intercept) < 0.0000000000000002 \*\*\*  
## applicant\_income\_000s 0.000752 \*\*\*  
## applicant\_race\_12 0.00001647209696119 \*\*\*  
## applicant\_race\_13 0.167282   
## applicant\_race\_14 0.114139   
## applicant\_race\_15 < 0.0000000000000002 \*\*\*  
## applicant\_race\_16 0.691956   
## applicant\_race\_17 0.143422   
## applicant\_sex2 0.00000000261895514 \*\*\*  
## applicant\_sex3 0.00000000000000258 \*\*\*  
## applicant\_sex4 0.00000000946136975 \*\*\*  
## loan\_amount\_000s < 0.0000000000000002 \*\*\*  
## loan\_type < 0.0000000000000002 \*\*\*  
## purchaser\_type < 0.0000000000000002 \*\*\*  
## lien\_status < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 487120 on 351722 degrees of freedom  
## Residual deviance: 331820 on 351708 degrees of freedom  
## AIC: 331850  
##   
## Number of Fisher Scoring iterations: 6

summary(model\_control)

##   
## Call:  
## glm(formula = Loan\_Approval ~ ., family = binomial(link = "logit"),   
## data = homeMortgage\_logit2\_train[, -c(6:8)])  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -6.6244 -1.3322 0.9625 1.0179 2.2122   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -0.47435595 0.05523116 -8.589 < 0.0000000000000002  
## applicant\_income\_000s -0.00007508 0.00001580 -4.752 0.0000020179615732  
## applicant\_race\_12 0.65743965 0.05676003 11.583 < 0.0000000000000002  
## applicant\_race\_13 0.12786003 0.05671133 2.255 0.024160  
## applicant\_race\_14 0.25941916 0.08413397 3.083 0.002046  
## applicant\_race\_15 0.83228553 0.05528409 15.055 < 0.0000000000000002  
## applicant\_race\_16 0.19986063 0.05726953 3.490 0.000483  
## applicant\_race\_17 0.42000252 0.24616854 1.706 0.087979  
## applicant\_sex2 0.02273623 0.00811813 2.801 0.005100  
## applicant\_sex3 0.15742716 0.02080746 7.566 0.0000000000000385  
## applicant\_sex4 -2.21663280 0.24039660 -9.221 < 0.0000000000000002  
## loan\_amount\_000s 0.00024223 0.00000715 33.880 < 0.0000000000000002  
##   
## (Intercept) \*\*\*  
## applicant\_income\_000s \*\*\*  
## applicant\_race\_12 \*\*\*  
## applicant\_race\_13 \*   
## applicant\_race\_14 \*\*   
## applicant\_race\_15 \*\*\*  
## applicant\_race\_16 \*\*\*  
## applicant\_race\_17 .   
## applicant\_sex2 \*\*   
## applicant\_sex3 \*\*\*  
## applicant\_sex4 \*\*\*  
## loan\_amount\_000s \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 487120 on 351722 degrees of freedom  
## Residual deviance: 447056 on 351711 degrees of freedom  
## AIC: 447080  
##   
## Number of Fisher Scoring iterations: 5

#analyze the deviance to recheck variables significance  
anova(model\_test)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: Loan\_Approval  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev  
## NULL 351722 487120  
## applicant\_income\_000s 1 104 351721 487016  
## applicant\_race\_1 6 38051 351715 448965  
## applicant\_sex 3 131 351712 448834  
## loan\_amount\_000s 1 1779 351711 447056  
## loan\_type 1 603 351710 446453  
## purchaser\_type 1 81711 351709 364742  
## lien\_status 1 32922 351708 331820

#compute Pseudo-R-squared  
nagelkerke(model\_test)

## $Models  
##   
## Model: "glm, Loan\_Approval ~ ., binomial(link = \"logit\"), homeMortgage\_logit2\_train"  
## Null: "glm, Loan\_Approval ~ 1, binomial(link = \"logit\"), homeMortgage\_logit2\_train"  
##   
## $Pseudo.R.squared.for.model.vs.null  
## Pseudo.R.squared  
## McFadden 0.318812  
## Cox and Snell (ML) 0.356954  
## Nagelkerke (Cragg and Uhler) 0.476152  
##   
## $Likelihood.ratio.test  
## Df.diff LogLik.diff Chisq p.value  
## -14 -77650 155300 0  
##   
## $Number.of.observations  
##   
## Model: 351723  
## Null: 351723  
##   
## $Messages  
## [1] "Note: For models fit with REML, these statistics are based on refitting with ML"  
##   
## $Warnings  
## [1] "None"

nagelkerke(model\_control)

## $Models  
##   
## Model: "glm, Loan\_Approval ~ ., binomial(link = \"logit\"), homeMortgage\_logit2\_train[, -c(6:8)]"  
## Null: "glm, Loan\_Approval ~ 1, binomial(link = \"logit\"), homeMortgage\_logit2\_train[, -c(6:8)]"  
##   
## $Pseudo.R.squared.for.model.vs.null  
## Pseudo.R.squared  
## McFadden 0.0822477  
## Cox and Snell (ML) 0.1076610  
## Nagelkerke (Cragg and Uhler) 0.1436120  
##   
## $Likelihood.ratio.test  
## Df.diff LogLik.diff Chisq p.value  
## -11 -20032 40064 0  
##   
## $Number.of.observations  
##   
## Model: 351723  
## Null: 351723  
##   
## $Messages  
## [1] "Note: For models fit with REML, these statistics are based on refitting with ML"  
##   
## $Warnings  
## [1] "None"

#evaluate models performance  
predit\_result\_test <- predict(model\_test,homeMortgage\_logit2\_valid,type='response')  
predit\_result\_control<- predict(model\_control,homeMortgage\_logit2\_valid[,-c(6:8)],type='response')  
  
fitted.results\_test <- ifelse(predit\_result\_test > 0.5,1,0)  
fitted.results\_control <- ifelse(predit\_result\_control > 0.5, 1, 0)  
  
  
misClasificError\_test <- mean(fitted.results\_test != homeMortgage\_logit2\_valid$Loan\_Approval)  
print(paste('Accuracy of Test Model',1-misClasificError\_test))

## [1] "Accuracy of Test Model 0.743651272020107"

misClasificError\_control <- mean(fitted.results\_control != homeMortgage\_logit2\_valid$Loan\_Approval)  
print(paste('Accuracy of Control Model',1-misClasificError\_control))

## [1] "Accuracy of Control Model 0.623955146649077"

\*Control variables are selected based on previous bivariate anlysis.The control model includes predictors: loan amounts, household income, applicant’s gender and race; while the test model further added county, loan type, purchaser type,lien status to the control model. For control model, all predictors are statistically significant(p<0.05) except applicant\_race\_17; while for the test model, all predictors are highly statistically significant(p<0.001) except applicant\_race\_13, applicant\_race\_14,applicant\_race\_16 and applicant\_race\_17. To evaluate the fittness of those two models,we used Psdudo R square and Information ratio. We found that the Pseudo R squared of our test model(0.318812) is much higher than our control model(0.0822477), , and the AIC of the test model is: 331850, which is also much lower than that of the control model:447080. The Psdudo R square and Information ratio proves test model has better fittness than control model. To test the predictability of two models, we splited the dataset to training and validation sets at 8:2, and trained and validated the models seperately. The result is the test model has higher accuracy rate(0.744) than the control model(0.624).

## 9. In your final selected model, are there any significant customer segments based on demographics? Explain.

as.formula(  
 paste0("Loan\_Approval~", round(coefficients(model\_test)[1],2), "",   
 paste(sprintf(" %+.2f\*%s ",   
 coefficients(model\_test)[-1],   
 names(coefficients(model\_test)[-1])),   
 collapse="")  
 )  
)

## Loan\_Approval ~ 1.09 + 0 \* applicant\_income\_000s + 0.29 \* applicant\_race\_12 -   
## 0.09 \* applicant\_race\_13 + 0.16 \* applicant\_race\_14 + 0.68 \*   
## applicant\_race\_15 + 0.03 \* applicant\_race\_16 + 0.44 \* applicant\_race\_17 +   
## 0.06 \* applicant\_sex2 + 0.19 \* applicant\_sex3 - 1.71 \* applicant\_sex4 +   
## 0 \* loan\_amount\_000s - 0.64 \* loan\_type + 0.86 \* purchaser\_type -   
## 1 \* lien\_status

\*From our final selected model, we have applicant income, applicant race, applicant sex are all demographic factors. Accroding to statistical significace of those variables, we found different segments will have different proabilityof loan originated. For example, applicant with race White or Asia are both statistically significant(p<0.001) in our model, and they all have higher log odds(of being originated) than the American Indian or Alaska Native segments when control other variables equal. Also, the three sex categorical variables are all statistically significant (p<0.001) in our model with different log odds(of being originated) compared with that of male when control other variables equal, which means sex could be also considered as segmentation factors.

## 10. Summarize your findings, and discuss what this could mean for the probability of getting a loan approved.

*When control all other variables being equal, White applicant and Asia applicant have higher prob to get loan approved than American Indian or Alaska Native;* Being a female has higher prob of getting a loan aproved compared with male, when holding all other variables equal; *Coventional loan is more likely to get approved compared to all other types of loan, which holding all other variables equal;* Purchaser type belongs to Fannie Mae (FNMA) has higher prob of getting a loan aproved compared with other purchaser types, when holding all other variables equal; \*Lien status is ‘Secured by a first lien’ is more get approved compared to all other lien status, which holding all other variables equal;

## 11. Reference

## Appendix