HMDA

Shiqi Lin, Rita Li, Yixuan Wang, Kankan Yang

2018/2/24

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("~/Desktop/Marketing Analytics")  
NYHMDA = read.csv("ny\_hmda\_2015.csv")

1.a. What is the data problem

First of all, descriptions of each variable is missing, which may cause some misunderstanding for users. Second of all, some variables are redundant, for instance, the dataset has both codes and names for the race/gender/ethnicity variables, which is unnecessary and requires huge amount of cleaning. Third, there are many missing data in some key variables, such as for applicant???s income data, out of roughly 60000 of 439654 data are missing. Forth, the data contains some outliers in terms of annual income and loan amount. Finally, we believe the dataset should specify the sampling method.

1.b. Managerial objective

The managerial objective is to specify the key determinants of obtaining 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.

1. Measurement type of each variable

Character: 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

Nominal: 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: 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

1. Table summarizing range/variation of each variable

###Action taken  
table(NYHMDA$action\_taken\_name)

##   
## Application approved but not accepted   
## 14180   
## Application denied by financial institution   
## 79697   
## Application withdrawn by applicant   
## 39496   
## File closed for incompleteness   
## 16733   
## Loan originated   
## 228054   
## Loan purchased by the institution   
## 61490   
## Preapproval request denied by financial institution   
## 4

###Agency name  
table(NYHMDA$agency\_abbr)

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

###Applicant ethnicity name  
table(NYHMDA$applicant\_ethnicity\_name)

##   
## Hispanic or Latino   
## 25073   
## Information not provided by applicant in mail, Internet, or telephone application   
## 43358   
## Not Hispanic or Latino   
## 320515   
## Not applicable   
## 50708

###Applicant income  
summary(NYHMDA$applicant\_income\_000s)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 1.0 58.0 90.0 140.1 142.0 9999.0 61003

###Applicant race  
table(NYHMDA$applicant\_race\_name\_1)

##   
## American Indian or Alaska Native   
## 1712   
## Asian   
## 27890   
## Black or African American   
## 28687   
## Information not provided by applicant in mail, Internet, or telephone application   
## 46279   
## Native Hawaiian or Other Pacific Islander   
## 1260   
## Not applicable   
## 50491   
## White   
## 283335

###Applicant gender  
table(NYHMDA$applicant\_sex\_name)

##   
## Female   
## 119877   
## Information not provided by applicant in mail, Internet, or telephone application   
## 27163   
## Male   
## 242175   
## Not applicable   
## 50439

###County name  
table(NYHMDA$county\_name)

##   
## Albany County Allegany County   
## 1192 9450 1054   
## Bronx County Broome County Cattaraugus County   
## 11177 4495 1780   
## Cayuga County Chautauqua County Chemung County   
## 2236 2763 2731   
## Chenango County Clinton County Columbia County   
## 1103 2131 1877   
## Cortland County Delaware County Dutchess County   
## 1186 903 7801   
## Erie County Essex County Franklin County   
## 25109 1197 1113   
## Fulton County Genesee County Greene County   
## 1256 1532 1633   
## Hamilton County Herkimer County Jefferson County   
## 224 1913 3464   
## Kings County Lewis County Livingston County   
## 33540 923 1903   
## Madison County Monroe County Montgomery County   
## 2187 23536 1082   
## Nassau County New York County Niagara County   
## 38797 25046 5496   
## Oneida County Onondaga County Ontario County   
## 6545 14277 4498   
## Orange County Orleans County Oswego County   
## 9918 1141 3607   
## Otsego County Putnam County Queens County   
## 1609 2959 37866   
## Rensselaer County Richmond County Rockland County   
## 4979 12153 7597   
## Saratoga County Schenectady County Schoharie County   
## 9129 4931 845   
## Schuyler County Seneca County St. Lawrence County   
## 657 979 1832   
## Steuben County Suffolk County Sullivan County   
## 3196 45525 1502   
## Tioga County Tompkins County Ulster County   
## 1367 2109 4614   
## Warren County Washington County Wayne County   
## 2515 1791 3339   
## Westchester County Wyoming County Yates County   
## 24371 1084 889

###HOEPA status  
table(NYHMDA$hoepa\_status\_name)

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

###Lien status  
table(NYHMDA$lien\_status\_name)

##   
## Not applicable Not secured by a lien   
## 61490 23620   
## Secured by a first lien Secured by a subordinate lien   
## 340272 14272

###Loan purpose  
table(NYHMDA$loan\_purpose\_name)

##   
## Home improvement Home purchase Refinancing   
## 46083 219174 174397

###Loan type  
table(NYHMDA$loan\_type\_name)

##   
## Conventional FHA-insured FSA/RHS-guaranteed   
## 337707 81263 3779   
## VA-guaranteed   
## 16905

###Owner occupancy  
table(NYHMDA$owner\_occupancy\_name)

##   
## Not applicable   
## 6041   
## Not owner-occupied as a principal dwelling   
## 42106   
## Owner-occupied as a principal dwelling   
## 391507

###Property type  
table(NYHMDA$property\_type\_name)

##   
## Manufactured housing   
## 7073   
## Multifamily dwelling   
## 5670   
## One-to-four family dwelling (other than manufactured housing)   
## 426911

###Purchaser type  
table(NYHMDA$purchaser\_type\_name)

##   
## Affiliate institution   
## 8806   
## Commercial bank, savings bank or savings association   
## 20063   
## Fannie Mae (FNMA)   
## 50247   
## Farmer Mac (FAMC)   
## 5   
## Freddie Mac (FHLMC)   
## 36979   
## Ginnie Mae (GNMA)   
## 39958   
## Life insurance company, credit union, mortgage bank, or finance company   
## 15427   
## Loan was not originated or was not sold in calendar year covered by register   
## 255196   
## Other type of purchaser   
## 11183   
## Private securitization   
## 1790

###Family income  
summary(NYHMDA$hud\_median\_family\_income)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 57200 69000 71300 78220 82700 109000 1667

###Loan amount  
summary(NYHMDA$loan\_amount\_000s)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 102.0 208.0 333.3 366.0 100000.0

1. Missing data
2. Identify key variables Actions in loans, race, ethnicity, gender, income, loan purpose, loan type, lien status, loan amount, property type
3. Change inputs that are not applicable to null

NYHMDA$Hispanic="0"  
NYHMDA[NYHMDA$applicant\_ethnicity==3,]$Hispanic=NA  
NYHMDA[NYHMDA$applicant\_ethnicity==4,]$Hispanic=NA  
NYHMDA$Male="0"  
NYHMDA[NYHMDA$applicant\_sex==3,]$Male=NA  
NYHMDA[NYHMDA$applicant\_sex==4,]$Male=NA  
NYHMDA$Race="0"  
NYHMDA[NYHMDA$applicant\_race\_1==6,]$Race=NA  
NYHMDA[NYHMDA$applicant\_race\_1==7,]$Race=NA  
NYHMDA$First\_lien="0"  
NYHMDA[NYHMDA$lien\_status==4,]$First\_lien=NA  
NYHMDA$Principal\_dwelling="0"  
NYHMDA[NYHMDA$owner\_occupancy==3,]$Principal\_dwelling=NA

1. Delete columns with unnecessary variables

NYHMDA\_clear<-subset(NYHMDA,select=-c(10,11,12,13,15,16,17,18,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,40,41,42,43,44,45,46,47,68,69,70,77))

1. Delete rows with missing data

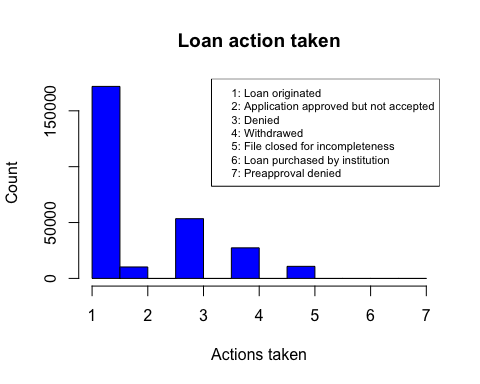
NYHMDA\_new<-na.omit(NYHMDA\_clear)  
summary(NYHMDA\_new)

## action\_taken   
## Min. :1.000   
## 1st Qu.:1.000   
## Median :1.000   
## Mean :1.886   
## 3rd Qu.:3.000   
## Max. :7.000   
##   
## action\_taken\_name   
## Application approved but not accepted : 10237   
## Application denied by financial institution : 53430   
## Application withdrawn by applicant : 27345   
## File closed for incompleteness : 10795   
## Loan originated :171854   
## Loan purchased by the institution : 0   
## Preapproval request denied by financial institution: 4   
## agency\_code agency\_abbr   
## Min. :1.000 CFPB:101372   
## 1st Qu.:5.000 FDIC: 10309   
## Median :7.000 FRS : 6577   
## Mean :6.738 HUD : 95533   
## 3rd Qu.:9.000 NCUA: 39758   
## Max. :9.000 OCC : 20116   
##   
## agency\_name applicant\_ethnicity  
## Consumer Financial Protection Bureau :101372 Min. :1.000   
## Department of Housing and Urban Development: 95533 1st Qu.:2.000   
## Federal Deposit Insurance Corporation : 10309 Median :2.000   
## Federal Reserve System : 6577 Mean :1.932   
## National Credit Union Administration : 39758 3rd Qu.:2.000   
## Office of the Comptroller of the Currency : 20116 Max. :2.000   
##   
## applicant\_ethnicity\_name  
## Hispanic or Latino : 18676   
## Information not provided by applicant in mail, Internet, or telephone application: 0   
## Not Hispanic or Latino :254989   
## Not applicable : 0   
##   
##   
##   
## applicant\_income\_000s applicant\_race\_1  
## Min. : 1.0 Min. :1.000   
## 1st Qu.: 60.0 1st Qu.:5.000   
## Median : 92.0 Median :5.000   
## Mean : 141.5 Mean :4.542   
## 3rd Qu.: 145.0 3rd Qu.:5.000   
## Max. :9999.0 Max. :5.000   
##   
## applicant\_race\_name\_1  
## American Indian or Alaska Native : 1247   
## Asian : 24475   
## Black or African American : 22996   
## Information not provided by applicant in mail, Internet, or telephone application: 0   
## Native Hawaiian or Other Pacific Islander : 1042   
## Not applicable : 0   
## White :223905   
## applicant\_sex   
## Min. :1.00   
## 1st Qu.:1.00   
## Median :1.00   
## Mean :1.33   
## 3rd Qu.:2.00   
## Max. :2.00   
##   
## applicant\_sex\_name  
## Female : 90295   
## Information not provided by applicant in mail, Internet, or telephone application: 0   
## Male :183370   
## Not applicable : 0   
##   
##   
##   
## county\_code county\_name hoepa\_status  
## Min. : 1.00 Suffolk County: 30646 Min. :1   
## 1st Qu.: 53.00 Nassau County : 26312 1st Qu.:2   
## Median : 65.00 Queens County : 23728 Median :2   
## Mean : 68.53 Kings County : 20387 Mean :2   
## 3rd Qu.: 91.00 Erie County : 17908 3rd Qu.:2   
## Max. :123.00 Monroe County : 16384 Max. :2   
## (Other) :138300   
## hoepa\_status\_name lien\_status   
## HOEPA loan : 54 Min. :1.000   
## Not a HOEPA loan:273611 1st Qu.:1.000   
## Median :1.000   
## Mean :1.145   
## 3rd Qu.:1.000   
## Max. :3.000   
##   
## lien\_status\_name loan\_purpose   
## Not applicable : 0 Min. :1.000   
## Not secured by a lien : 14585 1st Qu.:1.000   
## Secured by a first lien :248563 Median :1.000   
## Secured by a subordinate lien: 10517 Mean :1.887   
## 3rd Qu.:3.000   
## Max. :3.000   
##   
## loan\_purpose\_name loan\_type loan\_type\_name   
## Home improvement: 29732 Min. :1.000 Conventional :224394   
## Home purchase :137405 1st Qu.:1.000 FHA-insured : 38508   
## Refinancing :106528 Median :1.000 FSA/RHS-guaranteed: 1519   
## Mean :1.225 VA-guaranteed : 9244   
## 3rd Qu.:1.000   
## Max. :4.000   
##   
## msamd msamd\_name   
## Min. :10580 New York, Jersey City, White Plains - NY, NJ:101743   
## 1st Qu.:35004 Nassau County, Suffolk County - NY : 56958   
## Median :35614 Rochester - NY : 25449   
## Mean :32147 Buffalo, Cheektowaga, Niagara Falls - NY : 21831   
## 3rd Qu.:35614 Albany, Schenectady, Troy - NY : 21044   
## Max. :48060 Syracuse - NY : 14811   
## (Other) : 31829   
## owner\_occupancy owner\_occupancy\_name  
## Min. :1.000 Not applicable : 0   
## 1st Qu.:1.000 Not owner-occupied as a principal dwelling: 23152   
## Median :1.000 Owner-occupied as a principal dwelling :250513   
## Mean :1.085   
## 3rd Qu.:1.000   
## Max. :2.000   
##   
## preapproval preapproval\_name property\_type   
## Min. :1.000 Not applicable :213975 Min. :1.000   
## 1st Qu.:3.000 Preapproval was not requested: 51970 1st Qu.:1.000   
## Median :3.000 Preapproval was requested : 7720 Median :1.000   
## Mean :2.754 Mean :1.011   
## 3rd Qu.:3.000 3rd Qu.:1.000   
## Max. :3.000 Max. :2.000   
##   
## property\_type\_name  
## Manufactured housing : 3042   
## Multifamily dwelling : 0   
## One-to-four family dwelling (other than manufactured housing):270623   
##   
##   
##   
##   
## purchaser\_type   
## Min. :0.000   
## 1st Qu.:0.000   
## Median :0.000   
## Mean :1.581   
## 3rd Qu.:2.000   
## Max. :9.000   
##   
## purchaser\_type\_name  
## Loan was not originated or was not sold in calendar year covered by register:164755   
## Fannie Mae (FNMA) : 30010   
## Freddie Mac (FHLMC) : 20887   
## Commercial bank, savings bank or savings association : 16554   
## Ginnie Mae (GNMA) : 14076   
## Life insurance company, credit union, mortgage bank, or finance company : 12249   
## (Other) : 15134   
## respondent\_id sequence\_number hud\_median\_family\_income  
## 0000451965: 21579 Min. : 1 Min. : 57200   
## 0000852218: 13434 1st Qu.: 966 1st Qu.: 69700   
## 7197000003: 10921 Median : 7181 Median : 71300   
## 0000476810: 8775 Mean : 85140 Mean : 79719   
## 26-4599244: 7488 3rd Qu.: 59534 3rd Qu.: 82700   
## 0000501105: 6565 Max. :1206621 Max. :109000   
## (Other) :204903   
## loan\_amount\_000s number\_of\_1\_to\_4\_family\_units  
## Min. : 1.0 Min. : 6   
## 1st Qu.: 106.0 1st Qu.: 965   
## Median : 215.0 Median :1512   
## Mean : 294.6 Mean :1511   
## 3rd Qu.: 370.0 3rd Qu.:2007   
## Max. :24000.0 Max. :6345   
##   
## number\_of\_owner\_occupied\_units minority\_population population   
## Min. : 3 Min. : 0.34 Min. : 82   
## 1st Qu.: 775 1st Qu.: 8.37 1st Qu.: 3493   
## Median :1223 Median :17.39 Median : 4640   
## Mean :1242 Mean :28.67 Mean : 4809   
## 3rd Qu.:1669 3rd Qu.:37.27 3rd Qu.: 5966   
## Max. :6454 Max. :99.93 Max. :26588   
##   
## tract\_to\_msamd\_income Hispanic Male   
## Min. : 8.31 Length:273665 Length:273665   
## 1st Qu.: 89.05 Class :character Class :character   
## Median :108.55 Mode :character Mode :character   
## Mean :119.90   
## 3rd Qu.:134.51   
## Max. :367.61   
##   
## Race First\_lien Principal\_dwelling  
## Length:273665 Length:273665 Length:273665   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##

1. Analysis for key variables(histogram/density plot)

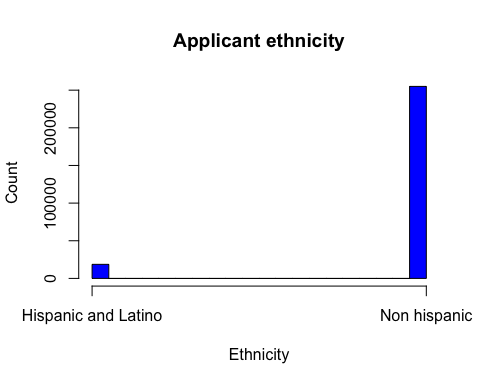
Variables deem necessary: Actions in loans, ethnicity, race, income, loan purpose, gender, lien status, loan amount 5.a Actions in loans

hist(NYHMDA\_new$action\_taken,col="blue",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=0.7)

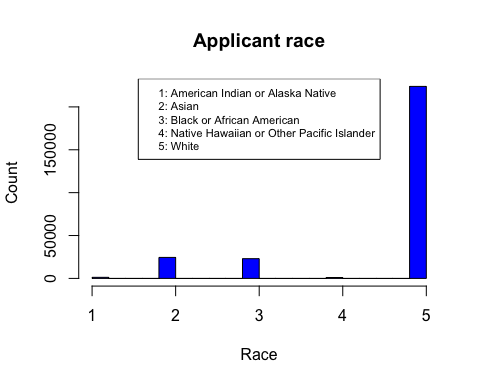
 Among all the samples, most of loans are originated.

5.b Ethnicity

hist(NYHMDA\_new$applicant\_ethnicity,col="blue",main="Applicant ethnicity",ylab="Count",xlab="Ethnicity",xaxt="n")  
axis(side=1,at=c(1,2),labels=c("Hispanic and Latino","Non hispanic"))

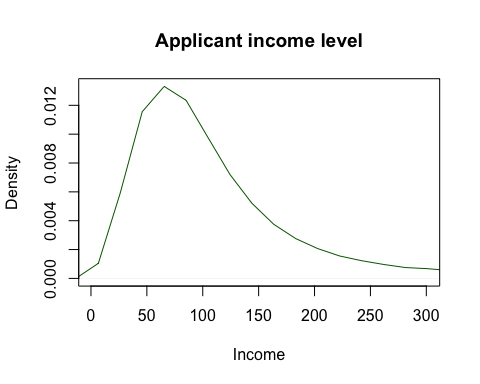
 Among all loan applicants, most of them are non hispanic or latino applicants. 5.c Race

hist(NYHMDA\_new$applicant\_race\_1,col="blue",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=0.7)

 Among all loan applicants, most of them are white people.

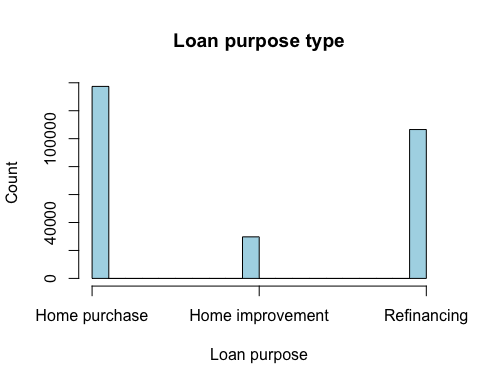
5.d Income

plot(density(NYHMDA\_new$applicant\_income\_000s),xlab="Income",ylab="Density",main="Applicant income level",xlim=c(1,300),col="darkgreen")

 Among all applicants, most people have annual income of around 60000 dollars.

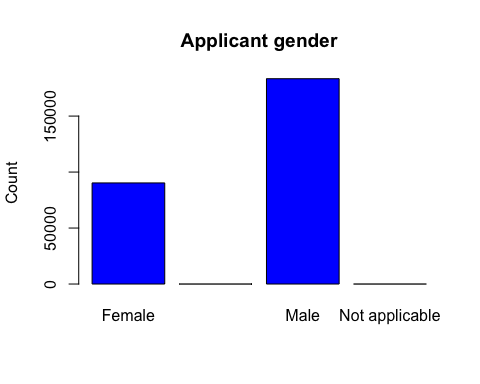
5.e Loan purpose

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"))

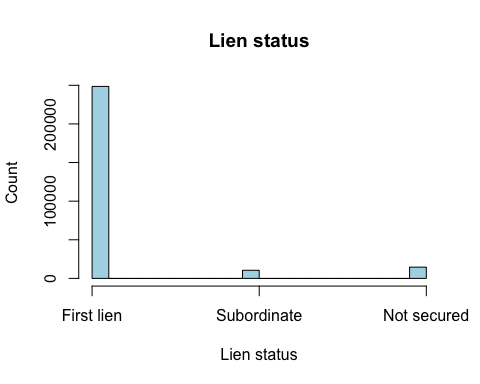
 In terms of loan purpose type, home purchase and refinancing are the most prevalent purposes.

5.f Gender

plot(NYHMDA\_new$applicant\_sex\_name,col="blue",main="Applicant gender",ylab="Count")

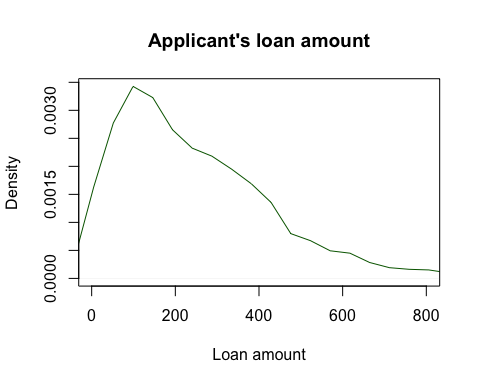
 The number of male applicants is almost twice the number of female applicants. 5.g Lien status

hist(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"))

 Most financial institutions are first lien debt holders of the mortgages.

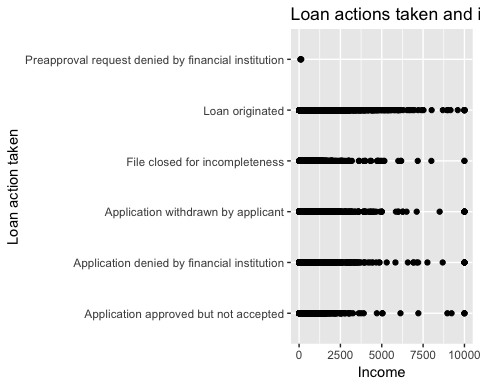
5.h Loan amount

plot(density(NYHMDA\_new$loan\_amount\_000s),xlab="Loan amount",ylab="Density",main="Applicant's loan amount",xlim=c(1,800),col="darkgreen")

 Most loan amount is clustered around 100,000 dollars and average amount is about 234,900 dollars.

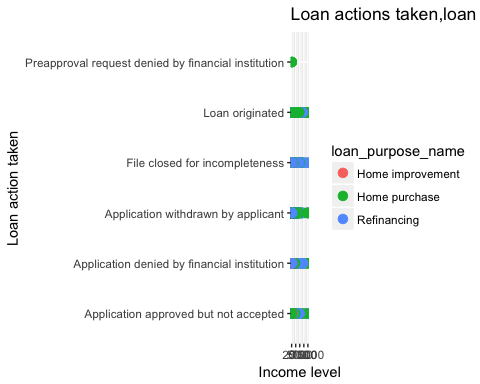
1. Bivariate frequency distributions 6.a Income level to loan action taken

library(ggplot2)  
ggplot(NYHMDA\_new,aes(x=NYHMDA\_new$applicant\_income\_000s,y=NYHMDA\_new$action\_taken\_name))+  
 geom\_point()+  
 labs(x="Income",y="Loan action taken",title="Loan actions taken and income level")



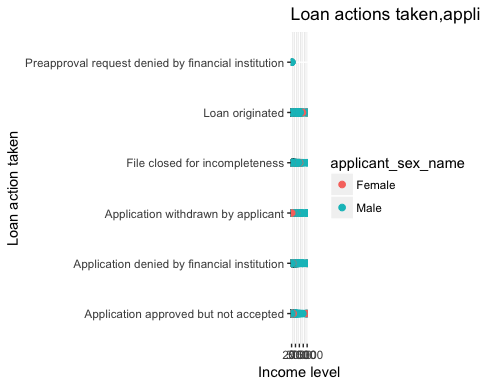
6.b Loan purpose, income level and loan actions taken

NYHMDA\_new$loan\_purpose\_name <- factor(NYHMDA\_new$loan\_purpose\_name)  
ggplot(data=NYHMDA\_new,aes(x=NYHMDA\_new$applicant\_income\_000s,y=NYHMDA\_new$action\_taken\_name,color=loan\_purpose\_name))+  
 geom\_point(size=3)+  
 labs(x="Income level",y="Loan action taken",title="Loan actions taken,loan purpose and income level")



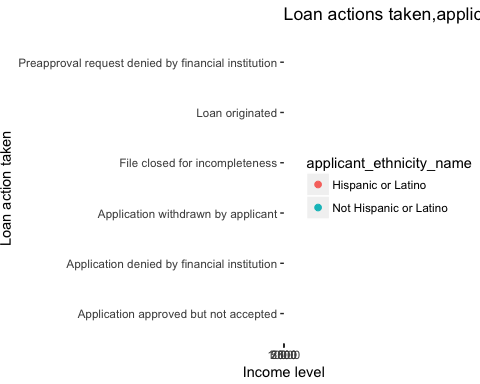
6.c Gender, income level and loan actions taken

NYHMDA\_new$applicant\_sex\_name <- factor(NYHMDA\_new$applicant\_sex\_name)  
ggplot(data=NYHMDA\_new,aes(x=NYHMDA\_new$applicant\_income\_000s,y=NYHMDA\_new$action\_taken\_name,color=applicant\_sex\_name))+  
 geom\_point(size=2)+  
 labs(x="Income level",y="Loan action taken",title="Loan actions taken,applicant gender and income level")



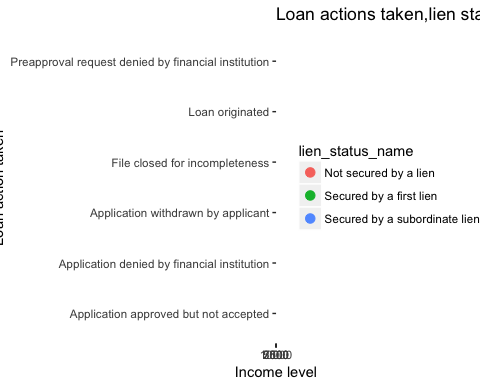
6.d Ethnicity, income level and loan actions taken

NYHMDA\_new$applicant\_ethnicity\_name <- factor(NYHMDA\_new$applicant\_ethnicity\_name)  
ggplot(data=NYHMDA\_new,aes(x=NYHMDA\_new$applicant\_income\_000s,y=NYHMDA\_new$action\_taken\_name,color=applicant\_ethnicity\_name))+  
 geom\_point(size=2)+  
 labs(x="Income level",y="Loan action taken",title="Loan actions taken,applicant ethnicity and income level")



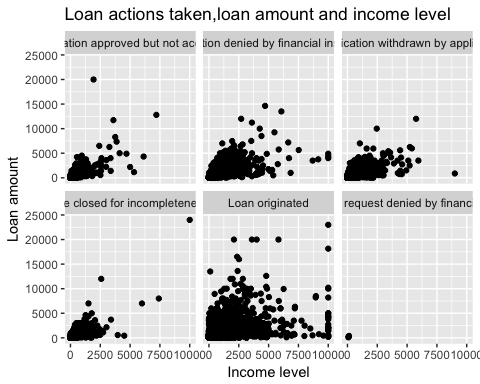
6.e Lien status, income level and loan actions taken

NYHMDA\_new$lien\_status\_name <- factor(NYHMDA\_new$lien\_status\_name)  
ggplot(data=NYHMDA\_new,aes(x=NYHMDA\_new$applicant\_income\_000s,y=NYHMDA\_new$action\_taken\_name,color=lien\_status\_name))+  
 geom\_point(size=3)+  
 labs(x="Income level",y="Loan action taken",title="Loan actions taken,lien status and income level")



6.f Loan amount and income level

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\_name)+  
 labs(x="Income level",y="Loan amount",title="Loan actions taken,loan amount and income level")



1. Data pattern indication

In terms of loan actions taken, we expects 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.

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. In terms of applicant gender and ethnicity, male applicants and non-Hispanic or Latino are more likely to get the loan in the sample. However, this conclusion is not very persuasive since there are much more male and non-Hispanic or Latino in the sample dataset. Lack of female applicants can be the fact that women are less likely to secure a loan or they have relatively lower income thus less likely to secure a loan.

For applicants whose loans are originated, none of them are not secured by a lien. However, for applicants whose loans are denied, some of them are not secured by a lien.

With regard to the relationship between loan amount and applicant income level, applicants who have higher income tend to apply for greater amount of loan, and these applications are more likely to be approved.

In conclusion, the more income and senior lien an applicant has, the more likely she/he will get the loan from financial institutions.