nyc\_housing

Anubhav Maini

February 21, 2018

Harshil Dwivedi, Monica Katoch, Anubhav Maini & Eric Moyal Marketing Analytics Professor Ranjan February 21, 2018, Git Link- <https://github.com/anumaini/housing_data>

New York Mortgage Decisions

The data problem is what variables are necessary to predict mortgage decisions based on the variables provided in “ny\_hmda\_2015.csv” such as race, gender, and income. The managerial objective is to obtain an accurate model to predict the outcome of mortgage decisions. Further, we can assess which variables are more pertinent to make this assessment.

1. Here’s how we assessed the dataset:

housing\_data <- read.csv(file.choose(), header=T)  
library(dplyr)

##   
## Attaching package: 'dplyr'

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

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

library(ggplot2)  
library(tidyr)

Using sapply to assess the class of data and creating a data frame for future reference.

#Understanding classes for all variables   
  
Variable\_names <- data.frame(sapply(housing\_data, class))  
Variable\_names <- rename(Variable\_names, data\_class= sapply.housing\_data..class.)  
Variable\_names$num <- 1:78  
print(Variable\_names)

## data\_class num  
## action\_taken integer 1  
## action\_taken\_name factor 2  
## agency\_code integer 3  
## agency\_abbr factor 4  
## agency\_name factor 5  
## applicant\_ethnicity integer 6  
## applicant\_ethnicity\_name factor 7  
## applicant\_income\_000s integer 8  
## applicant\_race\_1 integer 9  
## applicant\_race\_2 integer 10  
## applicant\_race\_3 integer 11  
## applicant\_race\_4 integer 12  
## applicant\_race\_5 integer 13  
## applicant\_race\_name\_1 factor 14  
## applicant\_race\_name\_2 factor 15  
## applicant\_race\_name\_3 factor 16  
## applicant\_race\_name\_4 factor 17  
## applicant\_race\_name\_5 factor 18  
## applicant\_sex integer 19  
## applicant\_sex\_name factor 20  
## application\_date\_indicator integer 21  
## as\_of\_year integer 22  
## census\_tract\_number numeric 23  
## co\_applicant\_ethnicity integer 24  
## co\_applicant\_ethnicity\_name factor 25  
## co\_applicant\_race\_1 integer 26  
## co\_applicant\_race\_2 integer 27  
## co\_applicant\_race\_3 integer 28  
## co\_applicant\_race\_4 integer 29  
## co\_applicant\_race\_5 integer 30  
## co\_applicant\_race\_name\_1 factor 31  
## co\_applicant\_race\_name\_2 factor 32  
## co\_applicant\_race\_name\_3 factor 33  
## co\_applicant\_race\_name\_4 factor 34  
## co\_applicant\_race\_name\_5 factor 35  
## co\_applicant\_sex integer 36  
## co\_applicant\_sex\_name factor 37  
## county\_code integer 38  
## county\_name factor 39  
## denial\_reason\_1 integer 40  
## denial\_reason\_2 integer 41  
## denial\_reason\_3 integer 42  
## denial\_reason\_name\_1 factor 43  
## denial\_reason\_name\_2 factor 44  
## denial\_reason\_name\_3 factor 45  
## edit\_status integer 46  
## edit\_status\_name factor 47  
## hoepa\_status integer 48  
## hoepa\_status\_name factor 49  
## lien\_status integer 50  
## lien\_status\_name factor 51  
## loan\_purpose integer 52  
## loan\_purpose\_name factor 53  
## loan\_type integer 54  
## loan\_type\_name factor 55  
## msamd integer 56  
## msamd\_name factor 57  
## owner\_occupancy integer 58  
## owner\_occupancy\_name factor 59  
## preapproval integer 60  
## preapproval\_name factor 61  
## property\_type integer 62  
## property\_type\_name factor 63  
## purchaser\_type integer 64  
## purchaser\_type\_name factor 65  
## respondent\_id factor 66  
## sequence\_number integer 67  
## state\_code integer 68  
## state\_abbr factor 69  
## state\_name factor 70  
## hud\_median\_family\_income integer 71  
## loan\_amount\_000s integer 72  
## number\_of\_1\_to\_4\_family\_units integer 73  
## number\_of\_owner\_occupied\_units integer 74  
## minority\_population numeric 75  
## population integer 76  
## rate\_spread numeric 77  
## tract\_to\_msamd\_income numeric 78

|  |  |  |  |
| --- | --- | --- | --- |
| Nominal | Ordinal | Interval | Ratio |
| agency\_code |  |  | applicant\_income\_000s |
| applicant\_ethnicity |  |  | census\_tract\_number |
| applicant\_race\_1 |  |  | hud\_median\_family\_income |
| applicant\_sex\_name |  |  | loan\_amount\_000s |
| co\_applicant \_ethnicity |  |  | number\_of\_1\_to\_4\_family\_units |
| county\_name |  |  | number\_of\_owner\_occupied\_units |
| denial\_reason\_1 |  |  | minority\_population |
| heopa\_status |  |  | population |
| lien\_status |  |  | tract\_to\_msamd\_income |
| loan\_purpose\_name |  |  |  |
| loan\_type\_name |  |  |  |
| owner\_occupancy |  |  |  |
| preapproval |  |  |  |
| property\_type |  |  |  |
| purchase\_type |  |  |  |
| respondent\_id |  |  |  |
| state\_code |  |  |  |
| msamd |  |  |  |

Understanding levels for factor type variables in the data set:

Variable\_factor <- Variable\_names%>% filter(data\_class == "factor")   
factor <- Variable\_factor$num

The table Variable\_names explains what each variable type is. And those variables with factors/levels are explained in table Variable\_factors.

1. Now, we move on to understand the statistics of Data:

ratio <- c(8, 23, 71, 72, 73, 74,75,76,78) #these are Ratio Variables  
#inspecting and learning about ratio variables   
variable\_stat <- data.frame(summary(housing\_data[, ratio], na.rm=TRUE))  
  
#Inspecting IQR's  
  
IQR(housing\_data$applicant\_income\_000s, na.rm=TRUE)

## [1] 84

IQR(housing\_data$applicant\_income\_000s, na.rm=TRUE)

## [1] 84

IQR(housing\_data$census\_tract\_number, na.rm=TRUE)

## [1] 1223.02

IQR(housing\_data$hud\_median\_family\_income, na.rm=TRUE)

## [1] 13700

IQR(housing\_data$loan\_amount\_000s, na.rm=TRUE)

## [1] 264

IQR(housing\_data$number\_of\_1\_to\_4\_family\_units, na.rm=TRUE)

## [1] 1044

IQR(housing\_data$number\_of\_owner\_occupied\_units, na.rm=TRUE)

## [1] 892

IQR(housing\_data$minority\_population, na.rm=TRUE)

## [1] 31.44

IQR(housing\_data$population, na.rm=TRUE)

## [1] 2453

IQR(housing\_data$tract\_to\_msamd\_income, na.rm=TRUE)

## [1] 43.65001

#Inspecting SD's  
sd(housing\_data$applicant\_income\_000s, na.rm=TRUE)

## [1] 268.4713

sd(housing\_data$census\_tract\_number, na.rm=TRUE)

## [1] 2427.44

sd(housing\_data$hud\_median\_family\_income, na.rm=TRUE)

## [1] 16235.41

sd(housing\_data$loan\_amount\_000s, na.rm=TRUE)

## [1] 1173.204

sd(housing\_data$number\_of\_1\_to\_4\_family\_units, na.rm=TRUE)

## [1] 790.5034

sd(housing\_data$number\_of\_owner\_occupied\_units, na.rm=TRUE)

## [1] 609.3794

sd(housing\_data$minority\_population, na.rm=TRUE)

## [1] 29.03251

sd(housing\_data$population, na.rm=TRUE)

## [1] 1881.876

sd(housing\_data$tract\_to\_msamd\_income, na.rm=TRUE)

## [1] 53.10745

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Mean | Standard Deviation | IQR | Median |
| applicant\_income\_000s | 140,200 | 268,471 | 84,000 | 90,000 |
| census\_tract\_number | 1387 | 2,427.44 | 1223.02 | 305 |
| hud\_median\_family\_income | 78,224 | 16,235.41 | 13,700 | 71,300 |
| loan\_amount\_000s | 333,300 | **1,173.20** | 264,000 | 208,000 |
| number\_of\_1\_to\_4\_family\_units | 1,512 | 790.50 | 1,044 | 1,520 |
| number\_of\_owner\_occupied\_units | 1,214 | 609.38 | 892 | 1,196 |
| minority\_population | 29.20 | 29.03 | 31.44 | 17.23 |
| population | 4,749 | 1881.88 | 2453 | 4,554 |
| tract\_to\_msamd\_income | 117.92 | 53.12 | 43.65 | 106.75 |

1. We handled missing data in two ways:

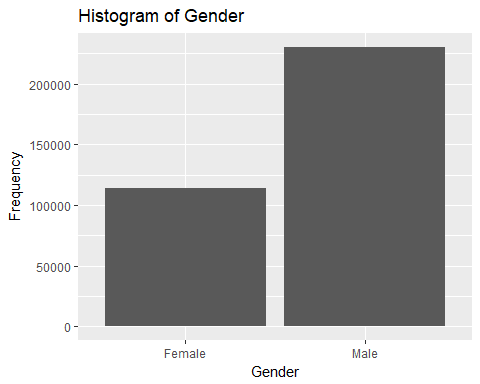
For data that was missing, we added na.rm=TRUE to the IQR() and sd() methods. This omits the data from the calculation so the method can accurately calculate the interquartile range and standard deviations.

Also, we can convert blanks to NA in the entire data set at the beginning of the beggining of the assessment.

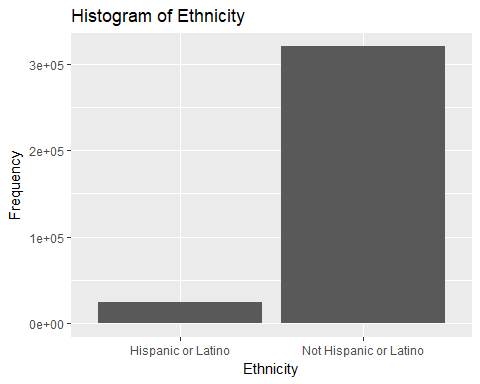
Variables <- variable.names(housing\_data) # converting all variable names to   
housing\_data <- data.frame(ifelse(housing\_data %in% c(""," ","NA"), NA, housing\_data))#Treating blanks and reconverting list to Data Frame   
# renaming the strings  
colnames(housing\_data) <- Variables

1. Assesing the variables using data vasualization techniques-

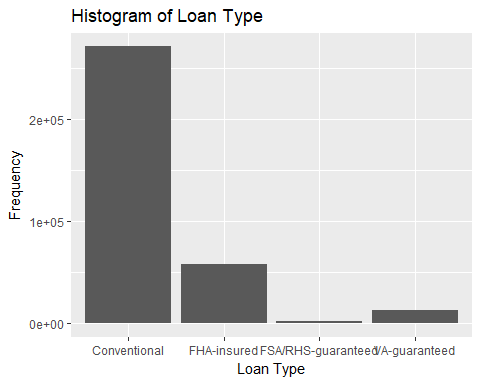
#gender filter  
housing\_data\_filtered <- filter(housing\_data,applicant\_sex\_name =="Male" | applicant\_sex\_name == "Female") %>%   
 droplevels()  
  
#ethnicity filter  
housing\_data\_filtered <- filter(housing\_data\_filtered,  
 applicant\_ethnicity\_name == "Not Hispanic or Latino" |   
 applicant\_ethnicity\_name == "Hispanic or Latino") %>%   
 droplevels()  
  
  
#histogram of gender  
housing\_data\_gender <- housing\_data\_filtered %>%   
 group\_by(applicant\_sex\_name) %>%  
 summarise(gender\_count = n())   
  
ggplot(housing\_data\_gender, aes(applicant\_sex\_name, gender\_count)) +  
 geom\_bar(stat = 'identity') +ggtitle("Histogram of Gender")+xlab("Gender")+ylab("Frequency")



#histogram of ethnicity  
housing\_data\_ethnicity <- housing\_data\_filtered %>%   
 group\_by(applicant\_ethnicity\_name) %>%  
 summarise(ethnicity\_count = n())   
  
ggplot(housing\_data\_ethnicity, aes(applicant\_ethnicity\_name, ethnicity\_count))+  
 geom\_bar(stat = 'identity') + ggtitle("Histogram of Ethnicity")+xlab("Ethnicity")+ylab("Frequency")



#histogram by loan type  
housing\_data\_loan <- housing\_data\_filtered %>%   
 group\_by(loan\_type\_name) %>%  
 summarise(loan\_count = n())   
  
ggplot(housing\_data\_loan, aes(loan\_type\_name, loan\_count)) +  
 geom\_bar(stat = 'identity') + ggtitle("Histogram of Loan Type")+xlab("Loan Type")+ylab("Frequency")

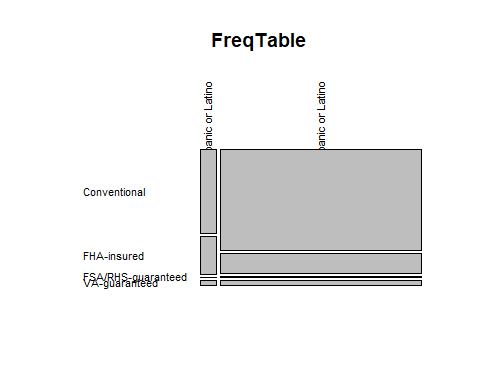
 Key Insights:

1. There are almost twice as many male applicants than female applicants
2. Significantly lower Hispanic or Latino Applicants in the pool( these only factor those who disclosed)
3. Conventional Loan applications are the highest followed by FHA insured, VA-guaranteed, and FSA/RHS guaranteed
4. Bivariate frequency distributions (tables or plots) for key variables

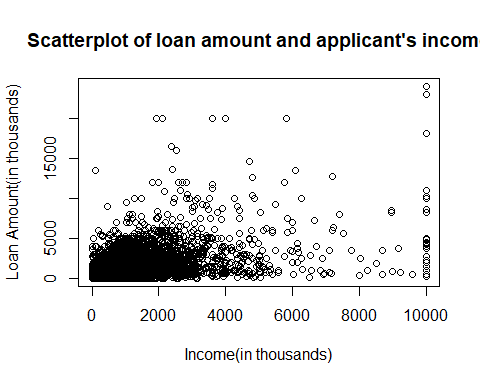
#frequency table for loan type and ethnicity  
head(factor(housing\_data\_filtered$applicant\_ethnicity\_name))

## [1] Not Hispanic or Latino Not Hispanic or Latino Not Hispanic or Latino  
## [4] Not Hispanic or Latino Not Hispanic or Latino Not Hispanic or Latino  
## Levels: Hispanic or Latino Not Hispanic or Latino

FreqTable <- table(housing\_data\_filtered$applicant\_ethnicity\_name,housing\_data\_filtered$loan\_type\_name)  
plot(FreqTable, las = 2)

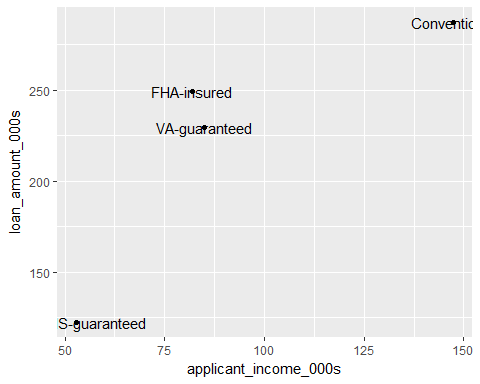


## We understand the distribution of data basis ethnicity and loan type. Conventional non Hispanic or Latino applicants are the largest in this data set. Other three loan types follow suit as per the bi-variate Frequecy table.   
  
# to observe the relation between loan amount and applicant's income  
plot(housing\_data\_filtered$loan\_amount\_000s ~ housing\_data\_filtered$applicant\_income\_000s, main="Scatterplot of loan amount and applicant's income",xlab="Income(in thousands)",ylab="Loan Amount(in thousands)")



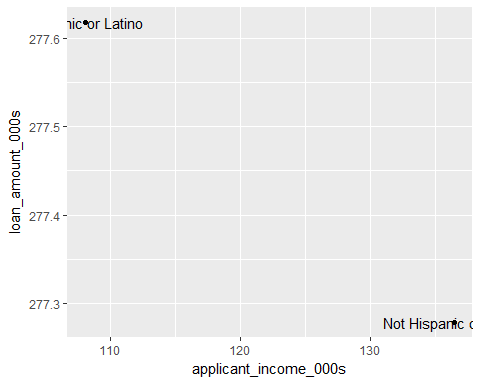
#As you can observe that there is a higher concentration of lower loan amounts as a factor of lower income. Although there are instances where the loan amounts for the same income are higher/audacious(in the real world), these outliers make a small portion of the data.   
  
# bivariate plot for relationship between applicant's income (mean) and loan amount (mean) by loan type  
housing\_data\_agg <- aggregate(housing\_data\_filtered, list(housing\_data\_filtered$loan\_type\_name), mean, na.rm=TRUE)

ggplot(housing\_data\_agg, aes(applicant\_income\_000s, loan\_amount\_000s)) + geom\_point() +  
 geom\_text(aes(label=housing\_data\_agg$Group.1))



#on working with averages for loan amount and income and factoring the loan type, we notice that conventional loan averages as a factor of mean income and loan are diagnoally dispered in the graph when compared to HS guarenteed. This means that the HS guarenteed applicants have lesser incomes and also apply for lower loan amounts, which in the real word makes sense. We can draw conclusions accordingly for the other two as they lie in between these plots.   
  
# bivariate plot for relationship between applicant's income (mean)   
# and loan amount (mean) by ethnicity  
housing\_data\_agg <- aggregate(housing\_data\_filtered, list(housing\_data\_filtered$applicant\_ethnicity\_name), mean, na.rm=TRUE)

ggplot(housing\_data\_agg, aes(applicant\_income\_000s, loan\_amount\_000s)) + geom\_point() +  
 geom\_text(aes(label=housing\_data\_agg$Group.1))



#on doing a similar analysis as above between income and loans on the basis of ethnicity, we learn that those applicants who identify as Hispanic or Latino have lower incomes and apply for high loan amounts. The contrary is observed for those who did not identify with this ethnicity.

1. We now understand how the data is spread out on the basis of income, gender, loan amount and ethnicity. Making assumptions about the probability of the getting the loan approved will require more advanced techniques such as creating models- both linear and r.part based, creating predictions, using bootstrapping techniques to improve these predictions. What we do know is the important variables especially those in the code called ratio(ratio variables).

We also know that we can group various collumns. Below are the assessment and the framework for the next step of predictions. This is just an example of how we will asses this data for creating predictive models and create them in the next submission. (only visible in R Markdown file)