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.

2. 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.

3. 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

4. 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.

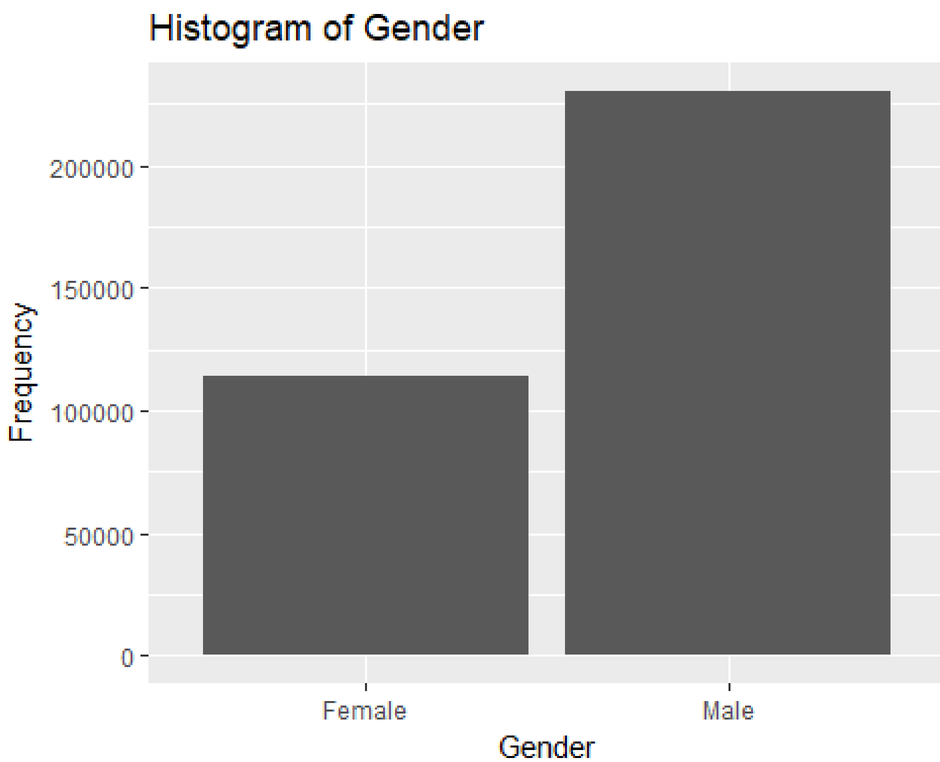
Also, we can convert blanks to NA in the entire data set at the beginning of the beggining of the assessment.

```
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
```

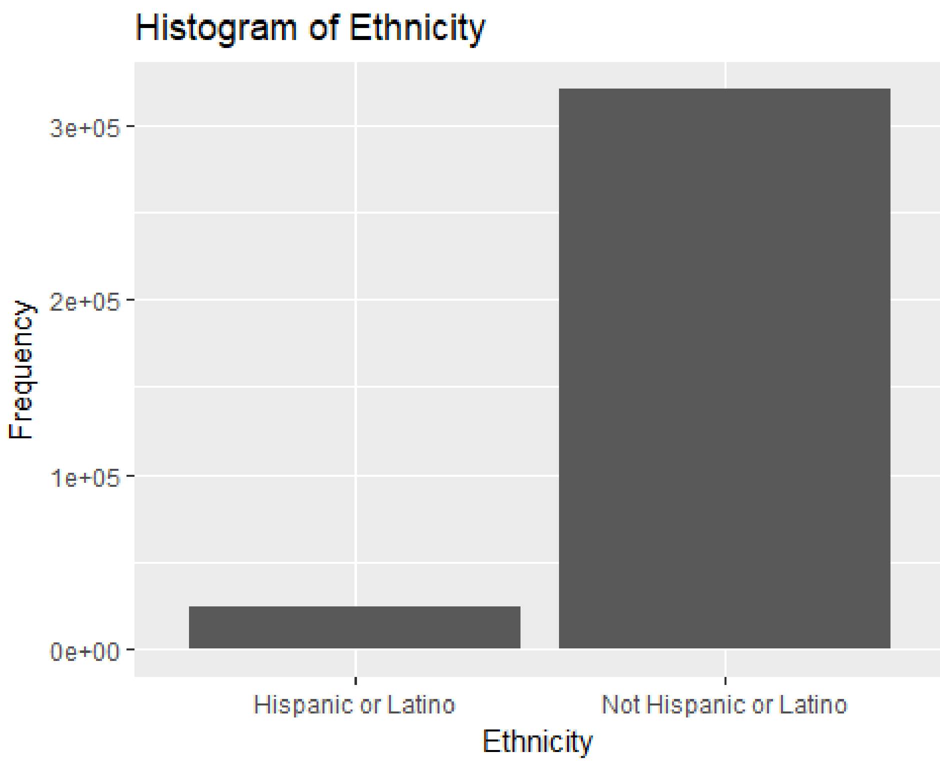
5. Assesing the variables using data vasualization techniques-

```
#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) %>%
```

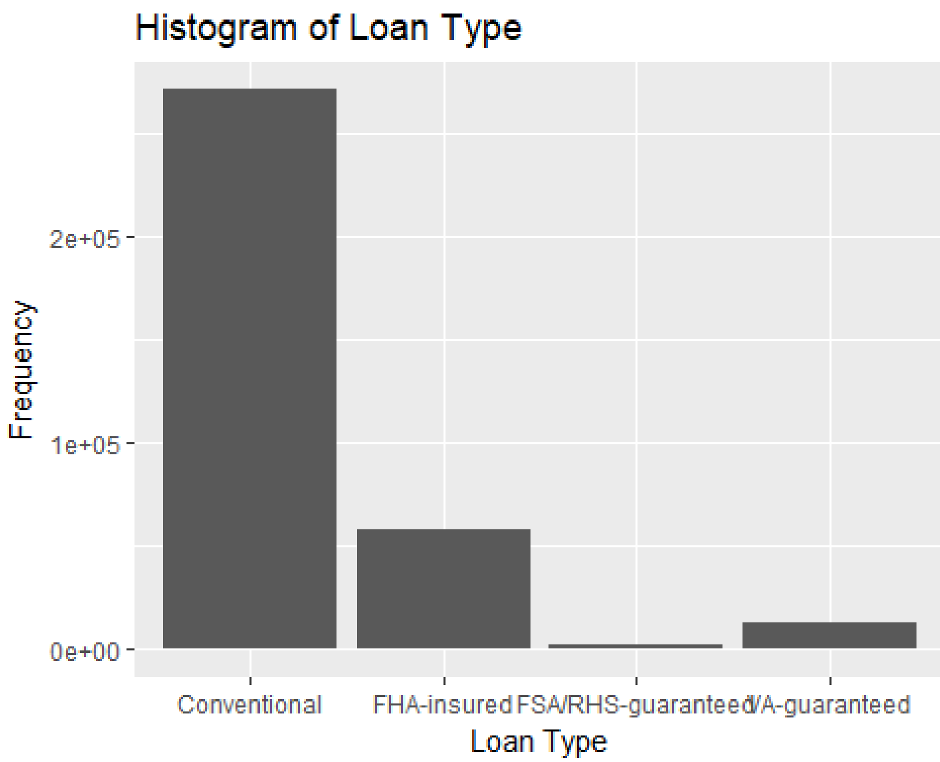
```
  summarise(gender_count = n())

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")
```

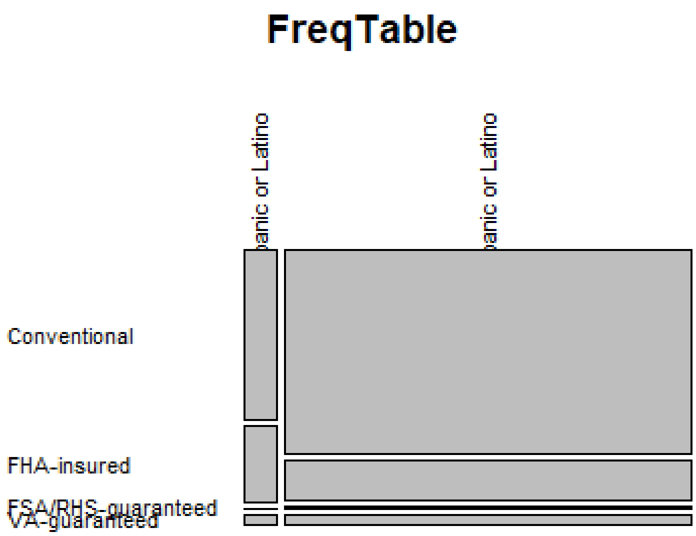


Key Insights:

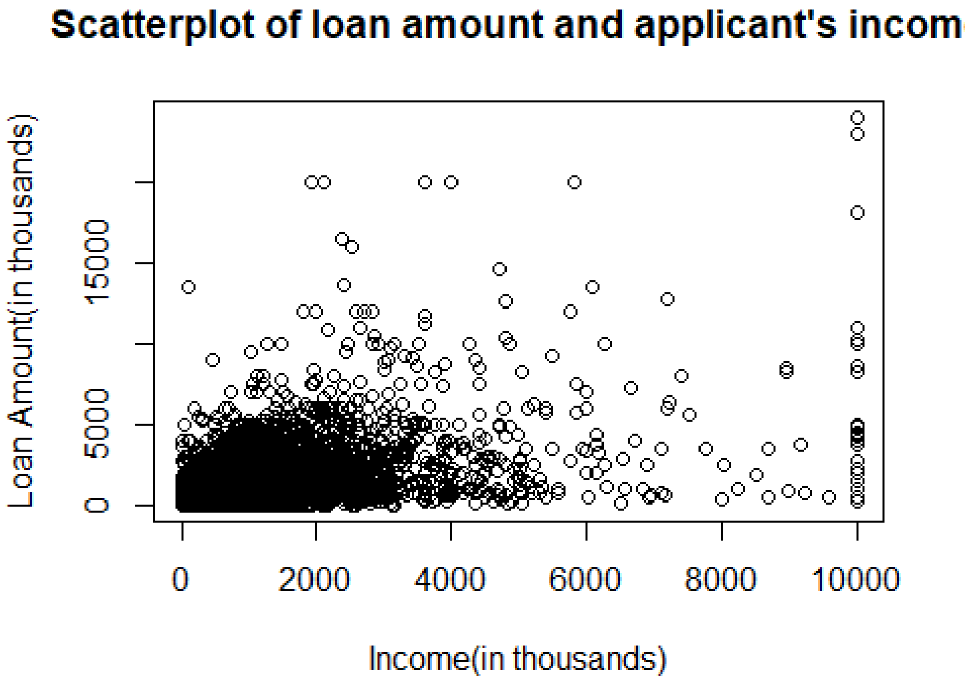
a) There are almost twice as many male applicants than female applicants
b) Significantly lower Hispanic or Latino Applicants in the pool(these only factor those who disclosed)
c) Conventional Loan applications are the highest followed by FHA insured, VA-guaranteed, and FSA/RHS guaranteed

6. Bivariate frequency distributions (tables or plots) for key variables

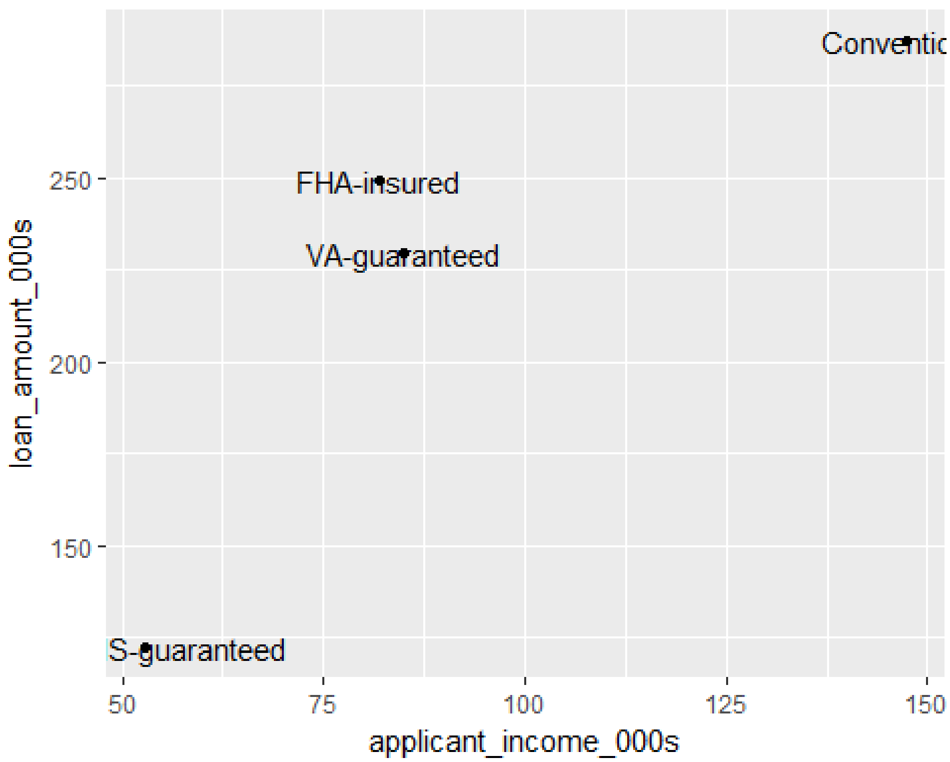
```
#frequency table for loan type and ethnicity
head(factor(housing_data_filtered$applicant_ethnicity_name))

## [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$applicant_ethnicity_name,housing_data_filtered$lo
an_type_name)
plot(FreqTable, las = 2)
```



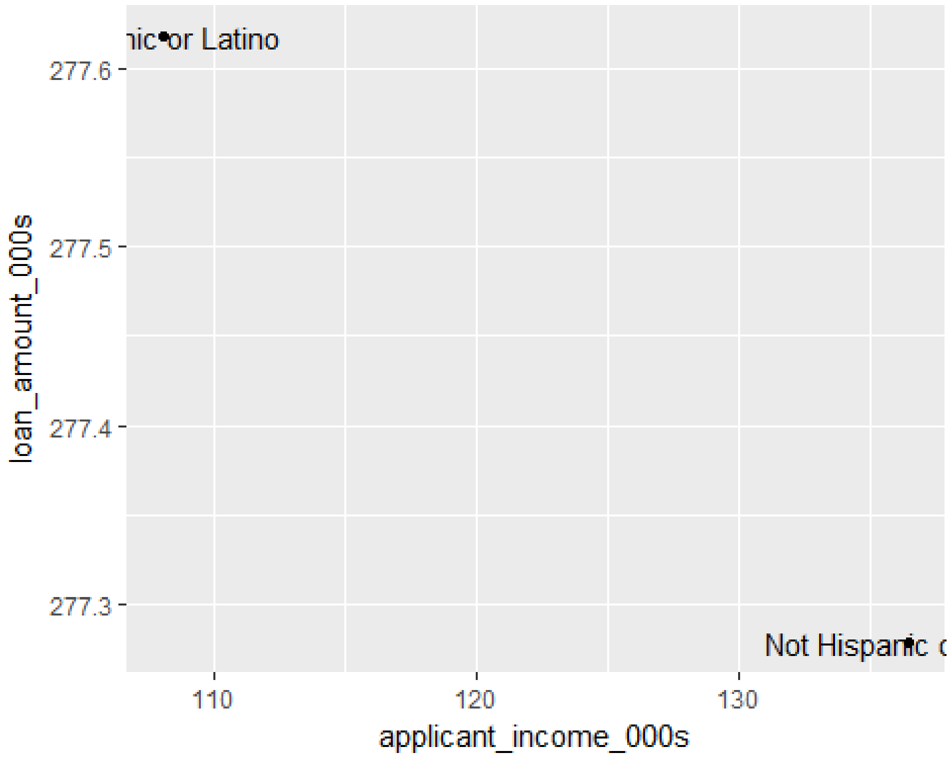

```
## We understand the distribution of data basis ethnicity and loan type.
Conventional non Hispanic or Latino applicants are the largest in this data
set. Other three loan types follow suit as per the bi-variate Frequecy table.

# 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.
```

7. 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)