HMDA Data Challenge

Author: **Naveen Kumar Aila**

Date: Mon Jul 25 2016

Table of Contents

Problem Statement:

To assist and help the “Change Financial bank” headquartered in Washington DC in taking a decision to enter the Home Loan market based on the historical data taken from Home Mortgage Disclosure Act (HMDA).

Data Munging:

Merging the application data and Institution data: hmda_init() function

In order to merge both the data sets, primary key is needed. Since “respondent name” has a unique id i.e. respondent_id. Along with that “as_of_year” and Agency_code can act as keys. Therefore, merging the two sets can be with above keys and by left joining the data and selecting the respondent name from the second data set. Thus by calling the hmda_init function, the expanded data set will be invoked.

```
hmda_init <- function()
{
  loans <- read.csv("2012_to_2014_loans_data.csv",na.strings = c("","NA"))
  institution <- read.csv("2012_to_2014_institutions_data.csv",na.strings =
c("","NA"))


  final_merged_data<- sqldf("select ln.*, inst.Respondent_Name_TS Respondent_Name
from loans ln left join institution inst on ln.Respondent_ID=inst.Respondent_ID and
ln.Agency_Code=inst.Agency_Code and ln.As_of_Year=inst.As_of_Year")

  return(final_merged_data)

}

final_merged_data = hmda_init()
```

Writing hmda_to_json() function.

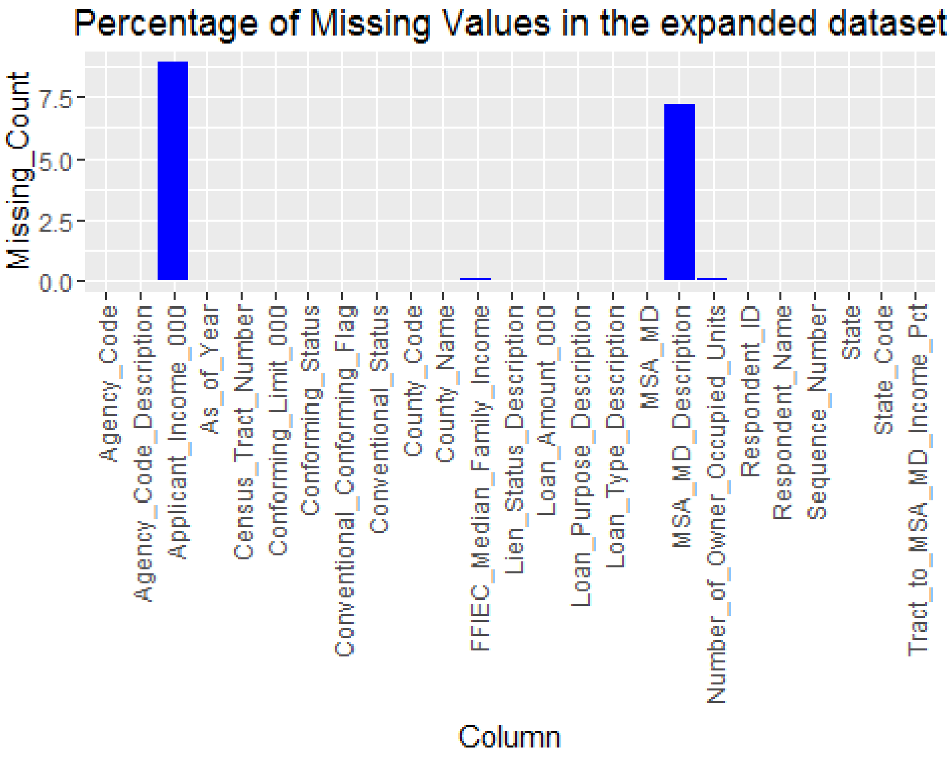
The below function allows a single state, confirmation flag in to the arguments. After passing the required parameters, the resultant data set will be invoked and will be stored or write to a JSON file.

```
hmda_to_json <- function(final_merged_data,states=FALSE,conf_flag=FALSE)
{
  if(states != FALSE)
  {
    final_merged_data = final_merged_data[final_merged_data$State %in% states,]

  }
  if(conf_flag != FALSE)
  {
    final_merged_data =
final_merged_data[final_merged_data$Conventional_Conforming_Flag %in% conf_flag,]

  }

  return(final_merged_data)
}

temp = final_merged_data[1:100000,]

return_val = hmda_to_json(temp,states = "WV",conf_flag = "Y")

JSONFile = toJSON(unname(split(return_val,1:nrow(return_val))))

write(JSONFile,file="JSONDATA.json")
```

Quality Check & Data Preprocessing:

Data for 2012_to_2014_loans_data.csv consists 1.3 Million records and 24 variables. Performed the following preprocessing steps on required variables:

- Variables Applicant_Income_000, Number_of_Owner_Occupied_Units, FFIEC_Median_Family_Income are converted from factor to numeric using relevant regular expressions.
- Variable Applicant_Income_000, Number_of_Owner_Occupied_Units, FFIEC_Median_Family_Income have the trailing zeros which are affecting the data type and have been removed.
- Applicant Income, Loan Amount , conditional_conventional_flag, Confirming limit, State, As_of_year, number of occupied units are found to be the important variables for majority of analysis with help of correlation and association rules along with numeric variables.
- Missing values in the data is not ignored, since this is unsupervised analysis. Missing values doesn’t affect the analysis. But in the further analysis missing values are handled.
- There are missing values (NA) values in the given data set. Out of all the variables, Applicant_Income_000 has highest percentage of missing values i.e. 8.5% of the data is missing for this variable.

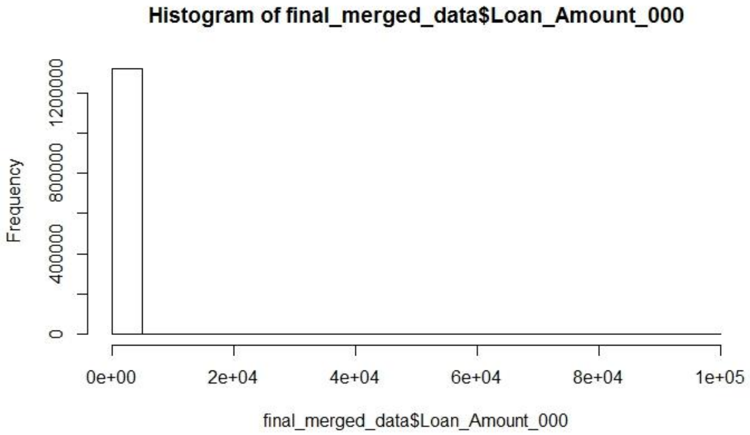


Univariate Analysis

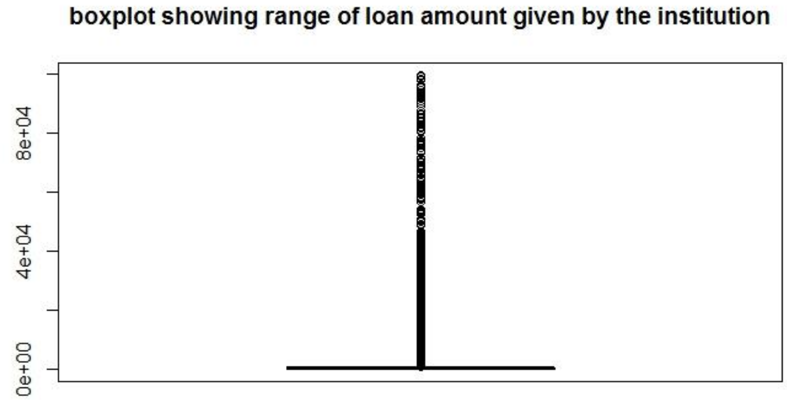
After plotting histograms for both Applicant_income000 and Loan_amount000 the data is looked right skewed. Due to the presence of outliers, the data is not normally distributed.

Below is univariate analysis for important variables.

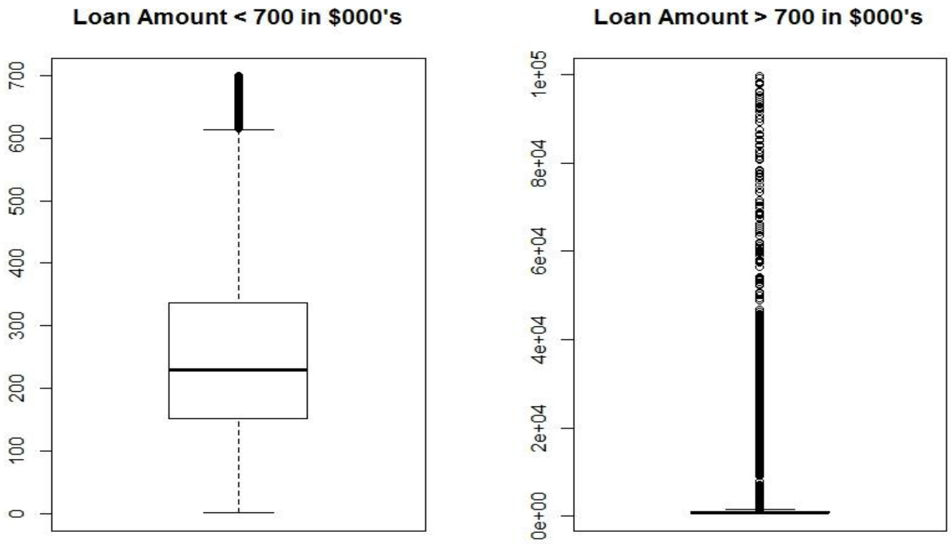
1)Loan_Amount_000



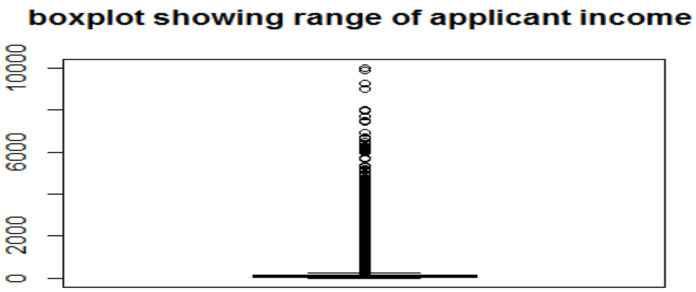
Below boxplot is not clear where the average Loan amount lies, so by taking subset of data is easy to analyze



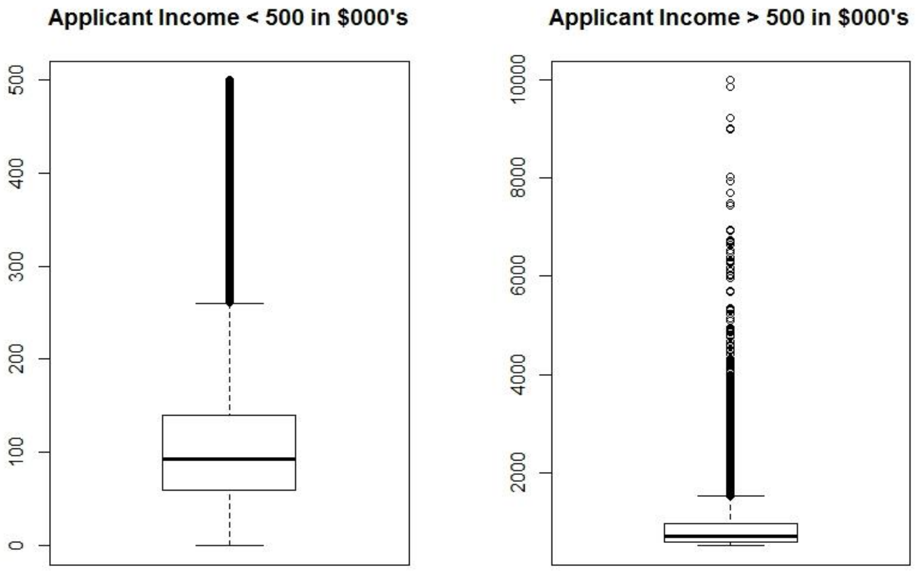
From below boxplot, the median of Loan Amount lies around 200k and the 75th quartile around 350k and there are outliers above 700k.
So studying data points > 700 gives some insights by studying this subset of data



2)Applicant_Income_000:

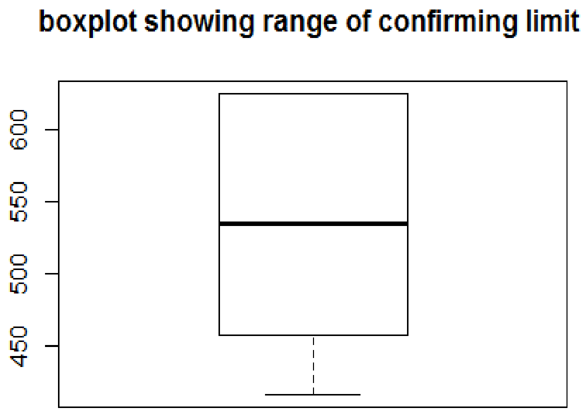


The above boxplot is hard to interpret, the median Loan amount lies around “0”. So by doing the analysis on subset of data i.e. by taking the 75th quartile as the base and subsetting the data in to two sets. The respective boxplots for the same is shown below.



From above boxplot, since the median of Applicant Income lies around 100 and the 75th quartile around 150 and there are outliers above 270. So studying data points > 270 gives some insights by studying this subset of data.

```
3)Conforming_Limit_000
```

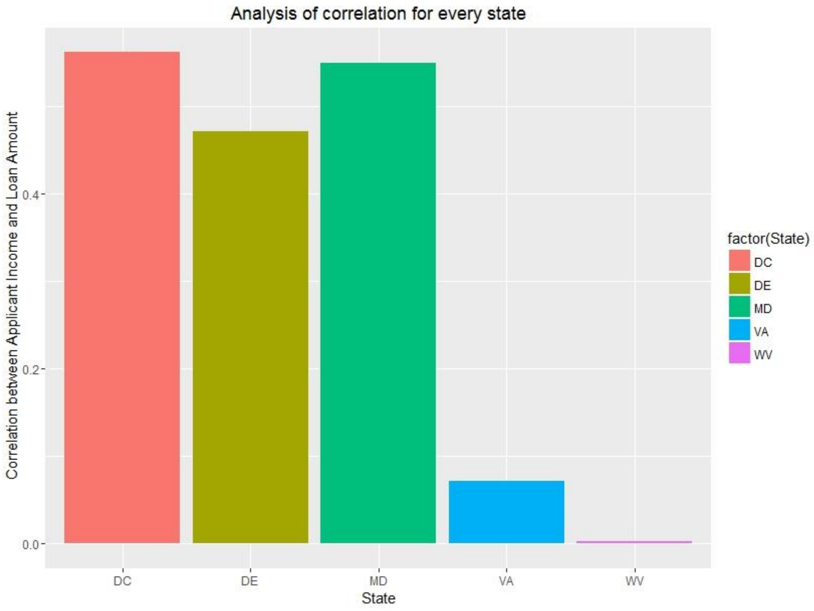


Above boxplot shows, the data distribution of Conforming_Limit_000 is little left skewed and median lies around 530.

Exploratory Data Analysis and Visualizations

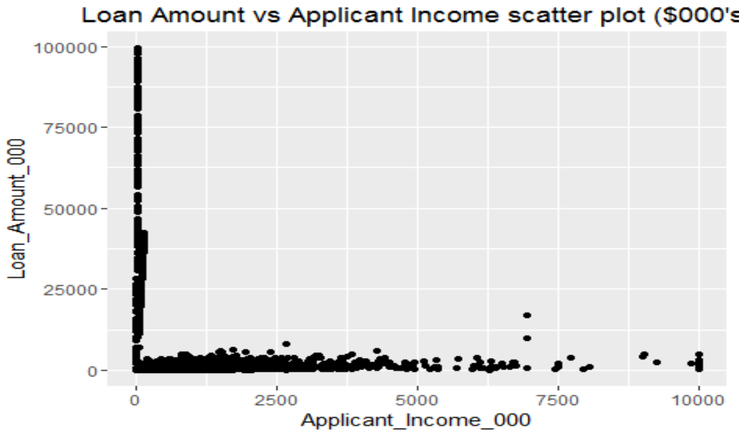
Bivariate Analysis & Finding Correlation between variables

```
Correlation between Income and Loan Amount by each state
Bar chart of Correlation between Applicant Income and Loan Amount
```



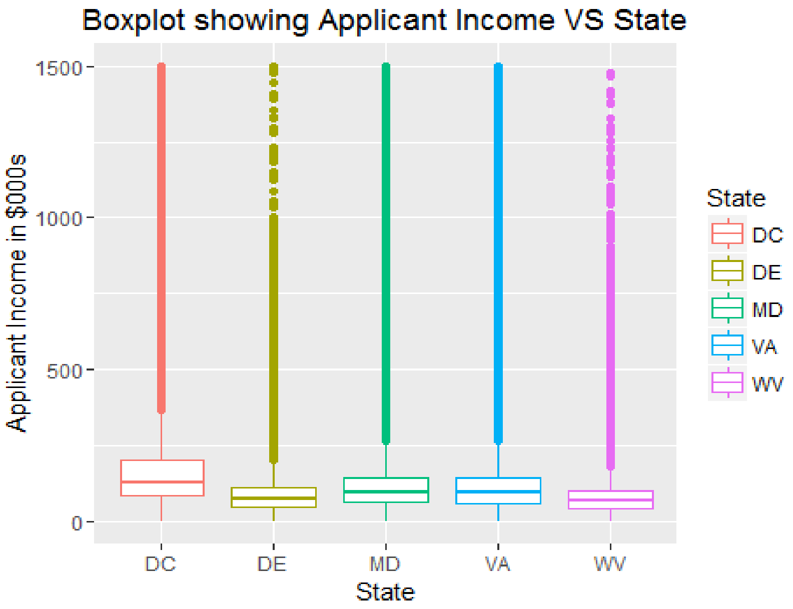
From the above chart, DC and DE states has high correlation between Applicant Income and loan amount. It indicates that Loan amount sanctioned has a positive correlation with applicant's income i.e. higher the applicant income higher the Loan amount sanctioned and vice versa.

Relation between Applicant Income and Loan Amount



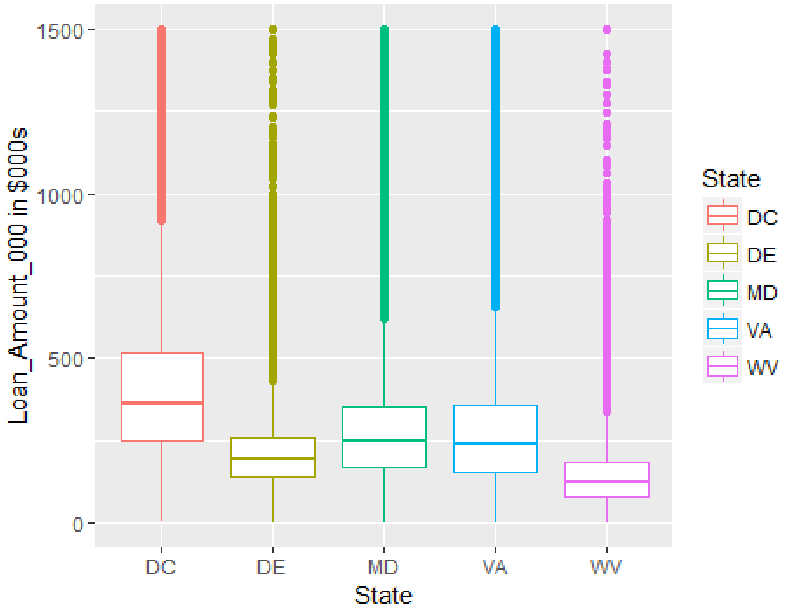
From the above plot, majority of Loan amount is given to the applicant with income is just greater than 0, i.e. around 100($000's)

Applicant Income by each state



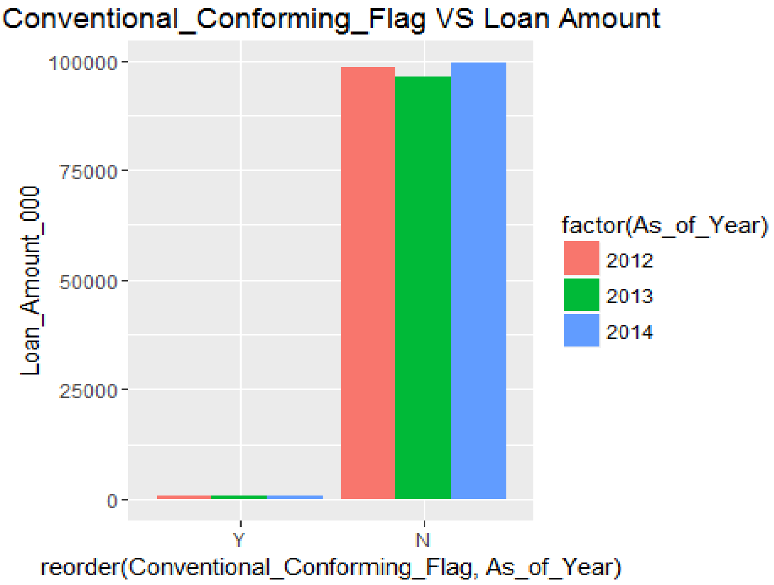
From the above boxplot, Average Applicant's income was greater for State "DC" from rest of the states which is around 125k$. States WV, DE have lower average income levels than MD, VA which have greater average applicant income and are almost same and there are many outliers (High Income groups)

Loan_Amount by each state



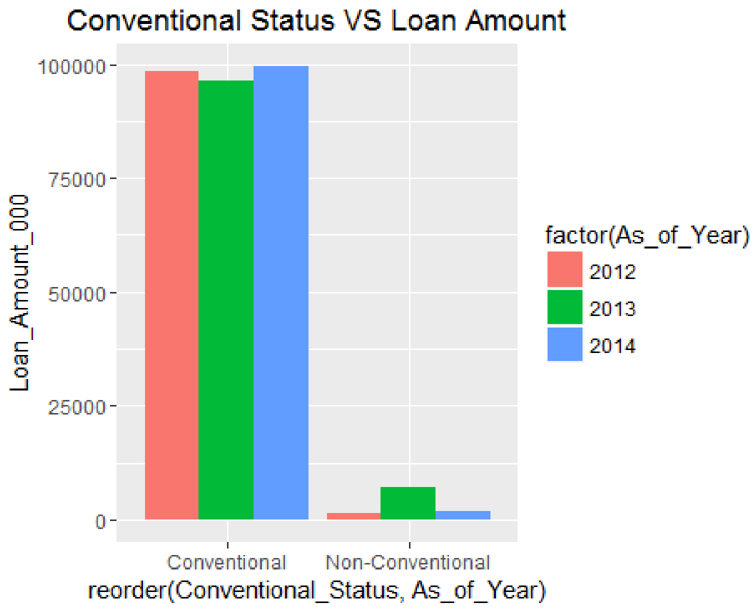
Above Boxplot shows the Loan amount given for every state. Similar to Applicant Income, "DC" state has highest average loan amount distributed.

Conventional_Conforming_Flag VS Loan amount (Then Conforming and conventional deep dive)



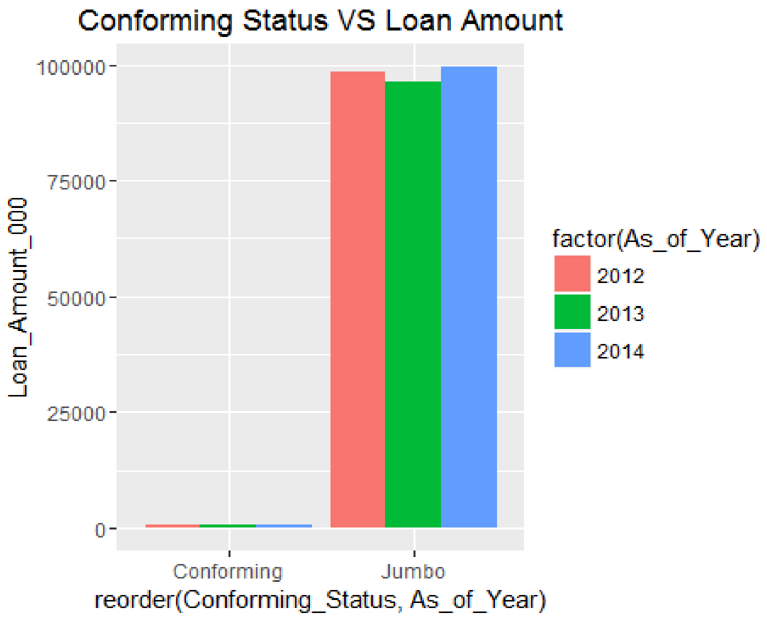
Above plot shows, majority of the loan amount involves either non-conventional or non-confirming

Conventional Status VS Loan Amount

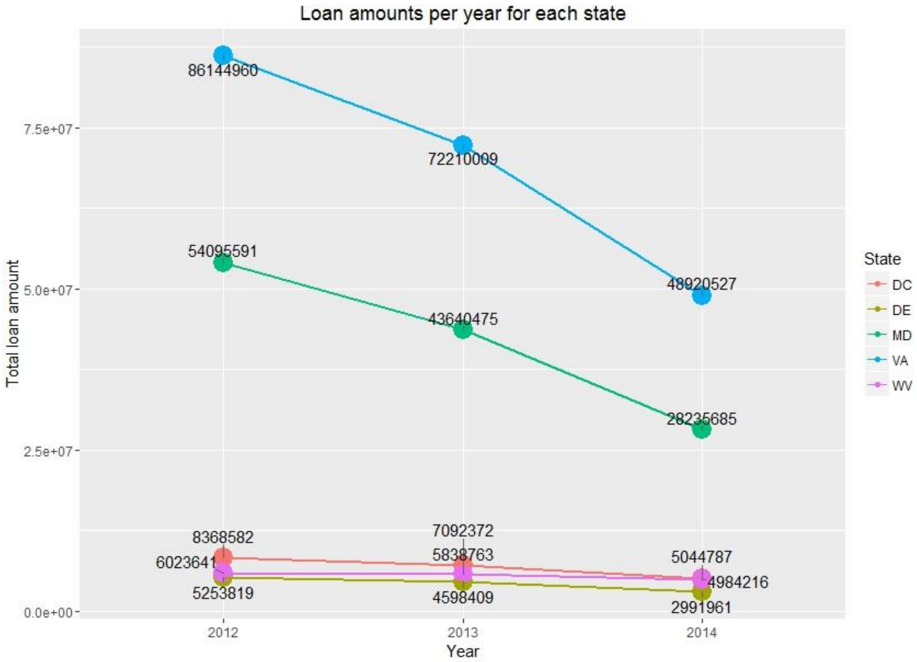


Above plot shows, the loan amount given was much higher with conventional i.e. not involved in any government program.

Conforming status VS Loan Amount

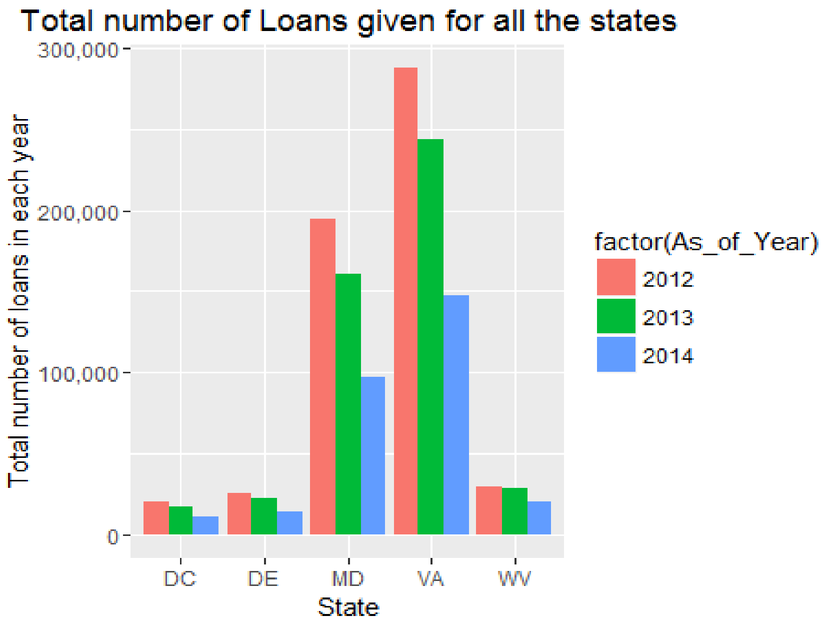


Loan amounts per year for each state



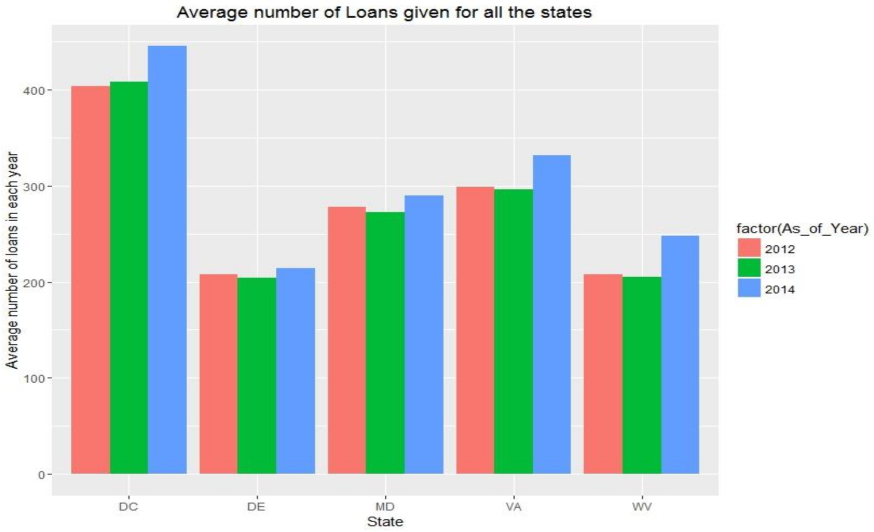
Above graph shows the total loan amount given each state for every year. From above graph we can observe that VA has given highest loan amount among all other states. on the other side, DE was the lowest. There is a trend that the loan amount provided by each state got decreasing every year

Total Number of Loans VS State by each year



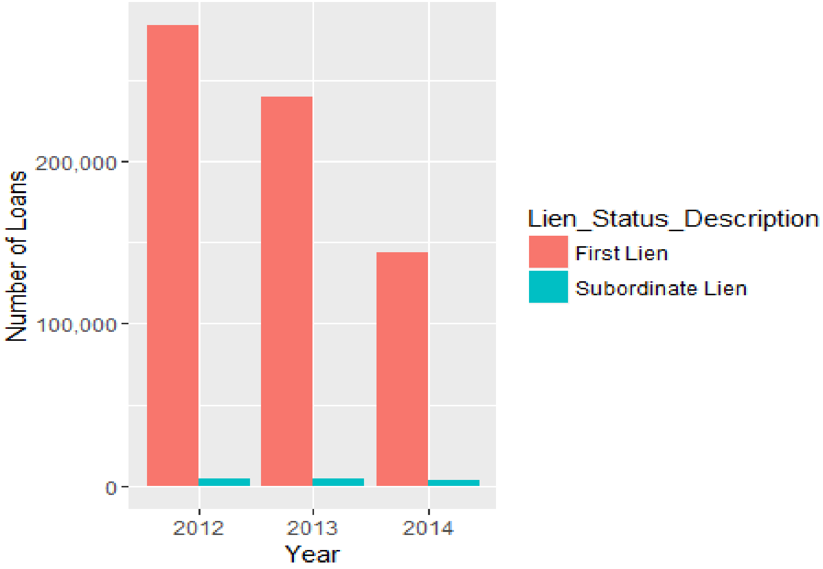
The above plot shows that Virginia has sanctioned the most number of loans between 2012 and 2014. Interestingly, the number of loans sanctioned is decreasing every year for each state.

Mean Loan amount VS State by each year



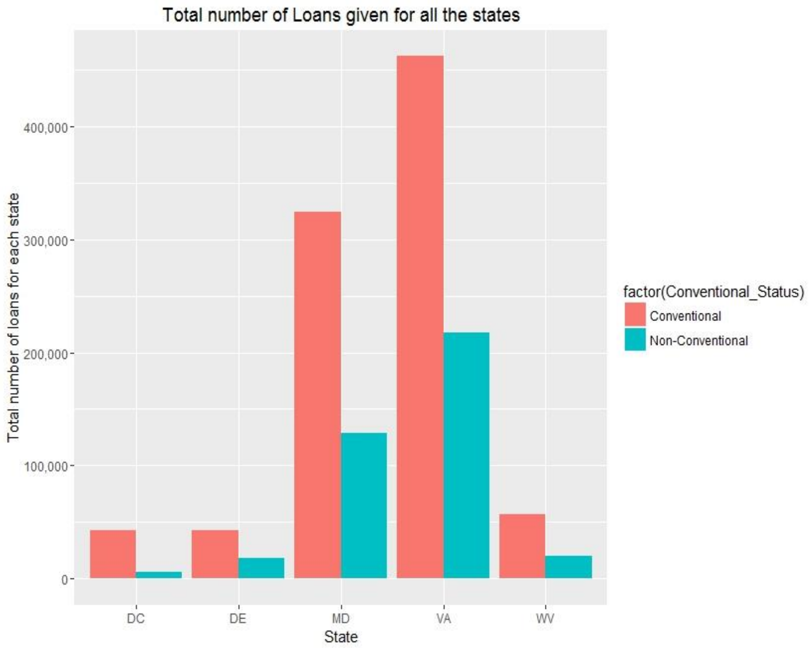
From the above plot, we can see DC state has a mean loan amount of $400,000 approximately in each year. Even though, the total number of loans sanctioned is least in DC, the average is high.

Number of first and Subordinate lien loans for each year



From the above plot, we can observe the number of least risk loans are decreasing. i.e. Increase of risk factor. Subordinate loans i.e. High risk loans remained almost constant throughout these years.

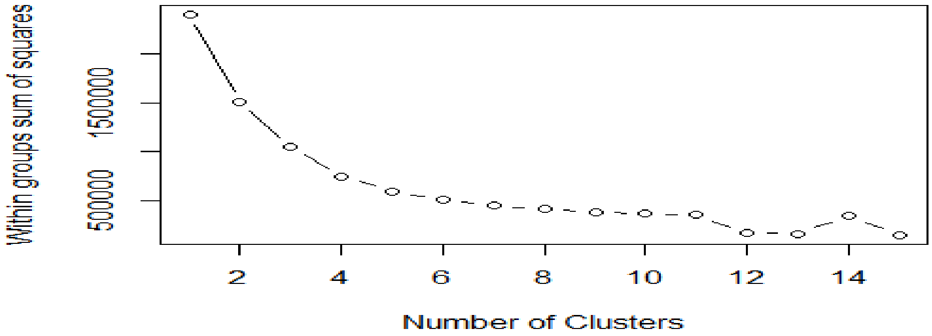
Number of Coventioanl& non-conventional loans for every State



From the above plot, the number of Conventional and non-conventional loans are high for VA state. Overall, the total number of coventional loans are given higher than non-conventiional loans in every state. MD & VA states gave Major contribution for both type of loans.

Cluster Analysis:

In order to identify homogenous groups of cases, i.e., observations, participants, respondents in a data, clustering analysis is done. In this particular data set applicant income and loan amount for different geographical locations differ. After this analysis, all the different types of income groups for particular regions depending up on their amount gets clustered. The clustered data can be useful for deep dive analysis based on income and loan amount for those particular geographical locations. All the analysis is done with K-means Clustering.



```
# From above we can conclude that we can cluster the data into 4 clusters, si
nce it has sharp bent.


table(fit$cluster, cstate$State)

##
##          DC     DE     MD     VA     WV
##   1  14721   4003  65835  93073   4076
##   2    522     77   1110   1206     83
##   3  32279  51856 345936 517955  69975
##   4      0      0      0    509     89
```

From the above table, 4 clusters are formed for all the states based on income group levels and loan amounts. i.e. The higher income groups with higher loan amount are clustered as one group and vice versa. Similary, Higher loan amount and lesser income level and vice versa.