HMDA DATA CHALLENGE

PALLAV ANAND

I have used R language and R studio platform for the analysis.

1. Data Munging

a) First step is setting the path for the data files to be loaded.

Code:

```
LoanPath='/Users/04pallav/Dropbox/Cap\ One\ 14th/data-challenge-data
master/2012_to_2014_loans_data.csv'
InstitutionPath='/Users/04pallav/Dropbox/Cap\ One\ 14th/data-challenge-data-
master/2012_to_2014_institutions_data.csv'

loans=read.csv(LoanPath)
institutions=read.csv(InstitutionPath)
```

Calculating summary statistics and basic Exploration of the two data sets.

```
dim(loans)
dim(institutions)
colnames(loans)
colnames(institutions)
summary(loans)
summary(institutions)
```

Merge the application data with institution data.

I merged the two datasets together so that each loan application has a "Respondent_Name" on the columns Agency_Code, Respondent_ID and As_of_Year.

```
mergedData <- merge(loans, institutions, by=c("Agency_Code","Respondent_ID", "As_of_Year"), all.x =
TRUE)
```

Create a new attribute that buckets "Loan_Amount_000" into reasonable groups

I find the break points for grouping by finding the range of the quantitative variable and iteration.

```
mergedData$Loan_Amount_000Cat<-cut(mergedData$Loan_Amount_000, c(0,10,100,500,1000,5000))
```

I observe here that many loans are in the order of 1 million to 5 million. I am a bit suspicious of these numbers as they strike me as too large. I will investigate whether or not these are data entry errors as the Loan_Amount are in thousands of dollars.

```
     Loan_Amount_000Cat
 (0,10]        :   6343
 (10,100]      : 134038
 (100,500]     :1073263
 (500,1e+03]   :  99344
 (1e+03,5e+03]:   7267
 NA's          :    903
```

b) **Next I** will **provide two functions**

hdma_init to read the data files and provide an expanded data set with the new bucketing attribute.

```
hmda_init <- function() {
 loans=read.csv(LoanPath)
 institutions=read.csv(InstitutionsPath)
 mergedData <- merge(loans, institutions, by=c("Agency_Code","Respondent_ID", "As_of_Year"), all.x =
TRUE)
 mergedData$Loan_Amount_000Cat<-cut(mergedData$Loan_Amount_000, c(0,10,100,500,1000,5000))
 return(mergedData)
}

mergedData=hmda_init()
```

I will use jsonlite package to covert data to json format. I am using jsonlite library which is specifically made for handling data to and from json. I am using dplyr library for filtering data.

I am saving the json file in R working directory by name of "data_filter.json"

```
library(jsonlite)
library(dplyr)


hmda_to_json= function(data, state, conventional_conforming,outputPath){
 if(missing(state) & missing(conventional_conforming)){data_filter=data }
 if (missing(state) & !missing(conventional_conforming)) { data_filter=data %>%
filter(Conventional_Conforming_Flag==conventional_conforming)}
 if (!missing(state) & missing(conventional_conforming)) { data_filter=data %>% filter(State==state)}
 if (!missing(state) & !missing(conventional_conforming)){ data_filter=data %>% filter(State==state &
Conventional_Conforming_Flag==conventional_conforming)}
 write(toJSON(data_filter), "data_filter.json")
 return(data_filter)
 }
```

I checked the function in various test cases

dim(hmda_to_json(mergedData))
dim(hmda_to_json(mergedData,"DC"))
dim(hmda_to_json(mergedData,"DC","Y"))
dim(hmda_to_json(mergedData,,"Y"))

2.Quality check

Loan_Amount_000

Our previous analysis on Loan_Amount_000 showed us that a lot of values are between 1 million and 5 million dollars which is a huge amount. This will need more investigation. I am **assuming it to be a data entry error** where people fail to realize that the Loan Amount is in 1000s. **So as a check on data quality, I will divide all values that are higher than 999,000$ by 1000 to get the actual loan amount.**

mergedData$Loan_Amount_000=ifelse(mergedData$Loan_Amount_000>999,mergedData$Loan_Amount_000/1000,mergedData$Loan_Amount_000)

After cleaning this is bucketing of Loan Amounts

```
     Loan_Amount_000Cat
 (0,10]       :  14885
 (10,100]     : 134914
 (100,500]    :1073263
 (500,1e+03]  :  98096
 (1e+03,5e+03]:      0
```

Respondent Names

As we want each loan application to have a unique Respondent Name. We cannot allow NA values in this column.
We can print out the number of Null values for the Respondent name column like this

print(paste("The number of Null Values in Respondent_Name_TS is",sum(is.na(mergedData$Respondent_Name_TS))))

Other quality checks

- **Find and remove Duplicate Records**

  ```
  mergedData=unique(mergedData)
  print(paste("The number of duplicate rows removed is",nrow(mergedData)-nrow(unique(mergedData))))
  ```

- **Checking the Conventional_Conforming_Flag**

  As our area of interest is conventional/conforming loans we can check the if these flags are correct and correct them if they are not.

```
a=mergedData %>% filter(Conventional_Status=='Conventional' & Conforming_Status=='Conforming')
if(all(a$Conventional_Conforming_Flag == 'Y')) {print("All Conventional_Conforming_Flags are correct ")}
else { print("Resetting the Conventional Conforming Flags")
 mergedData[mergedData$Conventional_Status=='Conventional' &
mergedData$Conforming_Status=='Conforming',"Conventional_Conforming_Flag"]}
```

- **Missing Values and outliers can be judged by inspection of summary results**

```
summary(mergedData)
```

Packing everything into data quality function

```
dataQualityCheck=function(testData) {

testData$Loan_Amount_000=ifelse(testData$Loan_Amount_000>999,testData$Loan_Amount_000/1000,testData$Loan_Amount_000)


 print(paste("The number of duplicate rows removed is",nrow(testData)-nrow(unique(testData))))
 testData=unique(testData)

 print(paste("The number of Null Values in Respondent_Name_TS is",
       sum(is.na(testData$Respondent_Name_TS ))))

 a=testData %>% filter(Conventional_Status=='Conventional' & Conforming_Status=='Conforming')
 if(all(a$Conventional_Conforming_Flag == 'Y')) {print("All Conventional_Conforming_Flags are correct ")}
 else { print("Resetting the Conventional Conforming Flags")
  testData[testData$Conventional_Status=='Conventional' &
 testData$Conforming_Status=='Conforming',"Conventional_Conforming_Flag"]}
}
```

I think it is important to monitor some other columns like “Applicant_Income_000” .
It is important to remember that from metadata definition columns like “Applicant_Income_000” and “Loan_Amount_000” are in thousands of dollars.

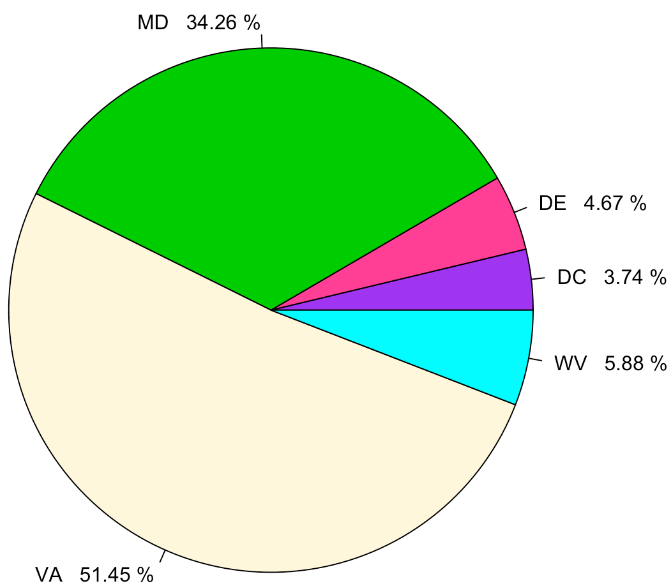
3. Data Narrative

Let us have a look at the market share in different states

```
a=mergedData %>% group_by(State) %>% summarize(LoanVolume=n())%>%
mutate(prcent = round(100*LoanVolume/sum(LoanVolume), 2))

pie(a$prcent,paste(a$State,' ',a$prcent,'%'),col=c("purple", "violetred1", "green3","cornsilk", "cyan",
"white"))
```

Proportion of Loan_Volume in Different States

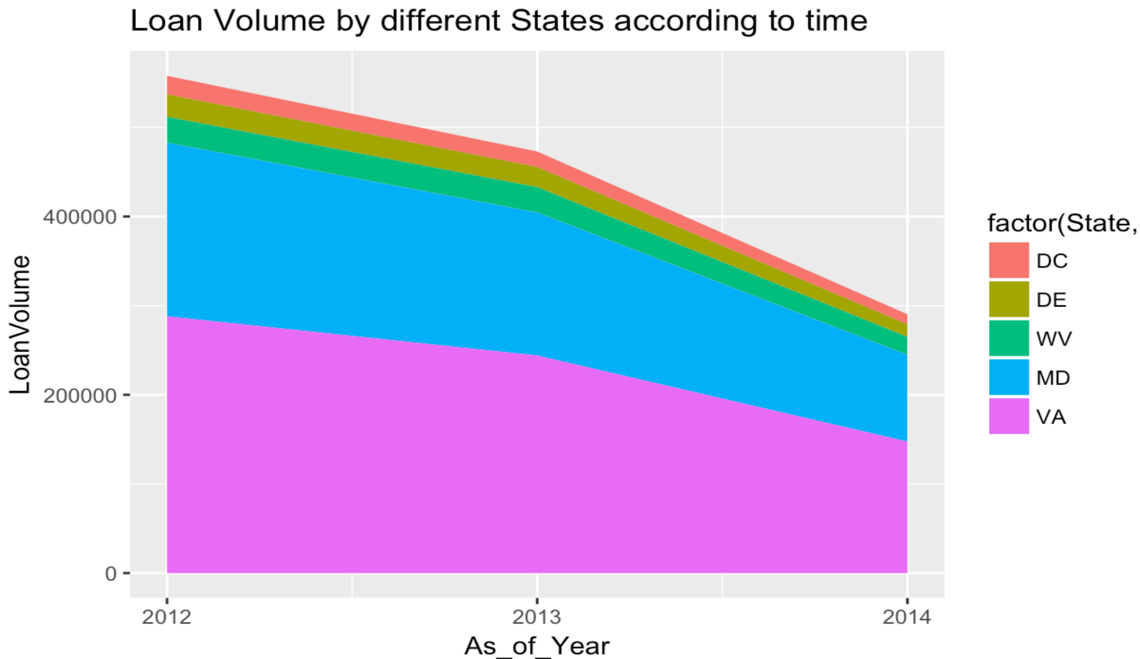


Inference

We see that Virginia and Maryland have around 85% of the market share in loans!

Let us explore how the market is distributed in different states according to time.

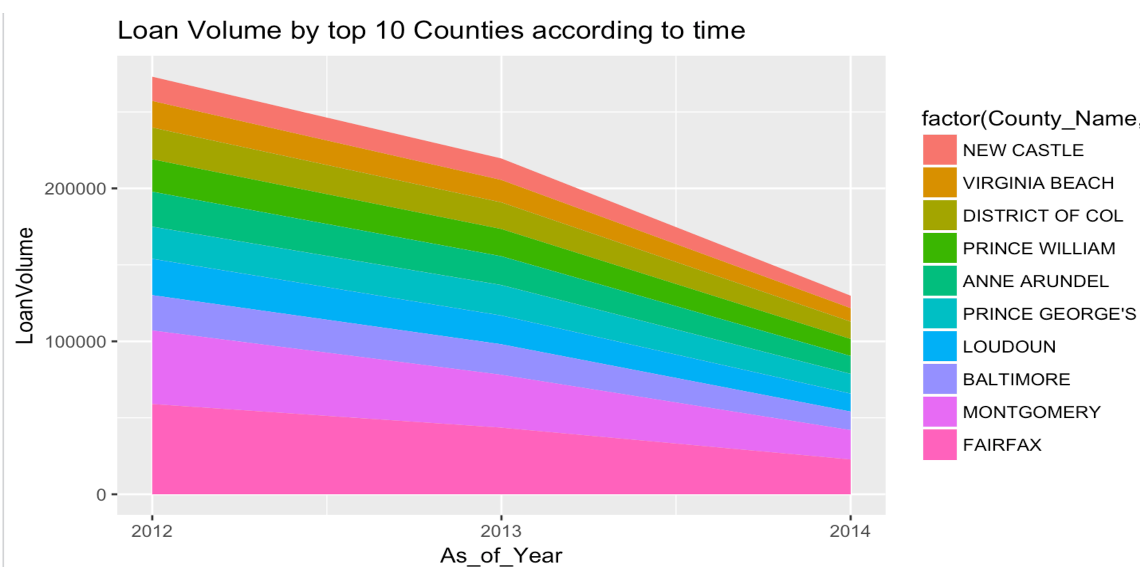
```
library(ggplot2)
a=mergedData %>% group_by(State,As_of_Year) %>% summarize(LoanVolume=n())
options(scipen=10000)
positions <- c("VA","MD","WV","DE","DC")
ggplot(a, aes(x = As_of_Year, y =LoanVolume,fill=factor(State,levels=rev(positions))))+
geom_area(stat ="identity",position ='stack')+
ggtitle("Loan Volume by different States according to time")+scale_x_continuous(breaks = c(2012:2014))
```



<u>Inference</u>
We see that the market is concentrated in a few states and also it seems to declining from 2012 to 2014.

<u>Let us see how the market is distributed by counties.</u>

#Finding the top 10 counties by Loan_volume
```
a=mergedData %>% group_by(County_Name) %>% summarize(LoanVolume=n())
a=a[order(-a$LoanVolume),][1:10,]
positions=a$County_Name
a=mergedData %>% filter(County_Name %in% positions) %>% group_by(County_Name,As_of_Year) %>%
summarize(LoanVolume=n())
ggplot(a, aes(x = As_of_Year, y =LoanVolume,fill=factor(County_Name,levels=rev(positions))))+
geom_area(stat ="identity",position ='stack')+
ggtitle("Loan Volume by different Counties according to time")+scale_x_continuous(breaks =
c(2012:2014))
```



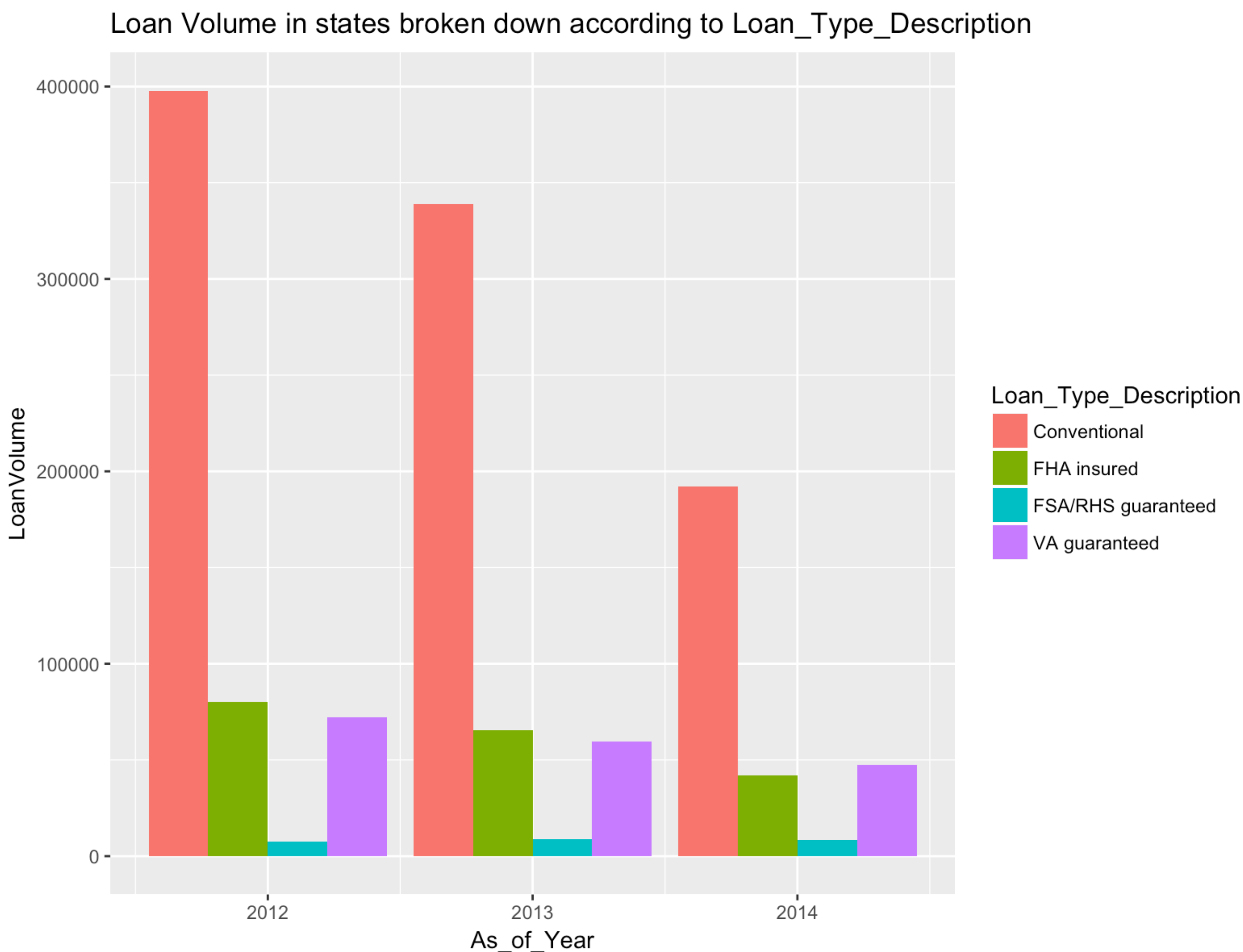
Inference

We see that Fairfax and Montgomery counties have the largest markets for loans.
From this plot also we can see the declining trend in volume of applications.

Let us see how loans are distributed across the States according to type of loan

```
a=mergedData %>% group_by(As_of_Year,Loan_Type_Description) %>% summarize(LoanVolume=n())

ggplot(a, aes(x = As_of_Year, y =LoanVolume,fill=Loan_Type_Description))+geom_bar(stat
="identity",position = "dodge")+ggtitle("Loan Volume in states broken down according to
Loan_Type_Description")
+scale_x_continuous(breaks = c(2012:2014))
```



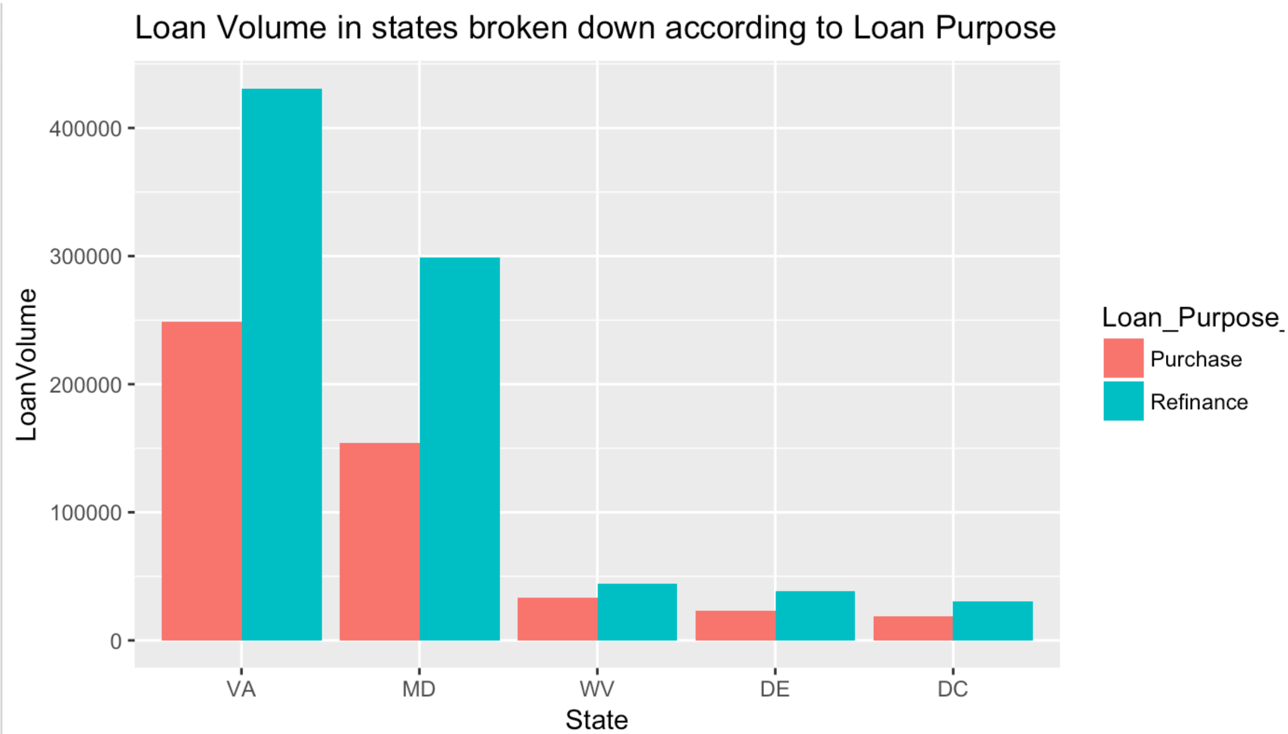
Inference

It seems that the market for Conventional home loans is much bigger than non conventional loans.

Let us explore how loans are distributed across the states according to purpose of loan

```
a=mergedData %>% group_by(State,Loan_Purpose_Description) %>% summarize(LoanVolume=n())
positions <- c("VA","MD","WV","DE","DC")

ggplot(a, aes(x = State, y =LoanVolume,fill=Loan_Purpose_Description))+geom_bar(stat
="identity",position = "dodge")+
scale_x_discrete(limits = positions)+ggtitle("Loan Volume in states broken down according to Loan
Purpose")
```

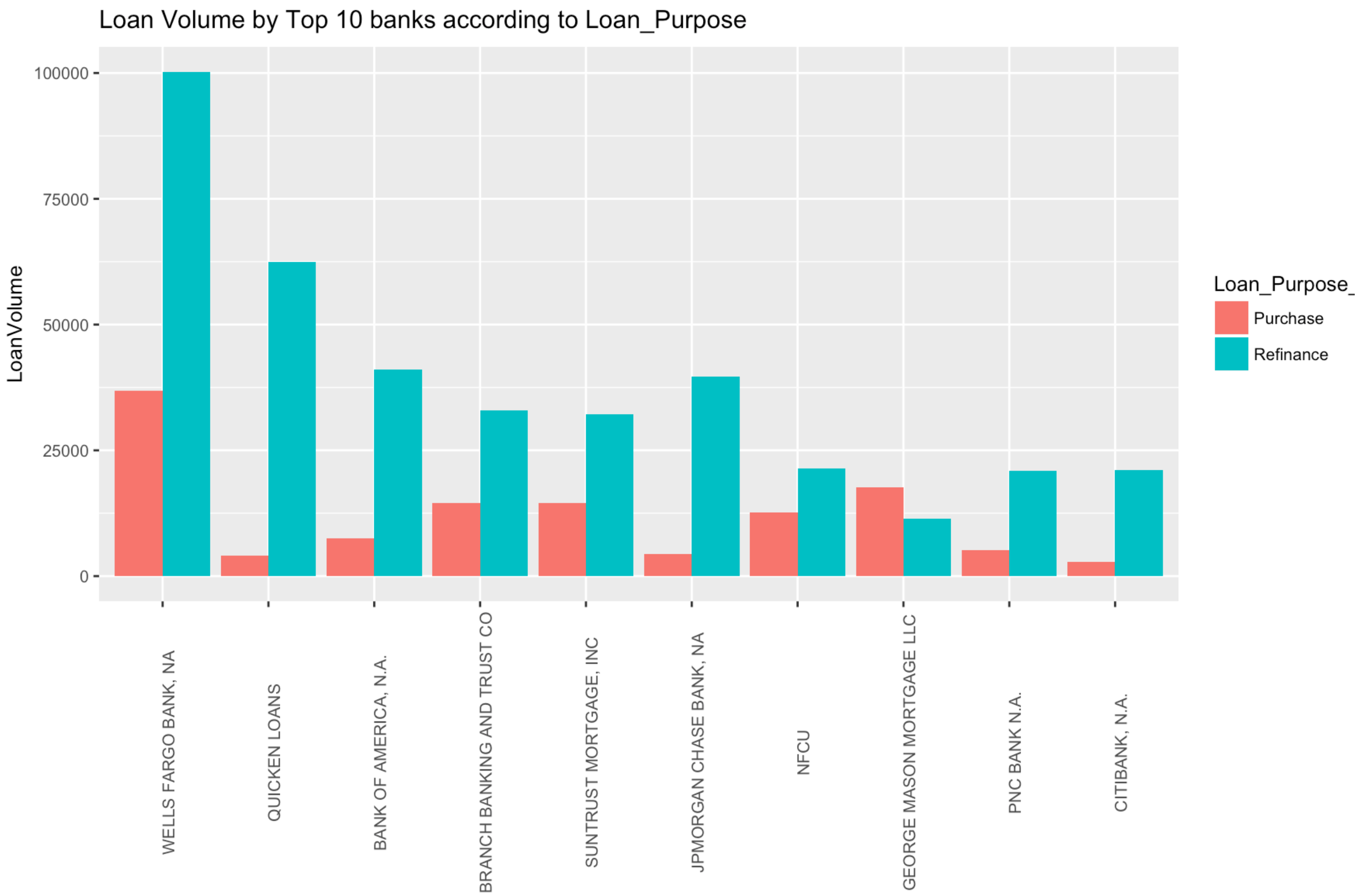


Inference

We see that across all the states Refinance loans are more in number than Purchase loans. It is possible than refinance loans are less risky that's why there is a bigger market for refinance loans than Purchase loans.

Let us have a look at the strategy of the biggest players in the market

```
####Lets find out the top 10 banks by Loan Volume over all the three years
a=mergedData %>% group_by(Respondent_Name_TS) %>% summarize(LoanVolume=n())
a=a[order(-a$LoanVolume),][1:10,]
positions=a$Respondent_Name_TS
a=mergedData %>% group_by(Respondent_Name_TS,Loan_Purpose_Description) %>%
summarize(LoanVolume=n())
ggplot(a, aes(x =Respondent_Name_TS,y =LoanVolume,fill=Loan_Purpose_Description))+
geom_bar(stat ="identity",position ='dodge')+scale_x_discrete(limits = positions)+
ggtitle("Loan Volume by Top 10 banks according to Loan_Purpose")+theme(text = element_text(size=10),
axis.text.x = element_text(angle=90, vjust=1))
+geom_text(aes(label=..count..) ,vjust = -1)
```



Inference

We see that the top players in the market have much more Refinance loans than Purchase Loans. This seems to confirm our hypothesis about the refinance loans being a bigger and safer market.

Recommendation

I think it will be a good idea to follow the strategy of the market leaders who are operating in the biggest markets and issuing more refinance loans than purchase loans. Change financial can follow a similar strategy and try to take a market share in the biggest markets.
It also seems that the market is on a decline right now so Change Financial should get more historical data and forecast the future trend before making a significant investment.