**CIS 5270 Project 2: Analysis of Marketplace Lending using Rstudio**

16

**Analysis of Marketplace Lending**

**Using Rstudio**

**Kaustubh Ambikesh Padhya CIN: 305861379**

**Vignesh Srinivas Ravi Shankar CIN: 305383629**

1. **Dataset URL’s:**

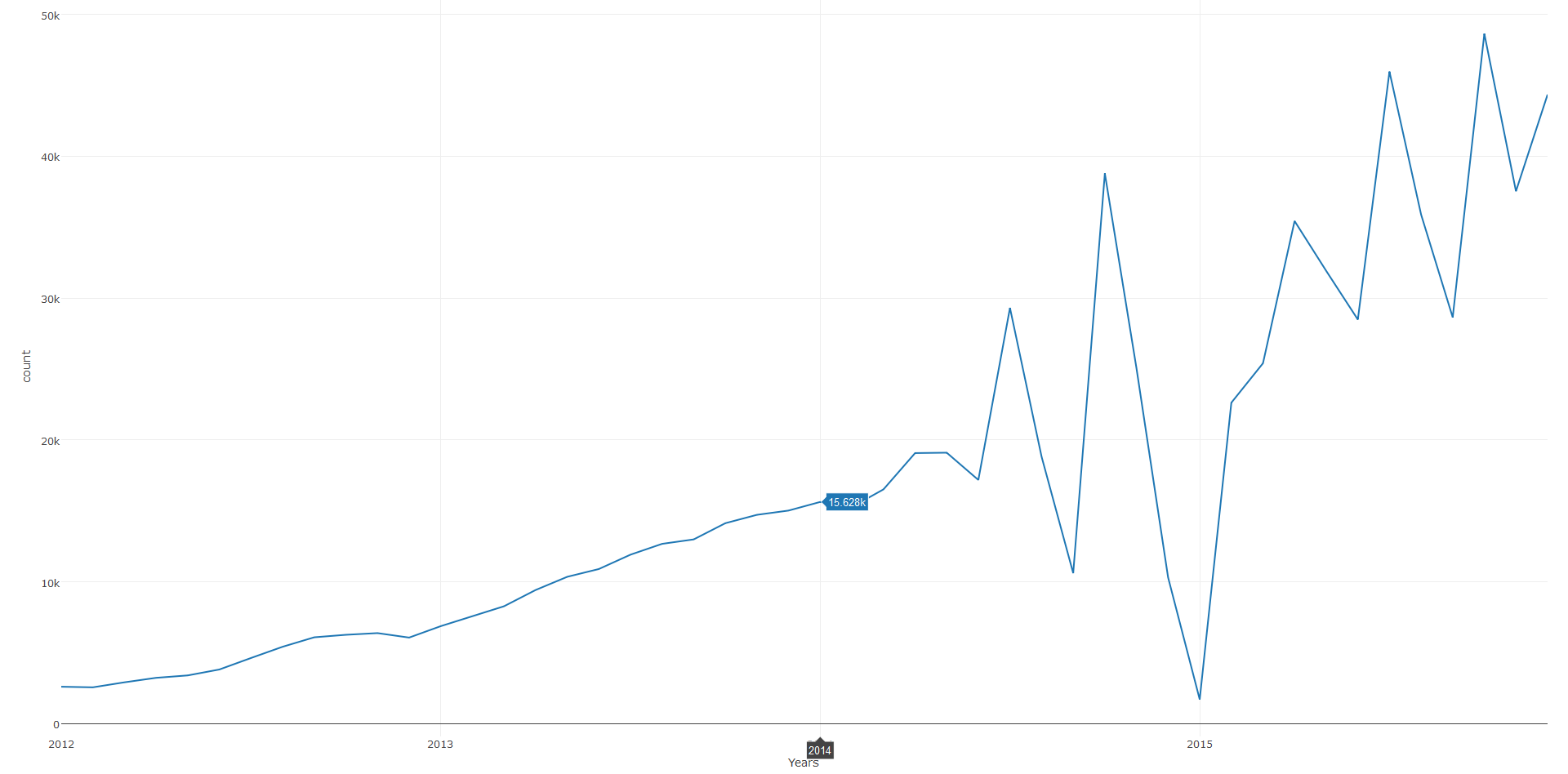
<https://www.lendingclub.com/info/download-data.action>

These URL’s contain the entire data regarding peer to peer Lending club Loans. The analysis is about Lending Club Loan data from year 2012 to 2015. The dataset has 111 columns and contains the information like, Loan amount, term, interest rate, monthly installment amount, current loan status, purpose of loan, city, Loan period etc., We are planning to analyze various figures like, most common purpose for taking loans, what type of loans have the maximum interest rate, State with maximum loans taken, percentage distribution of loan taken by people based on home ownership (own/rent/mortgage), Debt to income ratio of people taking loans for various purposes, average interest rate for loans issued over time, and reason for loan rejection. This analysis can be used as a guide for the future borrowers and investors who chose Lending Club for taking loan/invest.

1. **Data Cleaning:**

|  |  |
| --- | --- |
| **Problem** | **Un-cleaned and Cleaned Data** |
| **Deleting /Removing irrelevant columns** | The dataset contained 111 columns. We removed unwanted columns and cleaned the data which contains 26 columns. |
| **Splitting Columns** | The issue\_date column had a united date, but we required separated date by month and year. Hence we split the column into two. |
| **Deleting Columns/ Rows with empty values & updating it** | Our dataset did not contain any empty rows. |

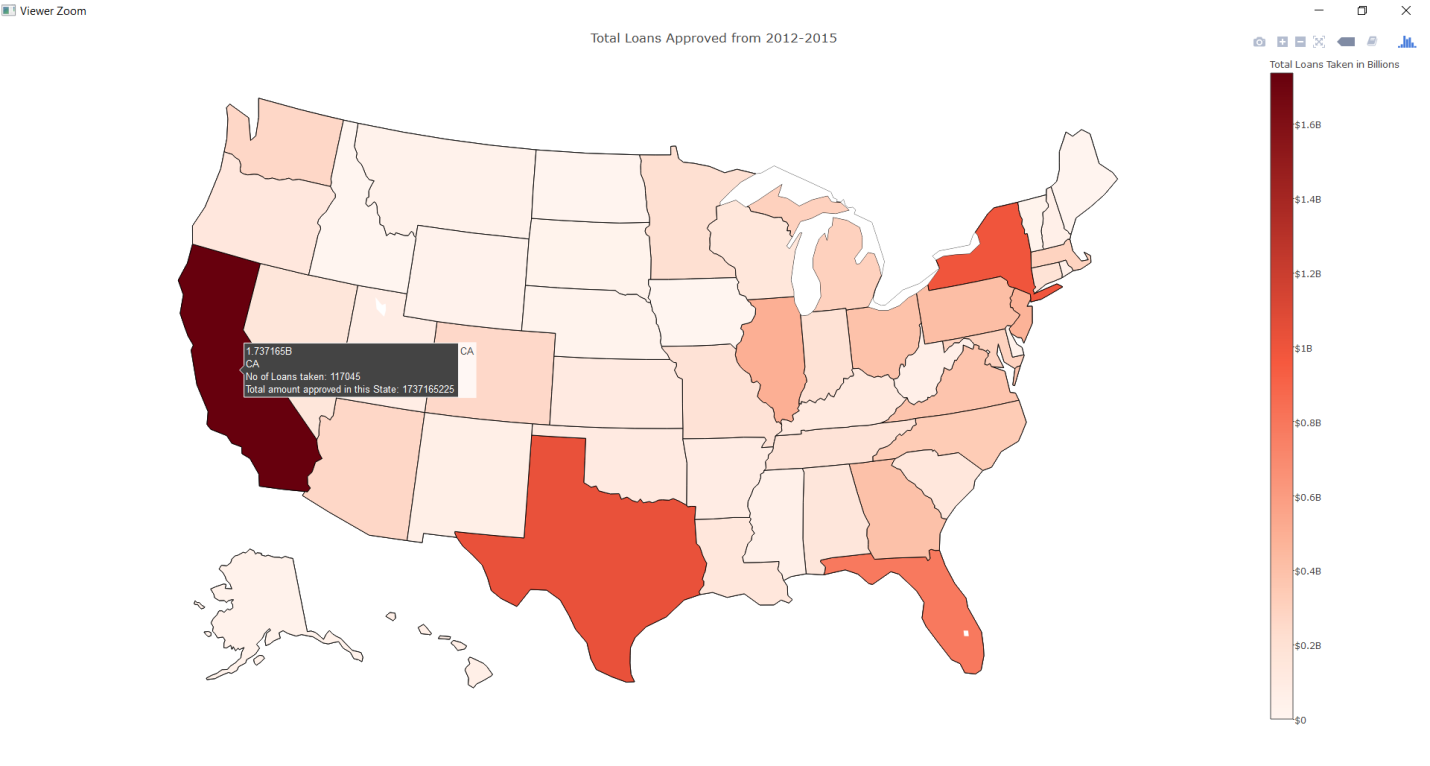
1. **Data Visualizations:**
2. **Total no. of loans issued from 2012 to 2015?**



(Highlights from R script – zoo, dplyr, plotly packages, unite, merge, line chart)

This Time series chart shows the monthly loans issued from 2012 to 2015. Here we can see a clear increasing trend in the no. of loans every year from 2602 in Jan 2012 to 44343 in Dec 2015. Also, looking at it we found seasonality in the graph which shows that there is a high demand of loans in the months of July, October and a Low demand in the month of September every year. Here, the spike in Loans in the month of July can be seen as the US financial year ends in the month of September and people are eligible for Tax exemption on Loans taken; Hence, in order to avail this benefit many people go for Loans in the month of July. Here, we suggest the Loan providers to run special benefit schemes in the months of July to October in order to capitalize on the demand.

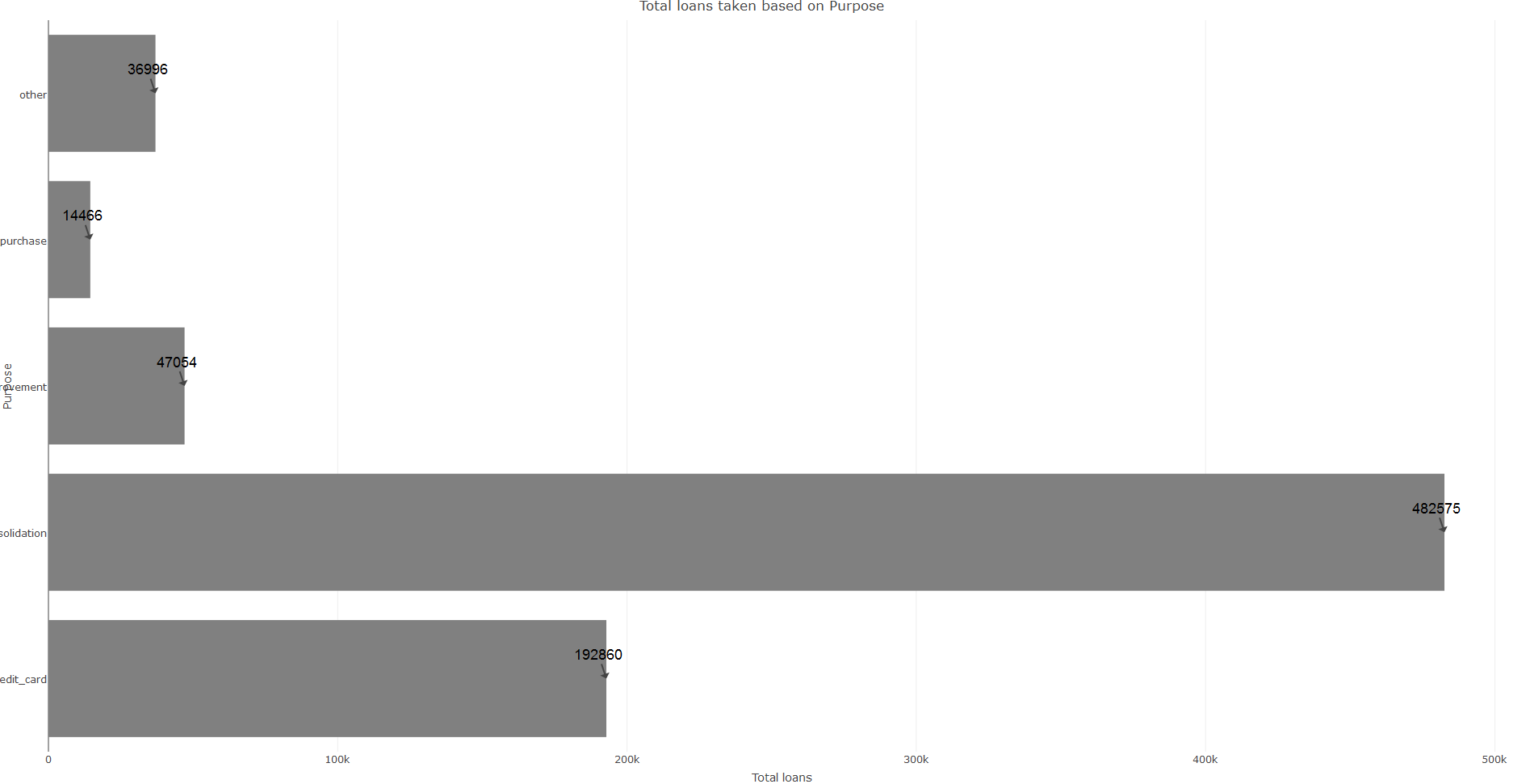
1. **Which are the states with maximum loans taken and total loan amount approved?**



(Highlights from R script – plotly, dplyr packages, sum, merge, geographic map)

The above Geographical map shows the states where maximum loans have been approved and taken from 2012 to 2015. It can be clearly seen that the State of California ranks the first in terms of Total Loans approved with 117045 loans and an amount of $1.73 Billion followed by Texas, New York and Florida. Hence, we suggest the Loan companies to Target these states first as the demand for Loans are more in these states and they should run their marketing campaigns accordingly.

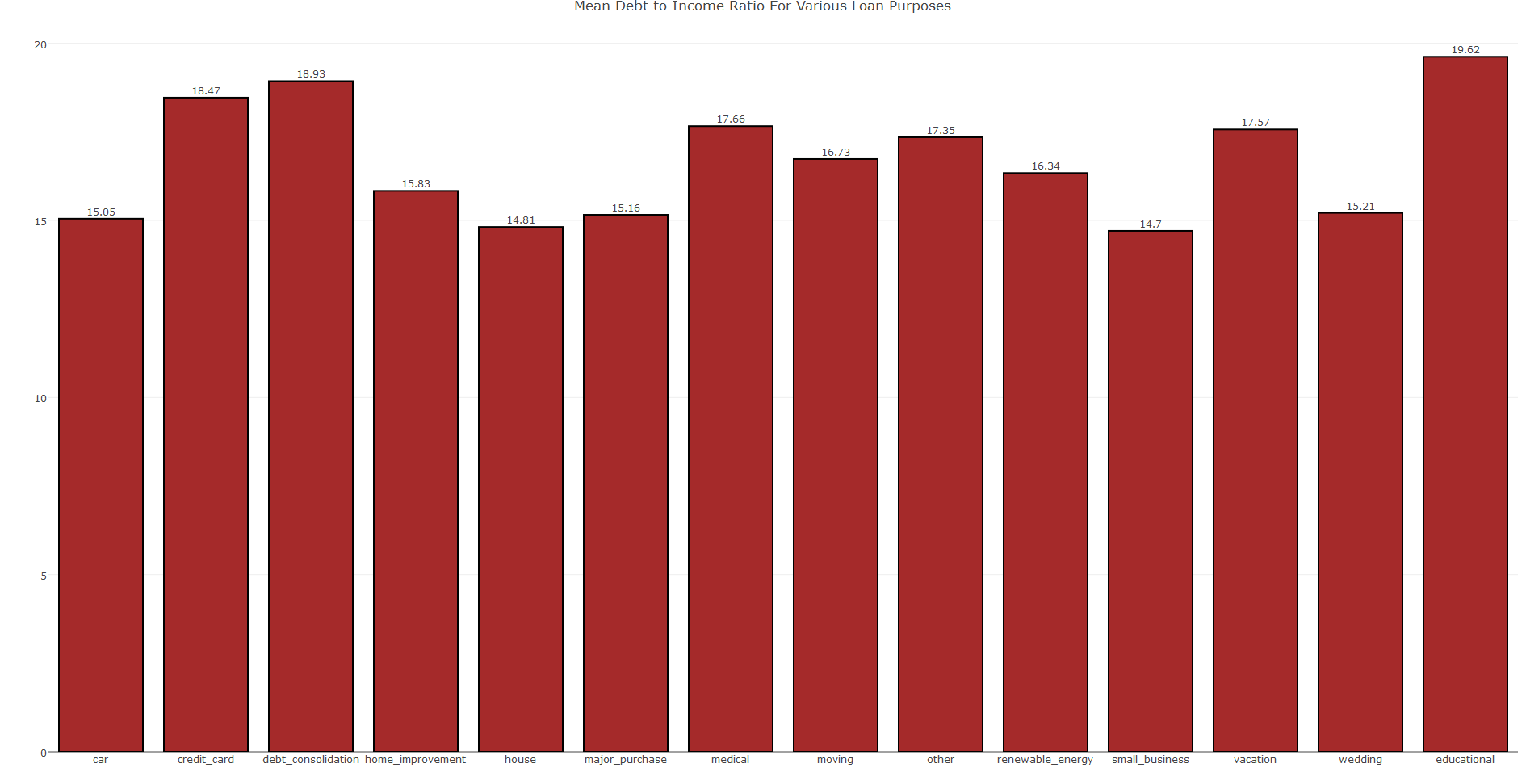
1. **What is the most common purpose people applied for loan in each year?**



(Highlights from R script - dplyr, plotly packages, horizontal bar chart )

The above Horizontal Bar chart shows the main purposes for which people take loans. From the visual it can be seen that Debt consolidation is the major reason why people take loans followed by credit card and home improvement. Hence, it is clear that majority opt for loans to pay off other loans and their credit card bills. Here, looking at this statistics we suggest the people to manage their expenses and finance in a better manner in order to avoid such situations which require refinancing. On the other hand, this analysis is helpful for loan providers as it should check the debt consolidation loans before approving in order to prevent them from increasing their Non performing Assets (NPA’s).

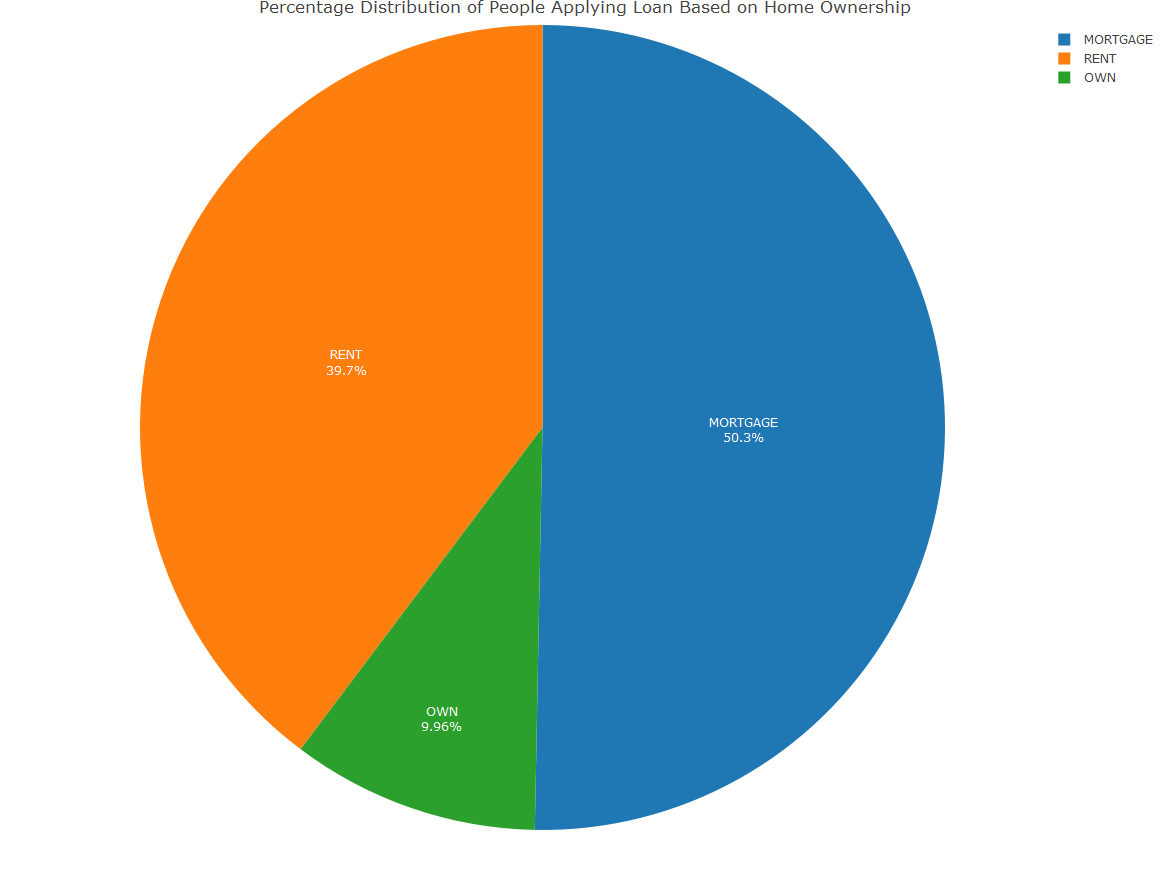
1. **What is the mean Debt to income ratio of people taking loans for all the purposes?**



(Highlights from R script – dplyr, plotly, mean, bar chart)

From the above Bar chart it can be seen that the mean Debt to income ratio is highest for Educational loans, followed by Debt consolidation and Credit card. Hence we can see that for Educational loans people generally have lower incomes, hence we suggest the Loan providers to offer Educational loans at a lower interest rate in order to help the needy. Also, it is seen that the mean Debt to income ratio for Small businesses is the least, which can be used as a plus by the loan companies and charge a higher interest rate.

1. **What is the percentage distribution of people taking loans according to their home ownership?**



(Highlights from R script – dplyr, plotly, pie chart)

The above pie chart shows the home ownership status of people taking loans. It can be seen that 50.3% of the people taking loans have their home as Mortgage, followed by Rent at 39.7% and Owned at 9.96%. Hence, it is clear from this that people who own a home generally do not opt for loans, as compared to the ones which are on Rent or Mortgage. By looking at these statistics we suggest the Loan providing companies to target people staying on Mortgage or Rent as these are more likely to go for Loans in comparison to the ones who own a home.

🡺**R Code for all the Analysis and Visualizations:**

setwd("F:/R project/R project")

install.packages('tidyr')

library(tidyr)

data12\_13<-read.csv("12\_13\_loan\_final.csv",header=T,sep=",")

data\_14<-read.csv("14\_loan\_final.csv",header=T,sep=",")

data\_15<-read.csv("15\_loan\_final.csv",header=T,sep=",")

data12\_15<-rbind(data12\_13,data\_14,data\_15)

data\_12\_15\_clean1<-separate(data12\_15,issue\_d,c("issue\_Month","issue\_Year"),sep="-")

data\_12\_15\_clean2<-data\_12\_15\_clean1[,1:26]

data\_cleaned<-data\_12\_15\_clean2

View(data\_cleaned)

**#Line chart for Total Loans from 2012 to 2015**

library(zoo)

united\_data<-unite(data\_cleaned,CombinedDate, issue\_Year, issue\_Month, sep="-")

df\_unite<-data.frame(united\_data)

yearly\_grp<-group\_by(df\_unite,CombinedDate) %>% summarise(count = n())

df\_yr<-data.frame(yearly\_grp)

Years<-as.yearmon(yearly\_grp$CombinedDate,"%y-%b")

df\_yr$id<- 1:nrow(df\_yr)

df\_yr1<-data.frame(Years)

df\_yr1$id<-1:nrow(df\_yr1)

merged\_date<-merge(df\_yr,df\_yr1,by='id')

sorted<-merged\_date[order(merged\_date$Years),]

sorted\_final<-sorted[,3:4]

library(plotly)

l<-plot\_ly(sorted\_final,x=~Years,y=~count,type='scatter',mode='lines') %>% layout(title = 'Total Loans taken over years', xaxis = list(autotick = F, dtick = 1))

print(l)

**# States with maximum loans approved on Geographical Map**

df\_3<-data.frame(data\_cleaned)

library(plotly)

grp3<-group\_by(df\_3,addr\_state) %>% summarise(cnt = n())

grp4<-group\_by(df\_3,addr\_state) %>% summarise(Total\_amount\_approved = sum(loan\_amnt))

merged<-merge(grp3,grp4,by="addr\_state")

merged\_ordr<-merged[with(merged,order(merged[,2], decreasing = T)),]

merged\_ordr$lst<-with(merged\_ordr,paste(addr\_state,"<br>","No of Loans taken:",cnt,"<br>","Total amount approved in this State:",Total\_amount\_approved))

map<- list(scope = 'usa',projection = list(type = 'albers usa'),showlakes = TRUE,lakecolor = toRGB('white'))

m<-plot\_geo(merged\_ordr,locationmode='USA-states') %>% add\_trace(z=~Total\_amount\_approved,text=~lst,locations=~addr\_state,color=~cnt,colors='Reds') %>% colorbar(title = "Total Loans Taken in Billions",tickprefix = '$') %>% layout(title = 'Total Loans Approved from 2012-2015',geo = map)

print(m)

**#Horizontal Bar plot for Loans according to Purpose**

library(dplyr)

df\_1<-data.frame(data\_cleaned)

pur\_year<-group\_by(df\_1,purpose) %>% summarise(count = n())

sorted1<-pur\_year[order(pur\_year$count,decreasing = T),]

top5\_sorted<-head(sorted1,5)

library(plotly)

hb<-plot\_ly(x=top5\_sorted$count,y=top5\_sorted$purpose,type='bar',orientation='h') %>% layout(title = "Total loans taken based on Purpose",xaxis = list(title = "Total loans"),yaxis = list(title = "Purpose"),annotations = list(x=top5\_sorted$count,y=top5\_sorted$purpose,text=top5\_sorted$count,showarrow = FALSE))

**#Bar chart of Mean Debt to income ratio of people taking loans according to purpose**

library(dplyr)

df\_2<-data.frame(data\_cleaned)

grp2<-group\_by(df\_2,purpose) %>% summarise(Mean\_DTI = mean(dti))

library(plotly)

b<-plot\_ly(grp2,x=grp2$purpose,y=grp2$Mean\_DTI,type="bar",text=grp2$purpose,marker = list(color = 'rgb(165,42,42)',line = list(color = 'rgb(0,0,0)', width = 2))) %>% layout(title = "Mean Debt to Income Ratio For Various Loan Purposes",xaxis = list(title = ""),yaxis = list(title = ""),annotations = list(x=grp2$purpose,y=grp2$Mean\_DTI,text=round(grp2$Mean\_DTI,digits=2),xanchor = 'center', yanchor = 'bottom',showarrow = FALSE))

print(b)

**#Pie chart for Loan according to Home Ownership**

library(dplyr)

library(plotly)

ho\_ow<-group\_by(data\_cleaned,home\_ownership) %>% summarise(count =n())

df\_5<-data.frame(ho\_ow)

ordr\_df5<-df\_5[with(df\_5,order(df\_5[,2], decreasing = T)),]

top\_df5<-head(ordr\_df5,3)

pie<-plot\_ly(top\_df5,labels=top\_df5$home\_ownership,values=top\_df5$count,type='pie',textposition = 'inside',textinfo = 'label+percent',insidetextfont = list(color = '#FFFFFF'),pull = list(color = '#FFFFFF', width = 3))

p<-layout(pie,title = 'Percentage Distribution of People Applying Loan Based on Home Ownership', xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE), yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE))

print(p)



**References:**

1. <https://www.lendingclub.com/info/download-data.action>
2. <https://plot.ly/r/getting-started/>
3. <https://commonbond.co/blog/what-is-marketplace-lending/>