FinalAssignment

Amol Gote

3/7/2020

library(ggplot2)   
library(readr)   
library(gridExtra)  
library(tidyverse)  
library(sf)

## Warning: package 'sf' was built under R version 3.6.3

library(maps)

## Warning: package 'maps' was built under R version 3.6.3

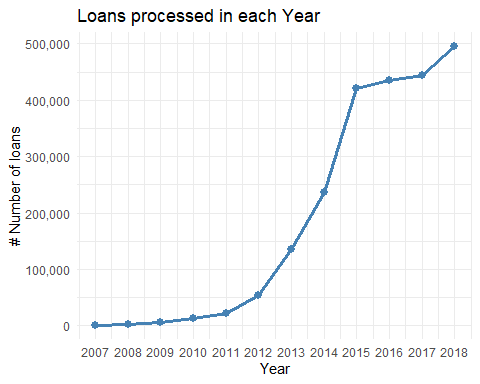
## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

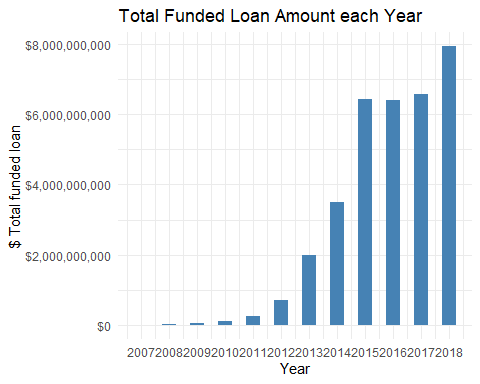
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

lendingClubLoanData <- read.csv("data/lending\_club\_loan\_data.csv")  
lendingClubLoanData <- lendingClubLoanData[, !(colnames(lendingClubLoanData) %in% c("id","member\_id", "url", "desc"))]  
lendingClubLoanData <- lendingClubLoanData[, !(colnames(lendingClubLoanData) %in% c("open\_acc\_6m", "open\_act\_il", "open\_il\_12m", "open\_il\_24m", "mths\_since\_rcnt\_il", "total\_bal\_il", "il\_util", "open\_rv\_12m", "open\_rv\_24m", "max\_bal\_bc", "all\_util", "total\_rev\_hi\_lim", "inq\_fi", "total\_cu\_tl", "acc\_open\_past\_24mths", "bc\_open\_to\_buy", "bc\_util", "mo\_sin\_old\_il\_acct", "mo\_sin\_old\_rev\_tl\_op", "mo\_sin\_rcnt\_rev\_tl\_op", "mo\_sin\_rcnt\_tl", "mths\_since\_recent\_bc", "mths\_since\_recent\_bc\_dlq", "mths\_since\_recent\_inq", "mths\_since\_recent\_revol\_delinq", "num\_accts\_ever\_120\_pd", "num\_actv\_bc\_tl", "num\_actv\_rev\_tl", "num\_bc\_sats", "num\_bc\_tl", "num\_il\_tl", "num\_op\_rev\_tl", "num\_rev\_accts", "num\_rev\_tl\_bal\_gt\_0", "num\_sats", "pct\_tl\_nvr\_dlq", "percent\_bc\_gt\_75", "tot\_hi\_cred\_lim", "total\_bal\_ex\_mort", "total\_bc\_limit", "total\_il\_high\_credit\_limit", "revol\_bal\_joint", "sec\_app\_earliest\_cr\_line", "sec\_app\_inq\_last\_6mths", "sec\_app\_mort\_acc", "sec\_app\_open\_acc", "sec\_app\_revol\_util", "sec\_app\_open\_act\_il", "sec\_app\_num\_rev\_accts", "sec\_app\_chargeoff\_within\_12\_mths", "sec\_app\_collections\_12\_mths\_ex\_med", "sec\_app\_mths\_since\_last\_major\_derog", "hardship\_reason", "hardship\_status", "deferral\_term", "hardship\_amount", "hardship\_start\_date", "hardship\_end\_date", "payment\_plan\_start\_date", "hardship\_length", "hardship\_dpd", "hardship\_loan\_status", "orig\_projected\_additional\_accrued\_interest", "hardship\_payoff\_balance\_amount", "hardship\_last\_payment\_amount", "disbursement\_method", "debt\_settlement\_flag", "debt\_settlement\_flag\_date", "settlement\_status", "settlement\_date", "settlement\_amount", "settlement\_percentage", "settlement\_term"))]  
lendingClubLoanData <- lendingClubLoanData[, !(colnames(lendingClubLoanData) %in% c("earliest\_cr\_line","inq\_last\_6mths","mths\_since\_last\_delinq","mths\_since\_last\_record","open\_acc","pub\_rec","revol\_bal","revol\_util","total\_acc","initial\_list\_status","out\_prncp","out\_prncp\_inv","total\_pymnt","total\_pymnt\_inv","total\_rec\_prncp","total\_rec\_int","total\_rec\_late\_fee","recoveries","collection\_recovery\_fee","last\_pymnt\_d","last\_pymnt\_amnt","next\_pymnt\_d","last\_credit\_pull\_d","collections\_12\_mths\_ex\_med","mths\_since\_last\_major\_derog","policy\_code","acc\_now\_delinq","tot\_coll\_amt","tot\_cur\_bal","inq\_last\_12m","avg\_cur\_bal","chargeoff\_within\_12\_mths","delinq\_amnt","mort\_acc","num\_tl\_120dpd\_2m","num\_tl\_30dpd","num\_tl\_90g\_dpd\_24m","num\_tl\_op\_past\_12m"))]  
lendingClubLoanData$orig\_year<-substr(lendingClubLoanData$issue\_d,5,8)  
write.csv(lendingClubLoanData, file = "data/lending\_club\_loan\_data\_final.csv", row.names=FALSE)  
lendingClubLoanData <- read.csv("data/lending\_club\_loan\_data\_final.csv")

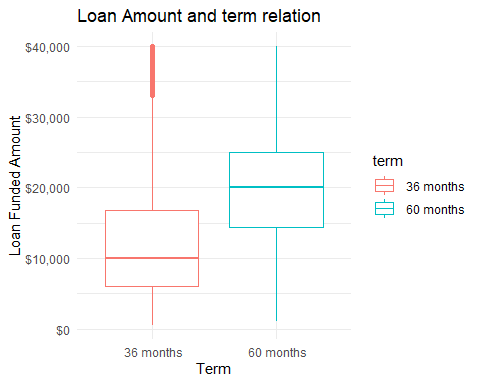
numberOfLoansByYear <- lendingClubLoanData %>%   
 group\_by(orig\_year)%>%  
 summarise(loanCountByYear=n())  
  
ggplot(data = numberOfLoansByYear)+  
 geom\_line(color="steelblue", size=1.2, aes(x=orig\_year,y=loanCountByYear))+  
 geom\_point(color="steelblue", size=2.5, aes(x=orig\_year,y=loanCountByYear))+  
 scale\_x\_continuous(breaks = numberOfLoansByYear$orig\_year) +  
 scale\_y\_continuous(labels = scales::comma\_format()) +  
 labs(x="Year",y="# Number of loans",title="Loans processed in each Year")+  
 theme\_minimal()



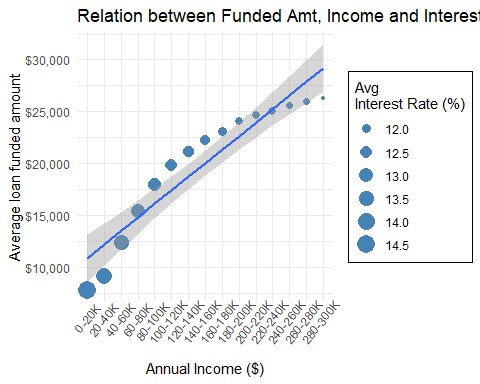
totalFundedAMountPerYear <- lendingClubLoanData %>%   
 group\_by(orig\_year)%>%  
 summarise(totalFundedAmount= sum(as.numeric(funded\_amnt)))  
  
ggplot(data = totalFundedAMountPerYear, aes(x=orig\_year, y=totalFundedAmount)) +  
 geom\_bar(stat="identity", width=0.5, fill = "steelblue") +  
 scale\_x\_continuous(breaks = numberOfLoansByYear$orig\_year) +  
 scale\_y\_continuous(labels = scales::dollar) +  
 labs(x = "Year", y = "$ Total funded loan",title="Total Funded Loan Amount each Year") +  
 theme\_minimal()



lendingClubLoanData %>%  
ggplot() +  
 geom\_boxplot(aes(x=term, y=funded\_amnt, color=term)) +   
 scale\_y\_continuous(labels = scales::dollar) +  
 labs(x = "Term", y = "Loan Funded Amount", title="Loan Amount and term relation") +  
 theme\_minimal()



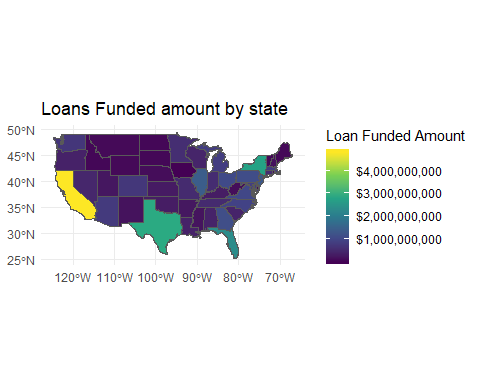
filteredLendingClubData <- lendingClubLoanData %>%  
 drop\_na(annual\_inc)  
  
filteredLendingClubData <- filteredLendingClubData %>%  
 filter(annual\_inc <= 300000)  
  
lbls <- c('0-20K','20-40K','40-60K','60-80K','80-100K','100-120K','120-140K','140-160K','160-180K','180-200K', '200-220K', '220-240K', '240-260K', '260-280K', '280-300K')  
groupedData <- filteredLendingClubData %>%   
 group\_by(incomeGroup = cut(annual\_inc, breaks= seq(0, 300000, by = 20000), right = TRUE, include.lowest = TRUE, labels = lbls) ) %>%   
 summarise(averageInterest= mean(int\_rate), averageLoanLoanFundedAmount = mean(funded\_amnt))   
  
ggplot(data =groupedData, aes(x=incomeGroup, y=averageLoanLoanFundedAmount)) +  
 geom\_point(colour="steelblue", shape=16, aes(size=averageInterest)) +  
 geom\_smooth(aes(incomeGroup, averageLoanLoanFundedAmount, group = 1), method = "lm") +  
 scale\_y\_continuous(labels = scales::dollar) +  
 labs(x="Annual Income ($)",y="Average loan funded amount",title="Relation between Funded Amt, Income and Interest Rate")+  
 guides(size=guide\_legend("Avg \nInterest Rate (%)")) +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle =50, hjust=0.75))+  
 theme(legend.background = element\_rect())



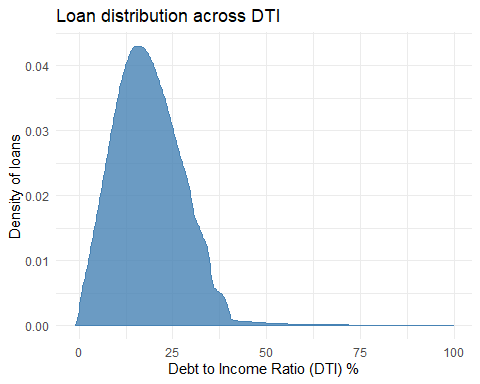
fullStateNames <- read.csv("data/states.csv")  
states <- st\_as\_sf(map("state", plot = FALSE, fill = TRUE))  
  
fundedAmountByState <- lendingClubLoanData %>%   
 group\_by(addr\_state)%>%  
 summarise(totalFundedAmount= sum(as.numeric(funded\_amnt)))  
  
fundedAmountByState <- fundedAmountByState %>%  
 inner\_join(fullStateNames, by = c("addr\_state" = "abbreviation"))  
  
  
states2 <- states %>% left\_join(fundedAmountByState, by = c("ID" = "state" ))

## Warning: Column `ID`/`state` joining factors with different levels,  
## coercing to character vector

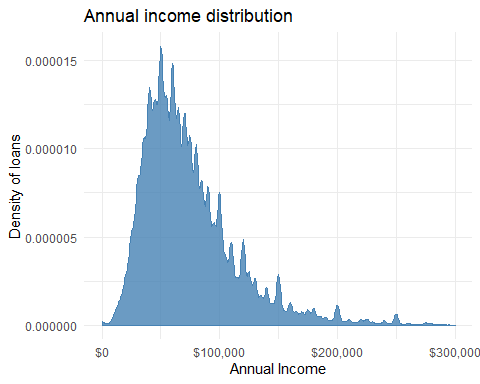
ggplot(data = states2) +   
 geom\_sf(aes(fill = totalFundedAmount)) +  
 scale\_fill\_viridis\_c("Loan Funded Amount", labels = scales::dollar) +  
 labs(title = "Loans Funded amount by state") +  
 theme\_minimal()



filteredLendingClubData <- lendingClubLoanData %>%  
 drop\_na(dti)%>%  
 filter(dti < 100)  
  
ggplot(data = filteredLendingClubData, aes(x = dti)) +  
 geom\_density(fill="steelblue", color="steelblue", alpha=0.8) +  
 labs(x="Debt to Income Ratio (DTI) %",y="Density of loans",title="Loan distribution across DTI") +  
 theme\_minimal()



filteredLendingClubData <- lendingClubLoanData %>%  
 drop\_na(annual\_inc)%>%  
 filter(annual\_inc < 300000)  
  
ggplot(data = filteredLendingClubData, aes(x = annual\_inc)) +  
 geom\_density(fill="steelblue", color="steelblue", alpha=0.8) +  
 scale\_x\_continuous(labels = scales::dollar) +  
 scale\_y\_continuous(labels = function(x) format(x, scientific = FALSE)) +  
 labs(x="Annual Income",y="Density of loans",title="Annual income distribution") +  
 theme\_minimal()



filteredLendingClubData <- lendingClubLoanData %>%  
 drop\_na(loan\_amnt)  
as.numeric(max(filteredLendingClubData$loan\_amnt))

## [1] 40000

as.numeric(min(filteredLendingClubData$loan\_amnt))

## [1] 500

loanGreaterThan100K <- filteredLendingClubData %>%  
 filter(loan\_amnt > 100000)  
  
uniqueLoanStatus <- unique(filteredLendingClubData$loan\_status)  
uniqueLoanStatus

## [1] Current   
## [2] Fully Paid   
## [3] Late (31-120 days)   
## [4] In Grace Period   
## [5] Charged Off   
## [6] Late (16-30 days)   
## [7] Default   
## [8] Does not meet the credit policy. Status:Fully Paid   
## [9] Does not meet the credit policy. Status:Charged Off  
## 9 Levels: Charged Off Current ... Late (31-120 days)

filteredLendingClubData <- lendingClubLoanData %>%  
 drop\_na(dti)  
as.numeric(max(filteredLendingClubData$dti))

## [1] 999

as.numeric(min(filteredLendingClubData$dti))

## [1] -1