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

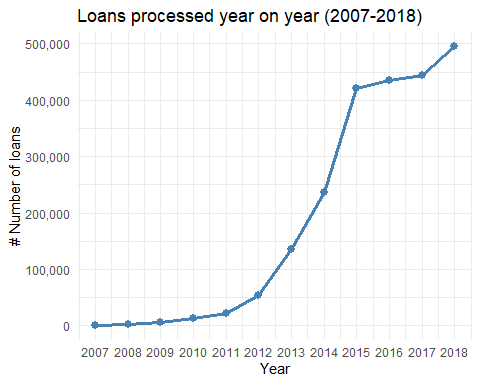
library(maps)

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

lendingClubLoanData <- read.csv("data/lending\_club\_loan\_data.csv")  
  
# Remove not required columns from the dataset, so as to minimize the dataset size.  
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 the dataset to csv file for analysis  
write.csv(lendingClubLoanData, file = "data/lending\_club\_loan\_data\_final.csv", row.names=FALSE)  
  
lendingClubLoanData <- read.csv("data/lending\_club\_loan\_data\_final.csv")

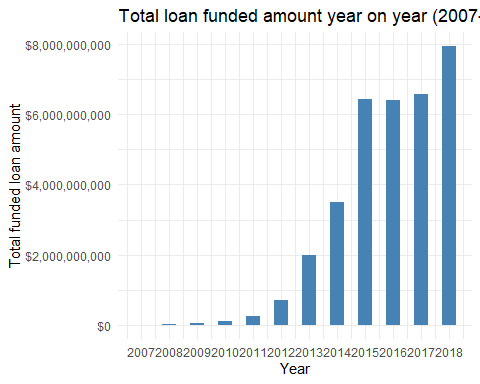
1. Below visualization shows number of loans funded each year from 2007 till 2018.
2. From 2007 till 2012, it had gradual progress in number of loans issued,
3. 2012 to 2015 number of loans issued has grown exponentially.
4. 2015 to 2017 saw a marginal hike in number of loans, 2017 to 2018 saw a considerable hike.
5. Number of loans issued in 2018 by lending club is close to 500,000 which is highest

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 year on year (2007-2018)")+  
 theme\_minimal()



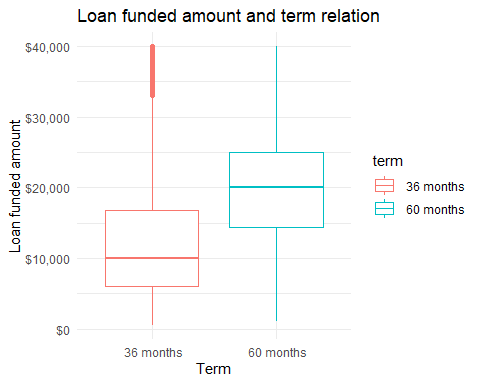
1. Below visualization shows total loan funded amount for each year from 2007 till 2018.
2. In general trend of total loan funded amount is in align to the number of loans issued as above visulization.
3. Lending club has funded loans maximum of $8 billion in 2018.
4. 2015 to 2017, total funded loan amount is flat.

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 amount",  
 title="Total loan funded amount year on year (2007-2018)") +  
 theme\_minimal()



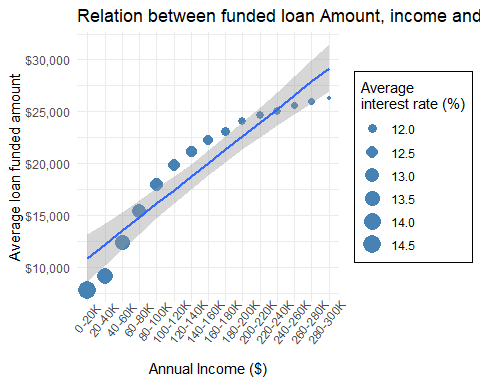
1. Below visualization shows relationship between loan funded amount and term.
2. Lending club offers only loans with 2 terms
   1. 36 Months
   2. 60 Months
3. For 36 months loan:
   1. Median loan funded amount is $10,000.
   2. Majority of the funded loan amount ranges from close $6000 to $16,000.
4. For 60 months loan:
   1. Median loan funded amount is $20000.
   2. Majority of the funded loan amount ranges from close $15,000 to $25,000.

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 funded amount and term relation") +  
 theme\_minimal()



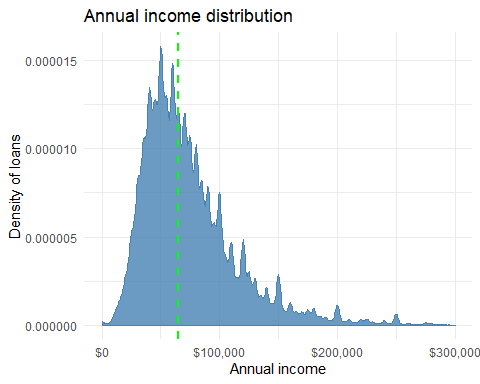
1. Below visulization shows relationship between funded loan amount, annual income and interest rate.
2. Visualization shows, that higher the annual income more is the funded loan amount.
3. As the annual income increase the interest rate drop by couple of percentage points.

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 loan Amount, income and interest rate")+  
 guides(size=guide\_legend("Average \ninterest rate (%)")) +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle =50, hjust=0.75))+  
 theme(legend.background = element\_rect())



1. Below visualization shows the distribution of annual income of people getting loans.
2. Majority of people are having annual income less than $200K.
3. Median annual income seems to around $60K.
4. Have excluded loans taken by people having annual income > 300K, as they are few in number and are outliars.

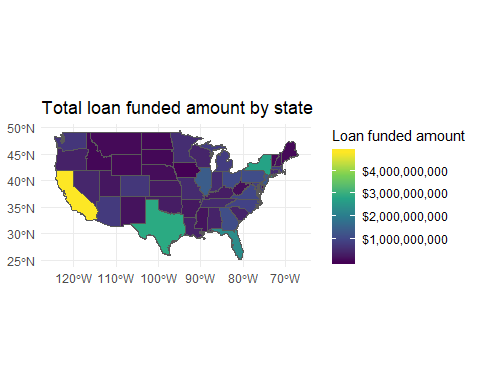
filteredLendingClubData <- lendingClubLoanData %>%  
 drop\_na(annual\_inc)%>%  
 filter(annual\_inc < 300000)  
  
ggplot(data = filteredLendingClubData, aes(x = annual\_inc)) +  
 geom\_density(fill="steelblue", color="steelblue", alpha=0.8) +  
 geom\_vline(aes(xintercept=median(annual\_inc)),color="green", linetype="dashed", size=1) +  
 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()

 1. Below visualization shows geographical distribution of total loan funded amount across various US states. 2. California has the highest total funded loan amount, followed by Texas, New York and Florida.

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" ))

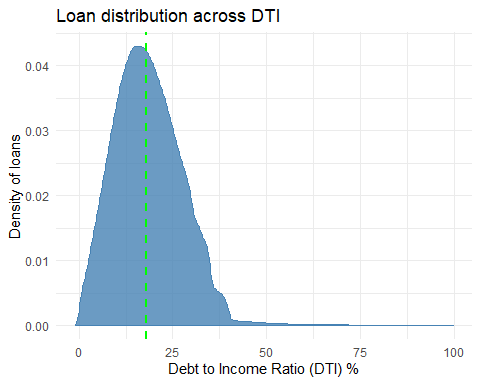
## Warning: Column `ID`/`state` joining factors with different levels,  
## coercing to character vector

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



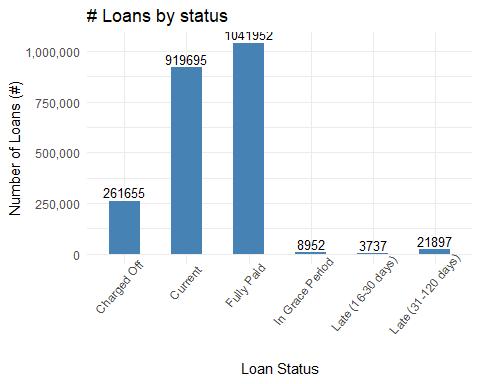
1. Below visualization shows the distributiion of DTI (Debt to Income Ratio (%)) for the people getting loan.
2. Lower the DTI, higher the probability of getting the loan approved.

filteredLendingClubData <- lendingClubLoanData %>%  
 drop\_na(dti)%>%  
 filter(dti < 100)  
  
ggplot(data = filteredLendingClubData, aes(x = dti)) +  
 geom\_density(fill="steelblue", color="steelblue", alpha=0.8) +  
 geom\_vline(aes(xintercept=median(dti)),color="green", linetype="dashed", size=1) +  
 labs(x="Debt to Income Ratio (DTI) %",y="Density of loans",title="Loan distribution across DTI") +  
 theme\_minimal()



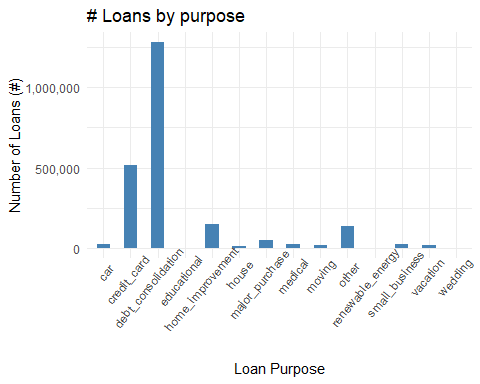
1. Below visulization shows loan distribution by the loan status.
2. Close to 900K loans are current and 1 Million loans are paid off.
3. Charged of loans are close to 2,61,655 out of total 2,260,668 loans. SO the charge of percentage is close to 12% (exact 11.57 %).
4. Most current loans will be eventually paid off.

loan\_statuses <- c("Current",  
 "Fully Paid",  
 "Late (31-120 days)",  
 "In Grace Period",  
 "Charged Off",  
 "Late (16-30 days)")  
  
numberOfLoansByLoanStatus <- lendingClubLoanData %>%   
 filter(loan\_status %in% loan\_statuses) %>%  
 group\_by(loan\_status)%>%  
 summarise(numberOfLoans = n())  
  
ggplot(data = numberOfLoansByLoanStatus, aes(x=loan\_status, y=numberOfLoans)) +  
 geom\_bar(stat="identity", width=0.5, fill = "steelblue") +  
 geom\_text(aes(label=numberOfLoans), vjust=-0.3, size=3.5) +  
 scale\_y\_continuous(labels = scales::comma\_format()) +  
 labs(x = "Loan Status", y = "Number of Loans (#)",title="# Loans by status") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle =50, hjust=0.75))



1. Below visualization shows the distribution of loans taken for various causes.
2. Debt consolidation is major primary purpose of taking loan from lending club.
3. Second major purpose of taking loan is for paying credit card bills.

numberOfLoansByPurpose <- lendingClubLoanData %>%   
 group\_by(purpose)%>%  
 summarise(numberOfLoans = n())  
  
ggplot(data = numberOfLoansByPurpose, aes(x=purpose, y=numberOfLoans)) +  
 geom\_bar(stat="identity", width=0.5, fill = "steelblue") +  
 scale\_y\_continuous(labels = scales::comma\_format()) +  
 labs(x = "Loan Purpose", y = "Number of Loans (#)",title="# Loans by purpose") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle =50, hjust=0.75))



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

## [1] 40000

# Minimum loan amount  
as.numeric(min(filteredLendingClubData$loan\_amnt))

## [1] 500

# Unique loan statuses   
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)  
# Max DTI  
as.numeric(max(filteredLendingClubData$dti))

## [1] 999

#Minimum DTI  
as.numeric(min(filteredLendingClubData$dti))

## [1] -1

# Total loan funded amount  
as.numeric(sum(filteredLendingClubData$funded\_amnt))

## [1] 33971525425

filteredLendingClubData <- lendingClubLoanData %>%  
 drop\_na(annual\_inc)%>%  
 filter(annual\_inc < 300000)  
# Annual income median  
median(filteredLendingClubData$annual\_inc)

## [1] 65000

filteredLendingClubData <- lendingClubLoanData %>%  
 drop\_na(dti)%>%  
 filter(dti < 100)  
# Dti median  
median(filteredLendingClubData$dti)

## [1] 17.82

#Number of loans  
nrow(lendingClubLoanData)

## [1] 2260668