Introduction of the dataset

Dataset belongs to company called Lending Club, it contains all the loans which have been issued by Lending Club from 2007 to 2018.

The Lending Club is a peer-to-peer lending service (it lends money to customers by matching lenders to borrowers), based in the United States. This company enables borrowers to create loan listings on its website by supplying details about themselves and the loans that they would like to request. On the basis of the borrower's credit score, credit history, desired loan amount and the borrower's debt-to-income ratio, Lending Club determines whether the borrower is credit worthy and assigns to its approved loans a credit grade that determines payable interest rate and fees.

Dataset contains close to 2 million loan records, precise count of records = 2,260,668

Dataset contains various attributes based on which the credit lending decision takes place, this includes

1. Deb to income ratio
2. Annual Income
3. Home ownership type (Rented, Mortgage, Owned)
4. Employment details.

Dataset also contains geographical information like

1. State from which loan has originated.
2. Zip code.

It also has issued loan details like

1. Funded Loan Amount
2. Issued month and year.
3. Loan Term (36 or 60 months)
4. Interest rate.

Post loan has been issued, loan performance needs to be tracked, and this dataset also contains those details

1. Loan Status (Current, Paid Off, Charged Off, Delinquent)
2. Loan Grades.
3. Delinquency details.

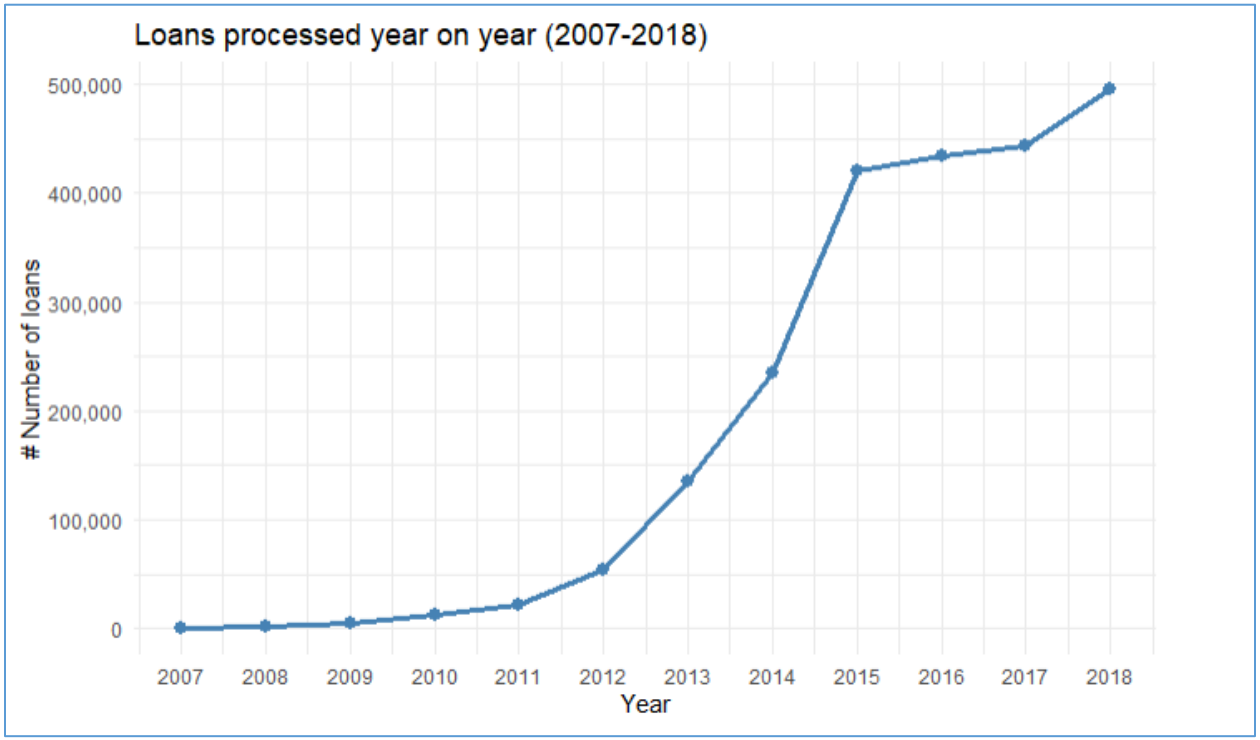
Data Analysis objective

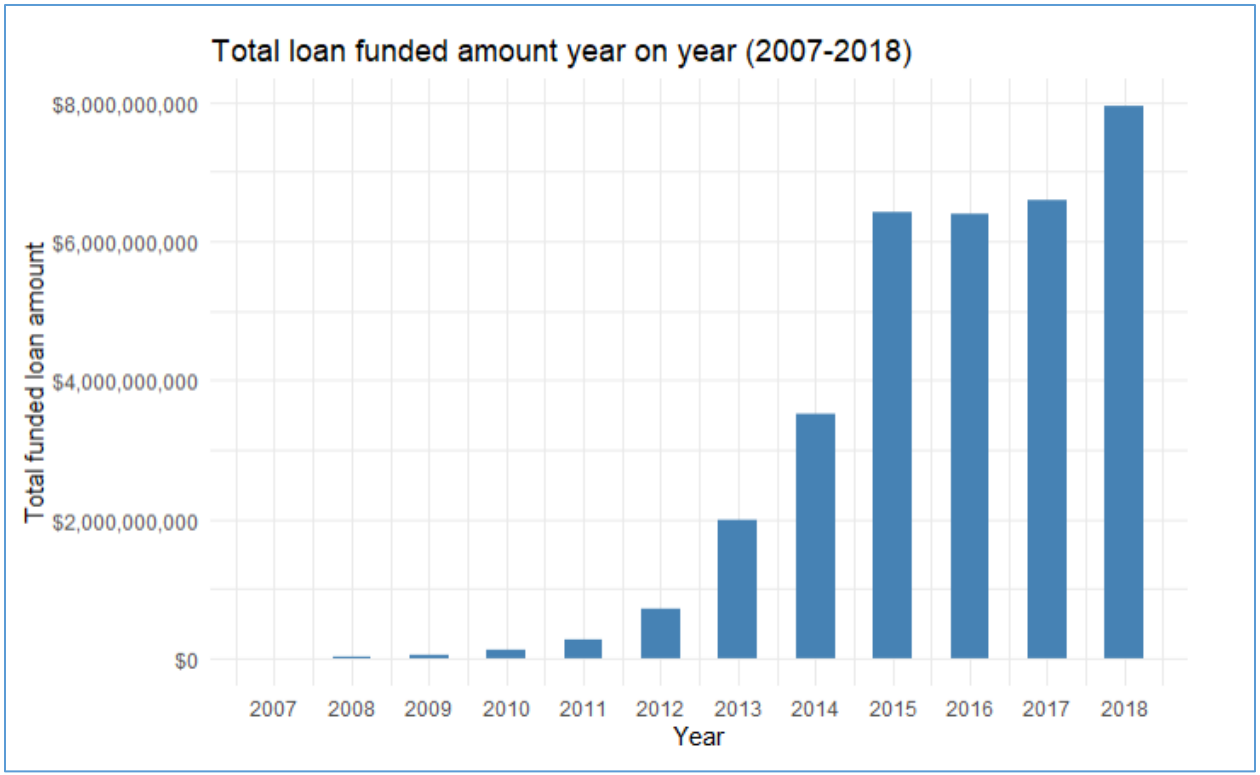
- How lending club has performed in terms of number of loans and funded loan amount?
- Identify details around the loan term and funded loan amount, typically bigger loans are associated with longer term.
- Typically lenders favor applicants with lower DTIs, analyze the dataset to check if this holds good. Identify the sweet spot number for DTI.
- Identify relationship between funded amount, annual income and interest rate.
- Identify geographically how loans are funded.

Visualizations

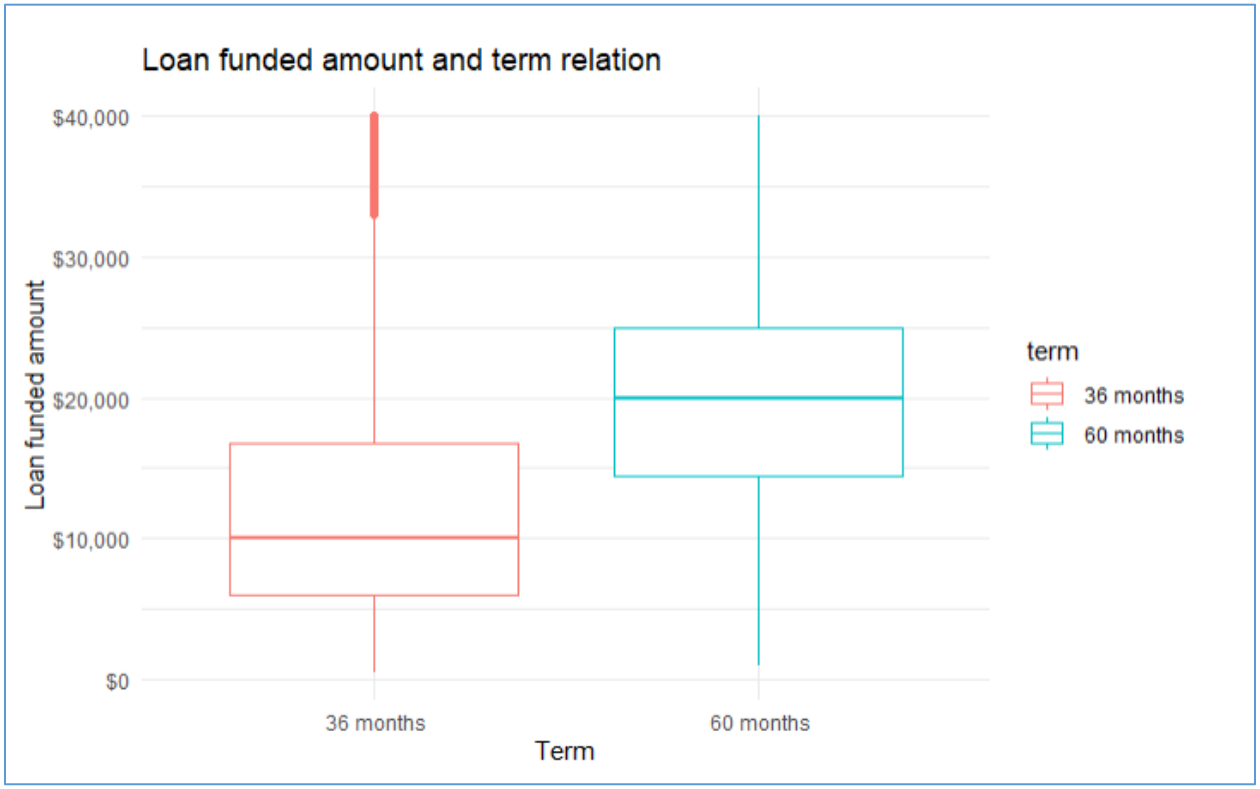
Have just pasted the visualization to provide unified view, there is separate section which contains the knitted markdown.

Year wise loan trends

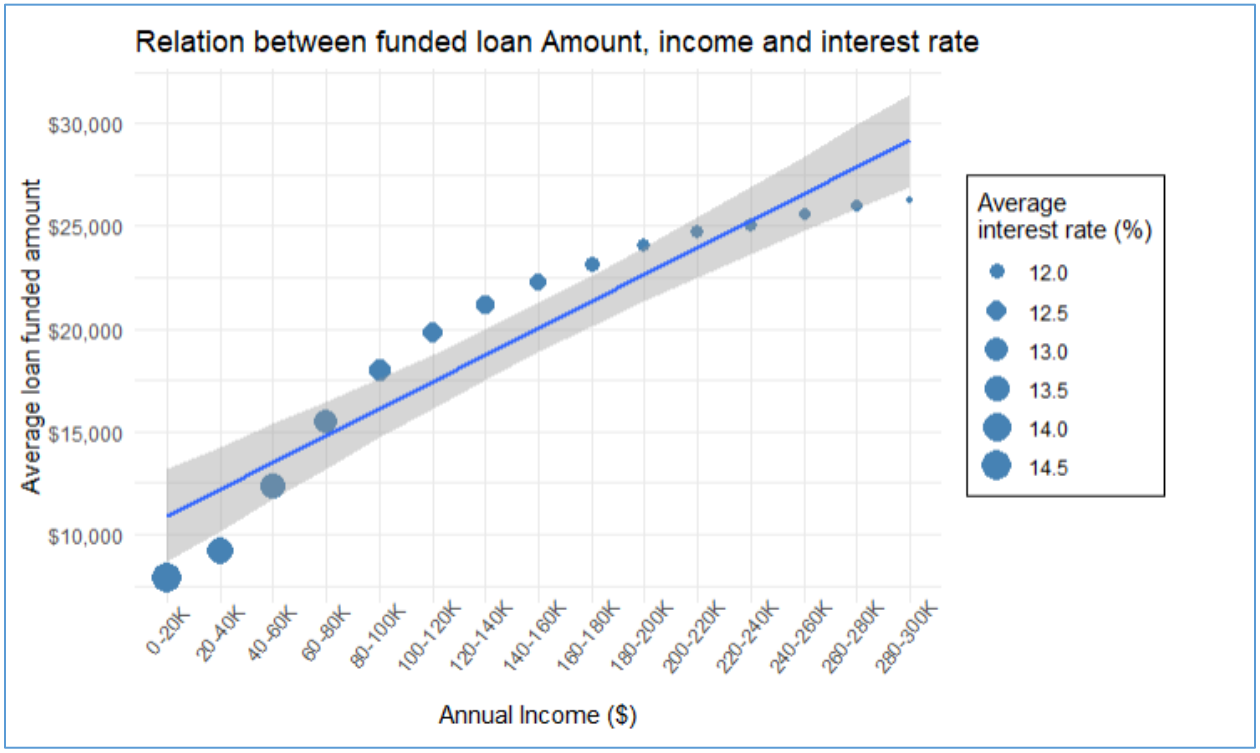


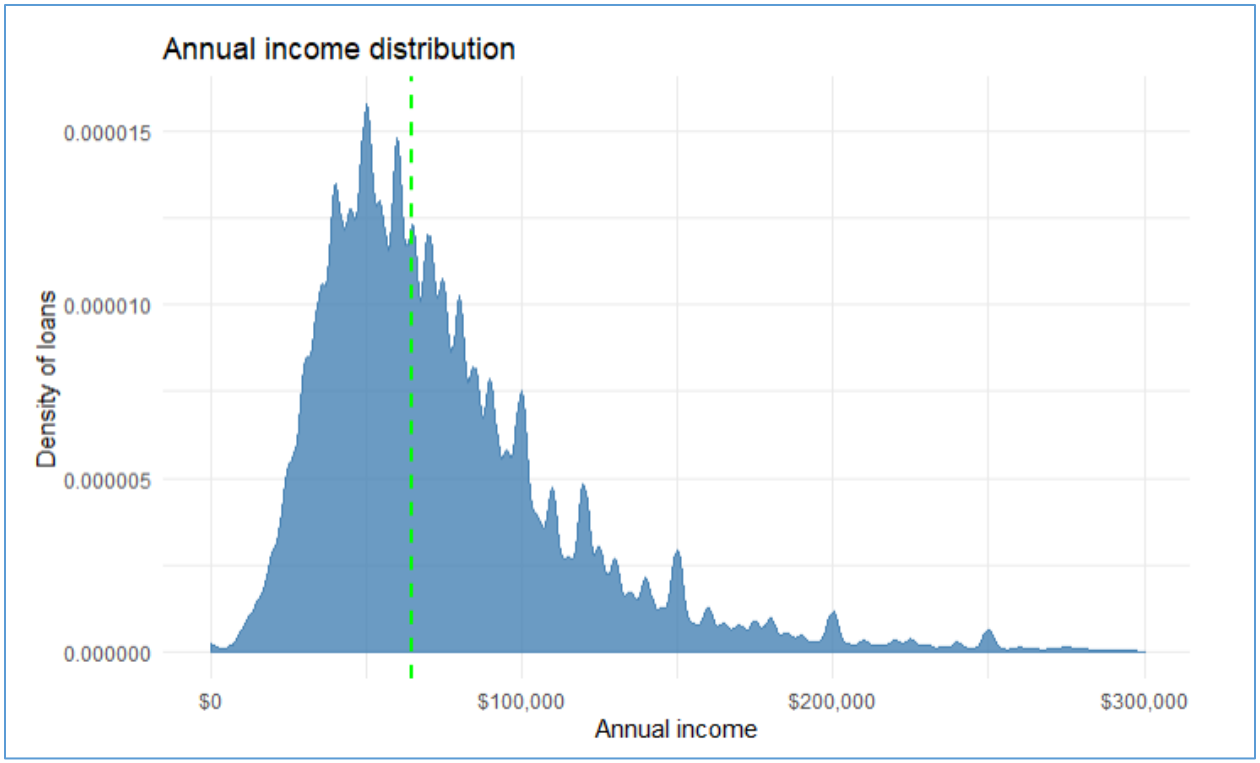

Total loan funded amount year on year (2007-2018)
$8,000,000,000
$6,000,000,000
$4,000,000,000
$2,000,000,000
$0
Total funded loan amount
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
Year

Loan Amount and term relation

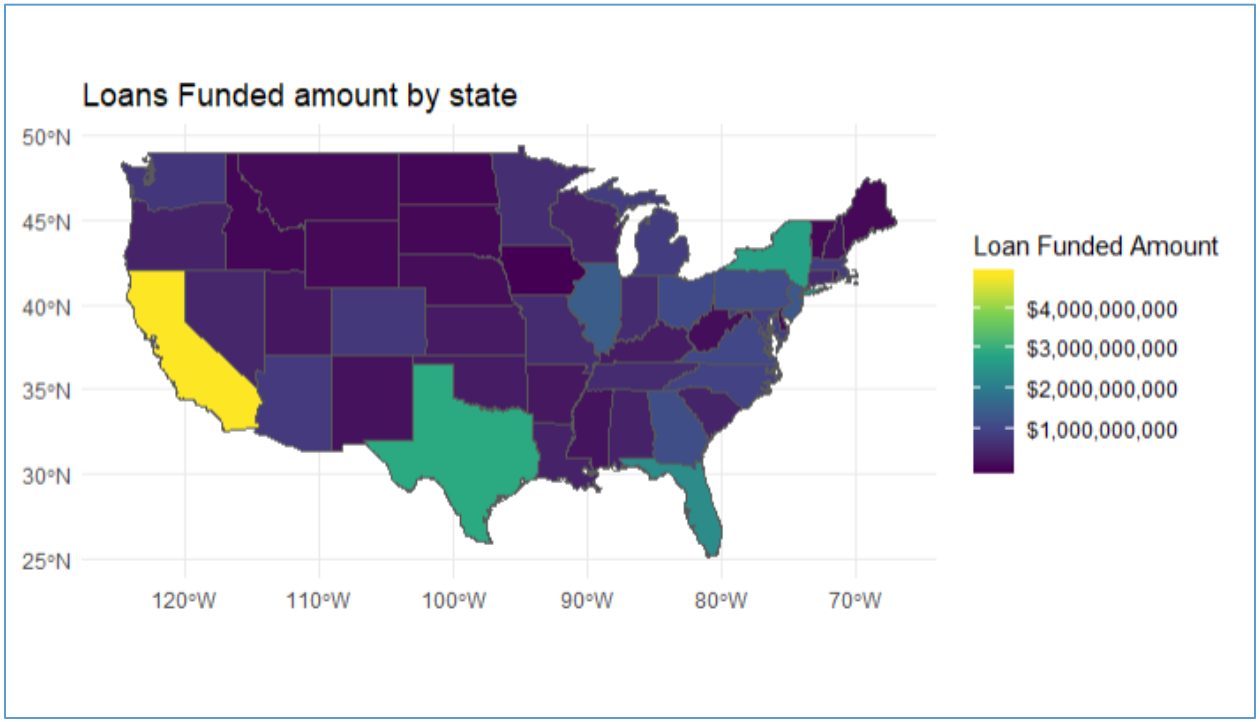


Loan funded amount, annual income and interest rate relation

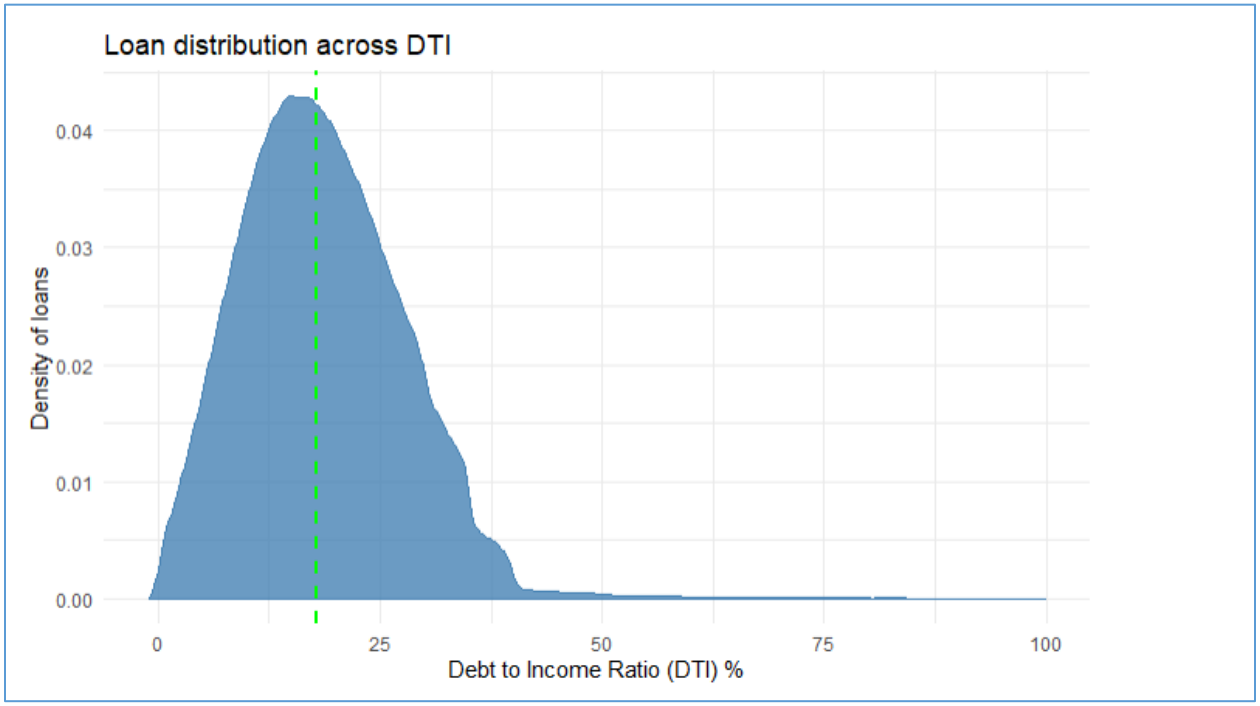



Annual income distribution
0.000015
0.000010
0.000005
0.000000
Density of loans
$0
$100,000
$200,000
$300,000
Annual income

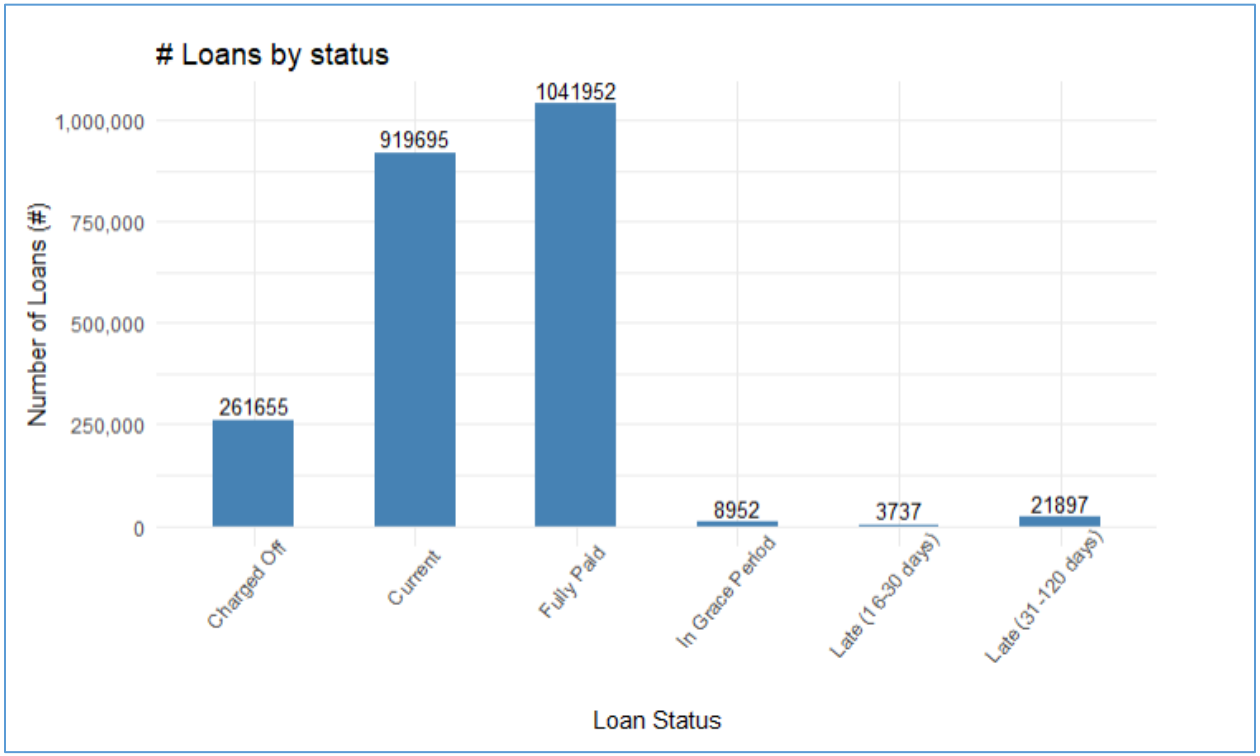
Loan funded amount by state.

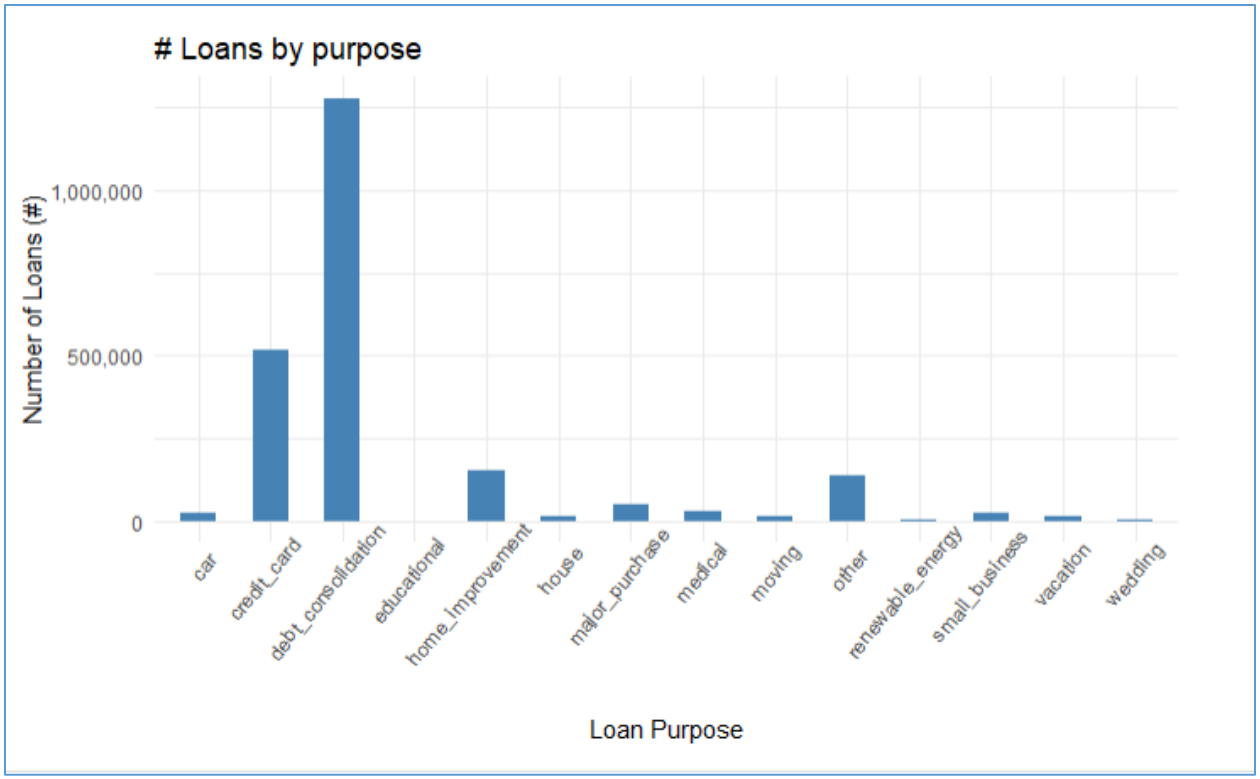


DTI Trend


Loan distribution across DTI
Density of loans
0.04
0.03
0.02
0.01
0.00
0
25
50
75
100
Debt to Income Ratio (DTI) %

Loan Status and Purpose




Loans by purpose
Number of Loans (#)
1,000,000
500,000
0
car
credit_card
debt_consolidation
educational
home_improvement
house
major_purchase
medical
moving
other
renewable_energy
small_business
vacation
wedding
Loan Purpose

R Code

R Markdown with charts

FinalAssignment

Amol Gote

3/7/2020

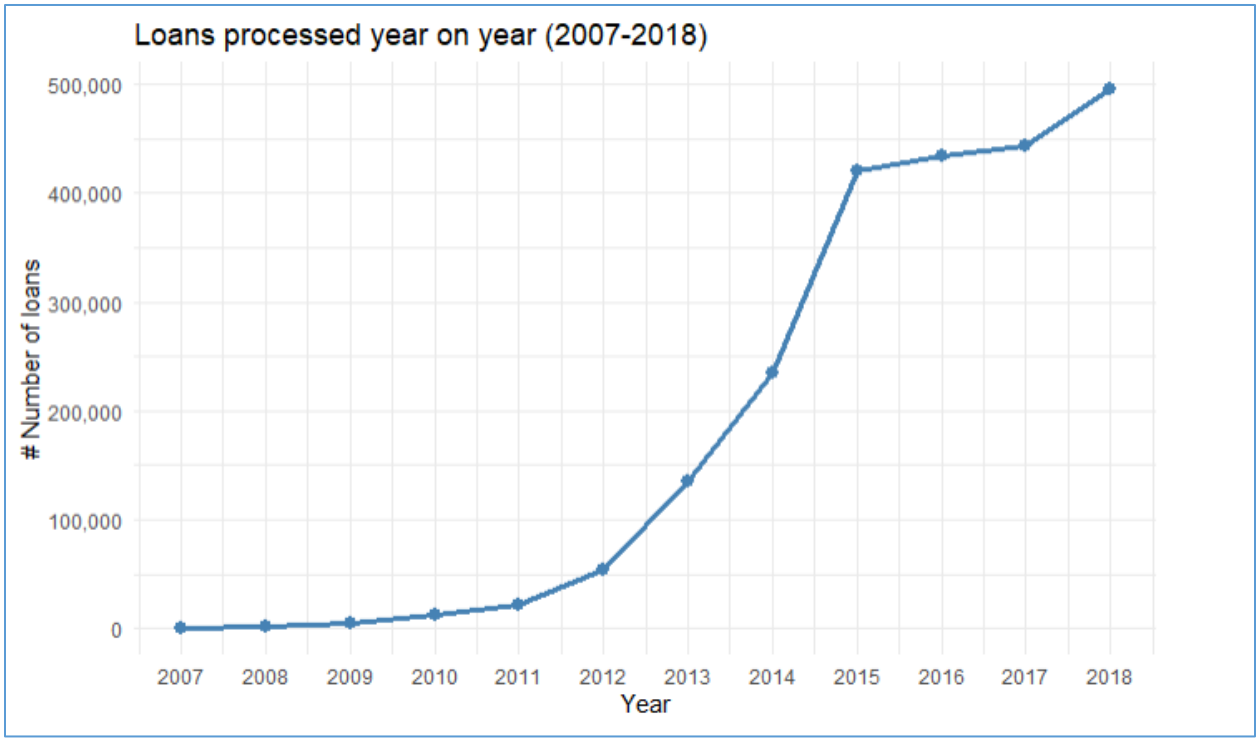
```
library(ggplot2)
library(readr)
library(gridExtra)
library(tidyverse)
library(sf)

library(maps)

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", "mt
```

```
hs_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", "nu
m_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", "t
otal_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_l
ine",
"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_py
mnt_inv",
"total_rec_prncp","total_rec_int","total_rec_late_fee","recoveries","collection_recovery_fee","las
t_pymnt_d",
"last_pymnt_amnt","next_pymnt_d","last_credit_pull_d","collections_12_mths_ex_med","mths_since_las
t_major_derog",
"policy_code","acc_now_delinq","tot_coll_amt","tot_cur_bal","inq_last_12m","avg_cur_bal","chargeof
```

```
f_within_12_mths",
"delinq_amnt","mort_acc","num_tl_120dpd_2m","num_tl_30dpd","num_tl_90g_dpd_24m","num_tl_op_past_12
m"))]

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")
```

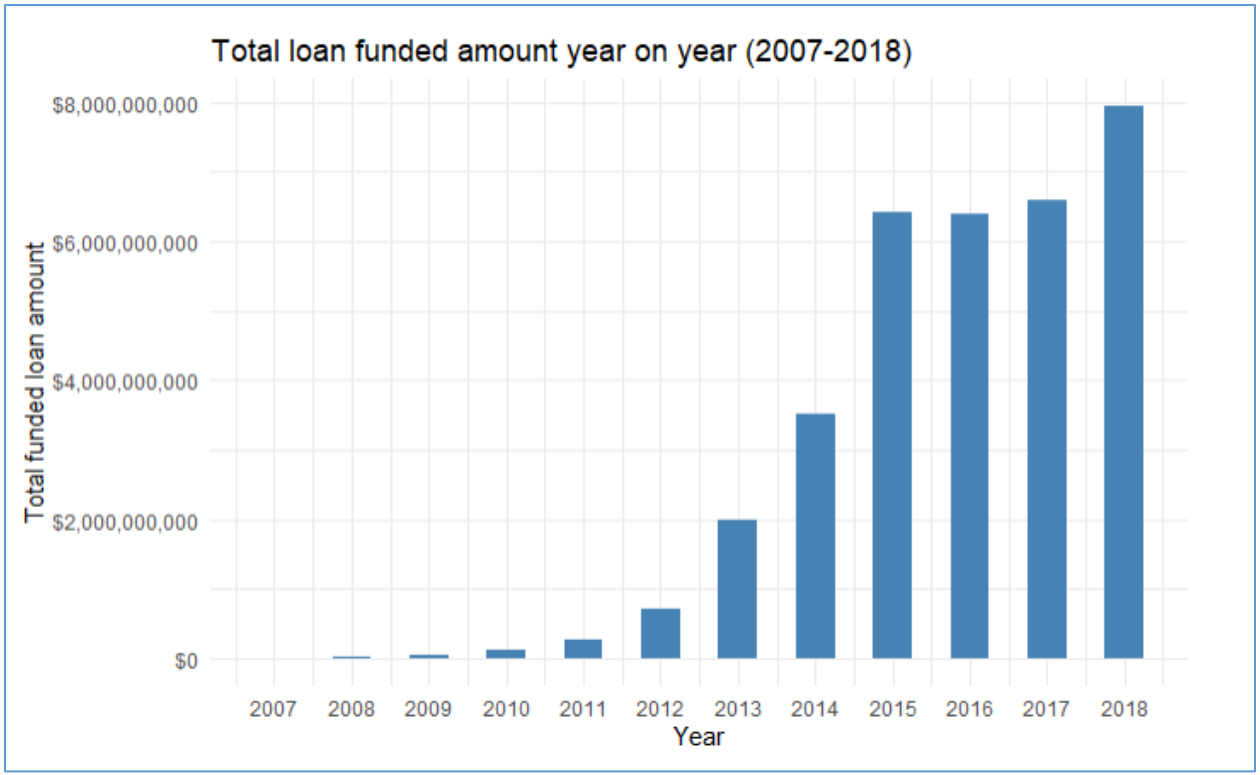
1. Below visualization shows number of loans funded each year from 2007 till 2018.
2. From 2007 till 2012, it had gradual progress in number of loans issued,
3. 2012 to 2015 number of loans issued has grown exponentially.
4. 2015 to 2017 saw a marginal hike in number of loans, 2017 to 2018 saw a considerable 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()
```


Loans processed year on year (2007-2018)
500,000
400,000
300,000
200,000
100,000
0
Number of loans
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
Year

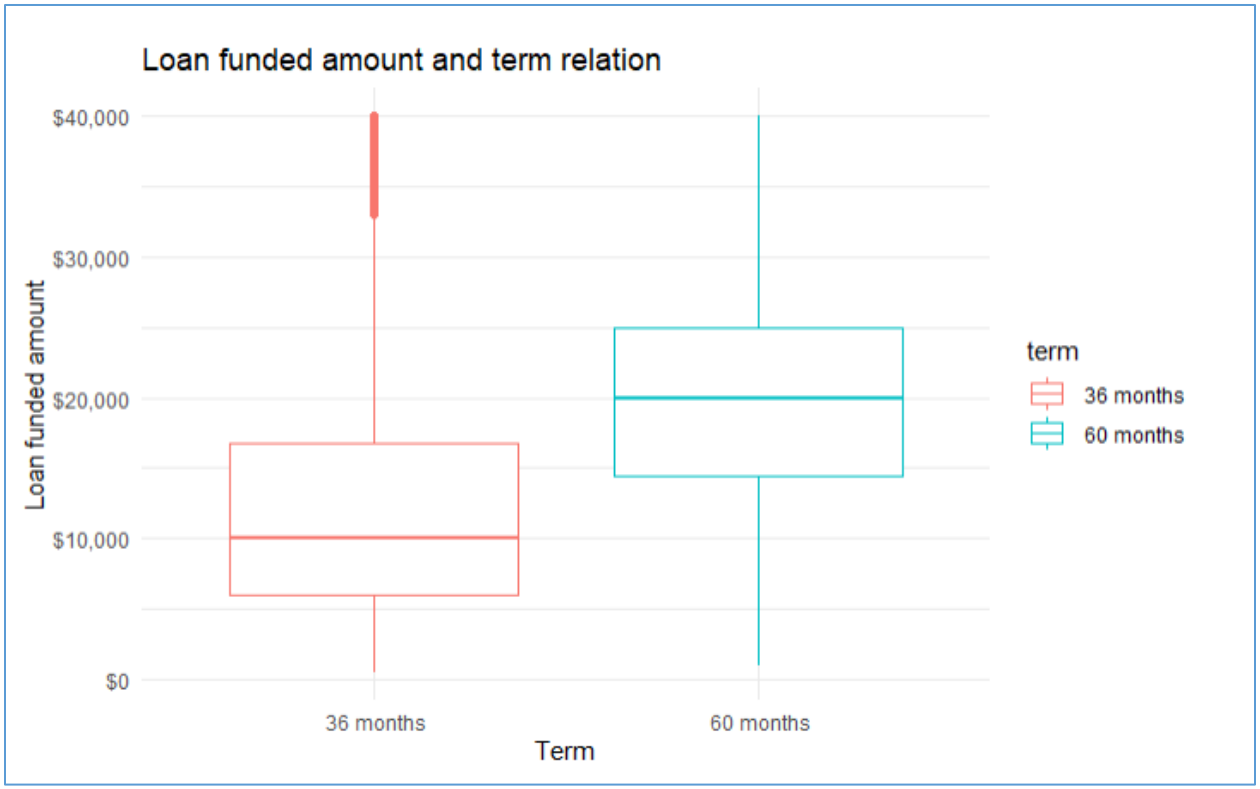
1. Below visualization shows total loan funded amount for each year from 2007 till 2018.
2. In general trend of total loan funded amount is in align to the number of loans issued as above visualization.
3. Lending club has funded loans maximum of $8 billion in 2018.
4. 2015 to 2017, total funded loan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


Total loan funded amount year on year (2007-2018)
$8,000,000,000
$6,000,000,000
$4,000,000,000
$2,000,000,000
$0
Total funded loan amount
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
Year

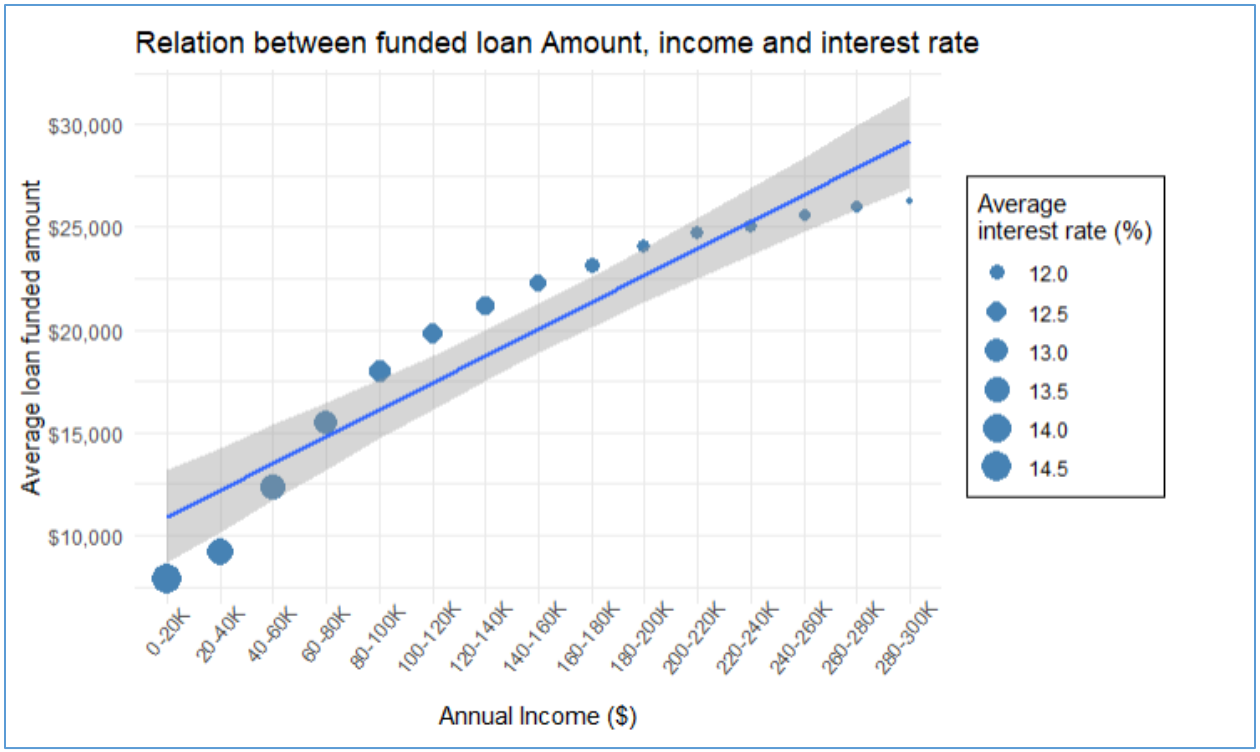
1. Below visualization shows relationship between loans funded amount and term.
2. Lending club offers only loans with 2 terms
    a. 36 Months
    b. 60 Months
3. For 36 months loan:
    c. Median loan funded amount is $10,000.
    d. Majority of the funded loan amount ranges from close $6000 to $16,000.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


Loan funded amount and term relation
$40,000
$30,000
$20,000
$10,000
$0
Loan funded amount
36 months
60 months
Term
term
36 months
60 months

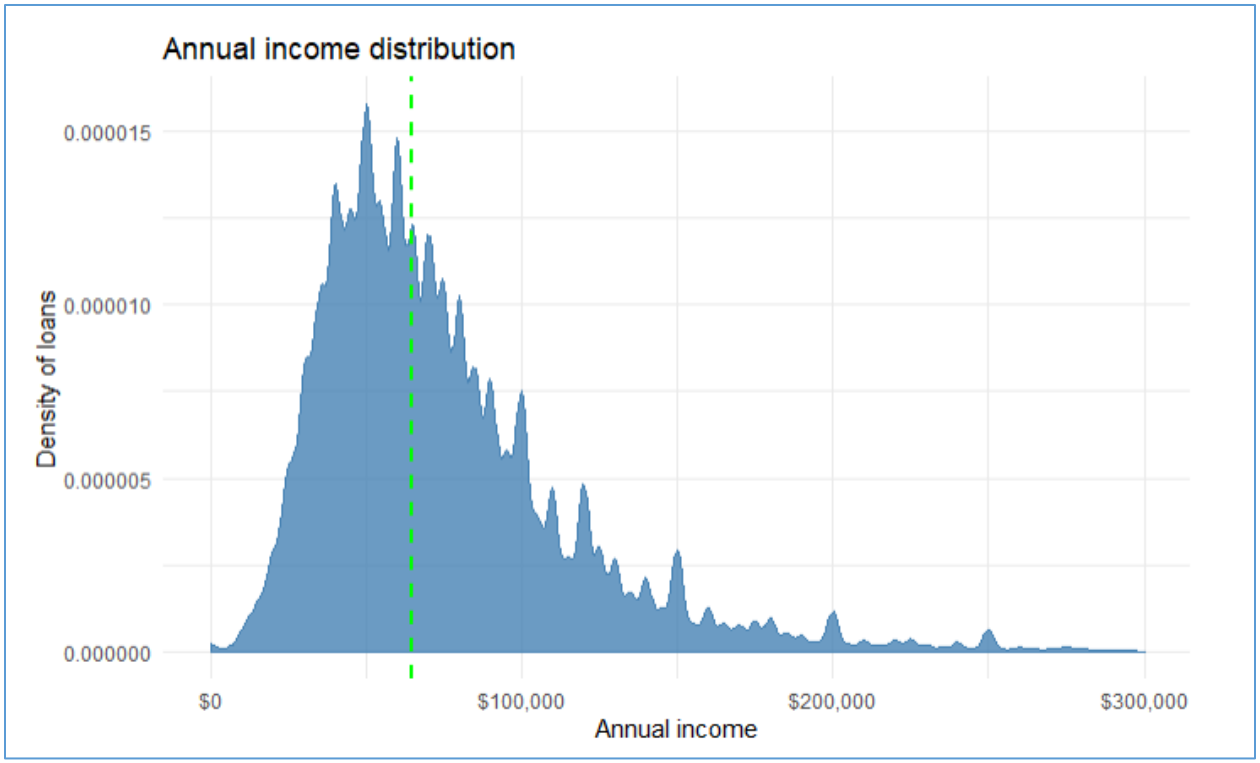
1. Below visualization shows relationship between funded loan amount, annual income and interest rate.
2. Visualization shows, that higher the annual income more is the funded loan amount.
3. As the annual income increase the interest rate drop by couple of percentage points.

```
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())
```


Relation between funded loan Amount, income and interest rate
Average loan funded amount
$30,000
$25,000
$20,000
$15,000
$10,000
0-20K
20-40K
40-60K
60-80K
80-100K
100-120K
120-140K
140-160K
160-180K
180-200K
200-220K
220-240K
240-260K
260-280K
280-300K
Annual Income ($)
Average interest rate (%)
12.0
12.5
13.0
13.5
14.0
14.5

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.

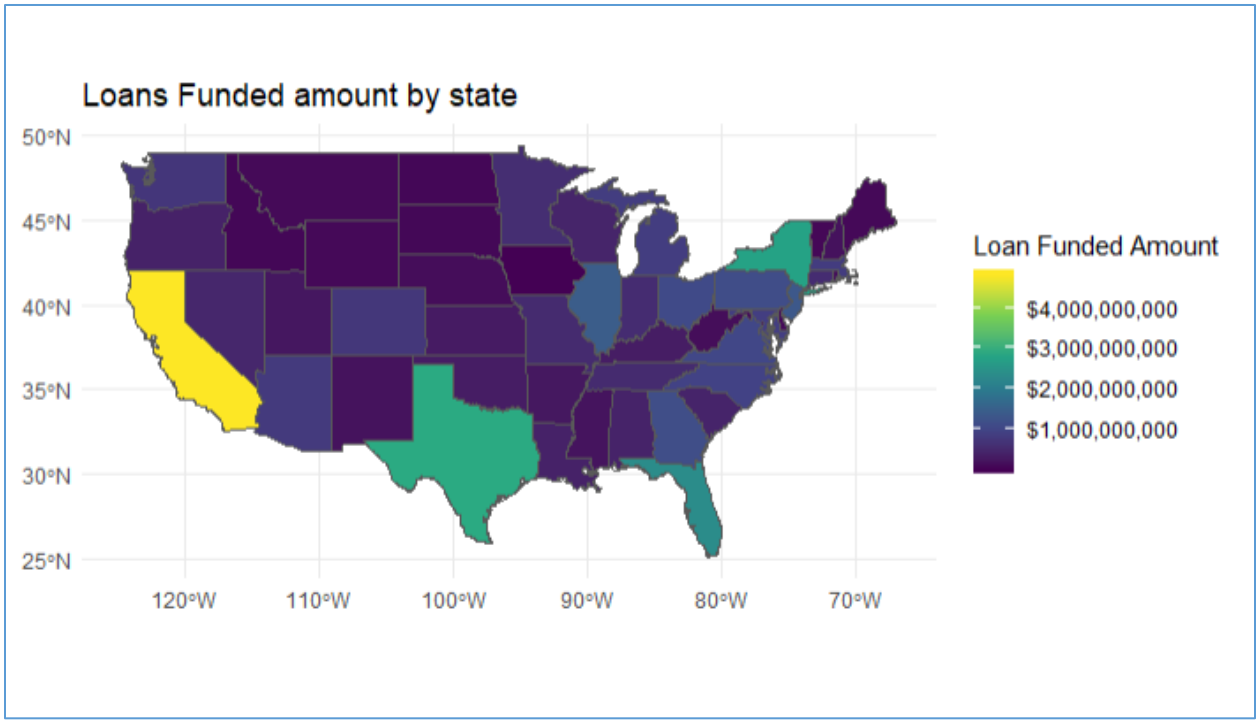
```
filteredLendingClubData <- lendingClubLoanData %>%
  drop_na(annual_inc)%>%
  filter(annual_inc < 300000)

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()
```


Annual income distribution
0.000015
0.000010
0.000005
0.000000
Density of loans
$0
$100,000
$200,000
$300,000
Annual income

1. Below visualization shows geographical distribution of total loan funded amount across various US states.

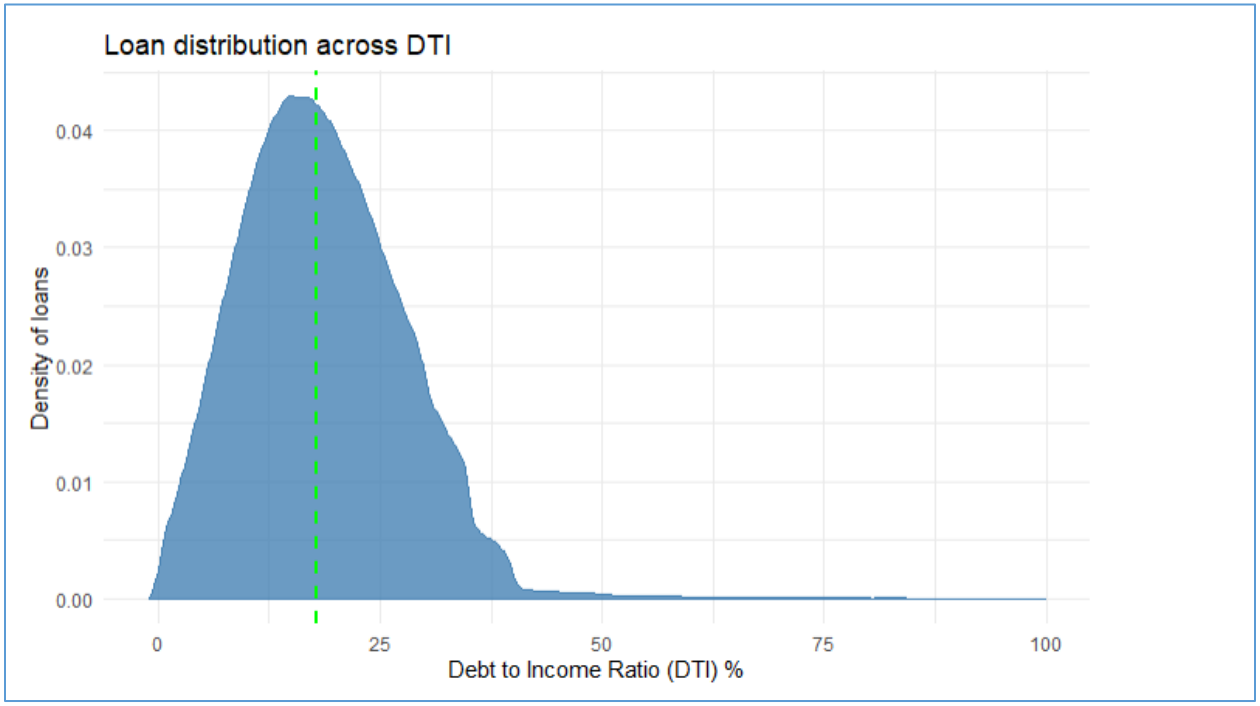
2. California has the highest total funded loan amount, followed by Texas, New York and Florida.

```
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" ))

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()
```


Loans Funded amount by state
50°N
45°N
40°N
35°N
30°N
25°N
120°W
110°W
100°W
90°W
80°W
70°W
Loan Funded Amount
$4,000,000,000
$3,000,000,000
$2,000,000,000
$1,000,000,000

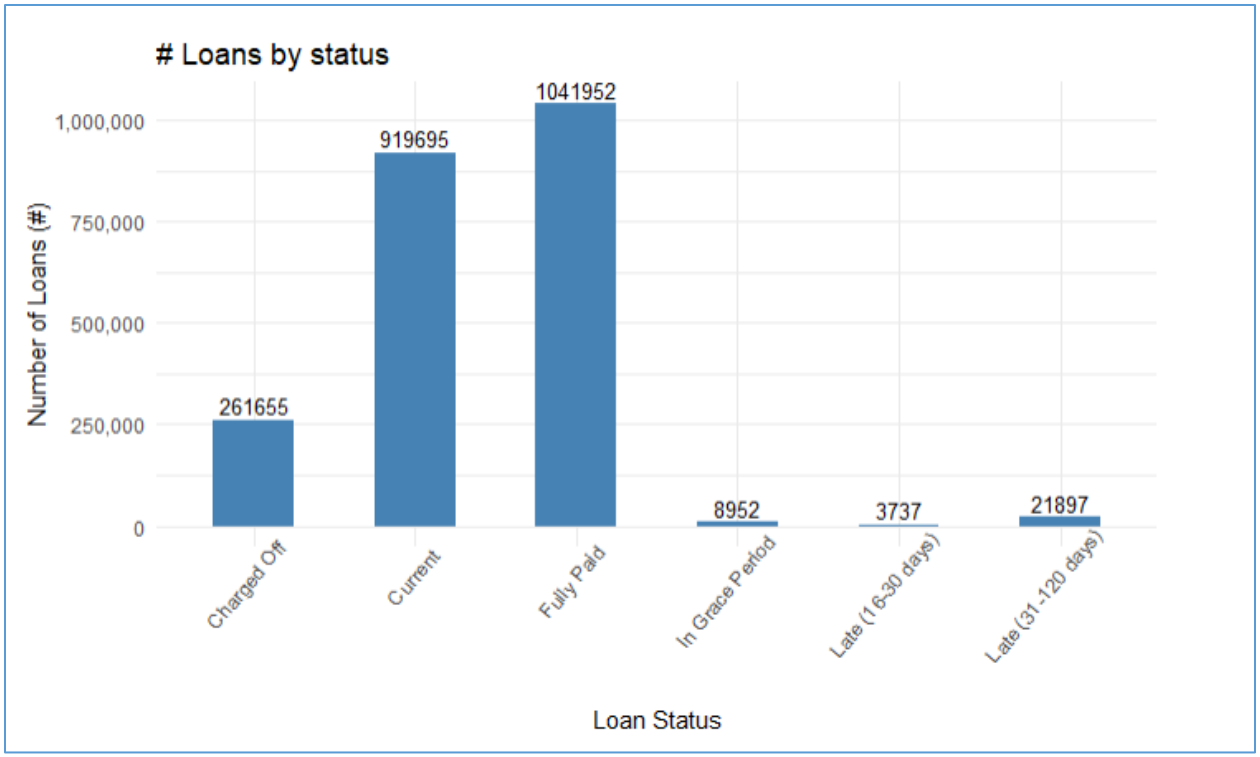
1. Below visualization shows the distribution of DTI (Debt to Income Ratio (%)) for the people getting loan.
2. Lower the DTI, higher the probability of getting the loan approved.

```
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()
```



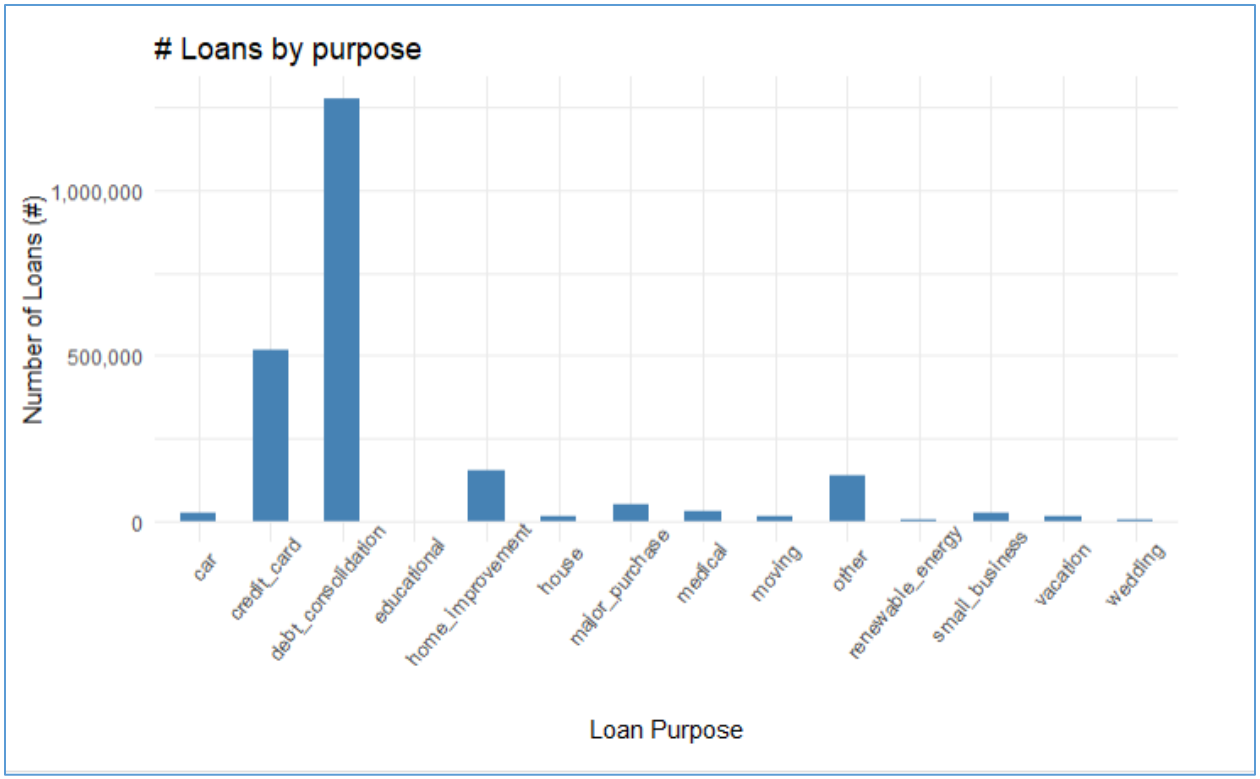
1. Below visualization shows loan distribution by the loan status.
2. Close to 900K loans are current and 1 Million loans are paid off.
3. Charged of loans are close to 2,61,655 out of total 2,260,668 loans. SO the charge of percentage is close to 12% (exact 11.57 %).
4. Most current loans will be eventually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


Loans by status
Number of Loans (#)
1,000,000
750,000
500,000
250,000
0
261655
919695
1041952
8952
3737
21897
Charged Off
Current
Fully Paid
In Grace Period
Late (16-30 days)
Late (31-120 days)
Loan Status

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))
```


Loans by purpose
Number of Loans (#)
1,000,000
500,000
0
car
credit_card
debt_consolidation
educational
home_improvement
house
major_purchase
medical
moving
other
renewable_energy
small_business
vacation
wedding
Loan Purpose

```r
filteredLendingClubData <- lendingClubLoanData %>%
  drop_na(loan_amnt) %>%
  drop_na(funded_amnt)
# Max loan amount
as.numeric(max(filteredLendingClubData$loan_amnt))
```

```
## [1] 40000
```

```r
# Minimum loan amount
as.numeric(min(filteredLendingClubData$loan_amnt))
```

```
## [1] 500
```

```r
# 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)
```

```r
filteredLendingClubData <- lendingClubLoanData %>%
  drop_na(dti)
# Max DTI
as.numeric(max(filteredLendingClubData$dti))
```

```
## [1] 999
```

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

```
## [1] -1
```

```r
# Total loan funded amount
as.numeric(sum(filteredLendingClubData$funded_amnt))
```

```
## [1] 33971525425
```

```r
filteredLendingClubData <- lendingClubLoanData %>%
  drop_na(annual_inc)%>%
  filter(annual_inc < 300000)
# Annual income median
median(filteredLendingClubData$annual_inc)
```

```
## [1] 65000
```

```r
filteredLendingClubData <- lendingClubLoanData %>%
  drop_na(dti)%>%
  filter(dti < 100)
# Dti median
median(filteredLendingClubData$dti)
```

```
## [1] 17.82
```

```r
#Number of loans
nrow(lendingClubLoanData)
```

```
## [1] 2260668
```

Link to dashboard

https://amol-gote.shinyapps.io/Week8/

Executive Summary

- 2.2 million Loans have been funded by lending club.
- Total funded loan amount is close to $ 33 billion
- Year on year there has been steady growth in number of loans funded and total funded loan amount.
    a. Year 2012 to 2015 saw the maximum growth.
    b. Highest number of loans issued was in 2018, close 500,000 loans
    c. Maximum total loan funded amount was in 2018 close to $8 billion.
- Loan distribution across term
    a. Lending club offers only loans with 2 terms 36 months and 60 months.
    b. For 36 months loan:
        i. Median loan funded amount is $10000.
        ii. Majority of the funded loan amount ranges from close $6000 to $16000.
    c. For 60 months loan:
        i. Median loan funded amount is $20000.
        ii. Majority of the funded loan amount ranges from close $15000 to $25000.
- Relationship between funded loan amount, annual income and interest rate
    a. Higher the annual income more is the funded loan amount.
    b. As the annual income increase the interest rate drop by couple of percentage points.
- Median annual income is $65,000. Majority of lenders are having annual income less than $200K.
- Lower the DTI higher the probability of getting the loan approved. Median DTI is close to 18% (17.82%). Above 38% DTI probability of loan getting funded is less.
- California has the highest total funded loan amount, followed by Texas, New York and Florida. Total Loan funded amount in California is greater than $4 billion.
- Loan performance

    a. Out of 2.2 million total loans, Close to 900K loans are current and 1 Million loans which are paid off.
    b. Charge of percentage is close to 12% (exact 11.57 %). So essentially 12% of loans are bad loans and remaining close 88% are good loans.
- Loan Purpose
    a. Debt consolidation is major primary purpose of taking loan from lending club.
    b. Second major purpose of taking loan is for paying credit card bills.