CSC 465 Final Project

Project: Loan Club Data

Reporter: Yingping Li

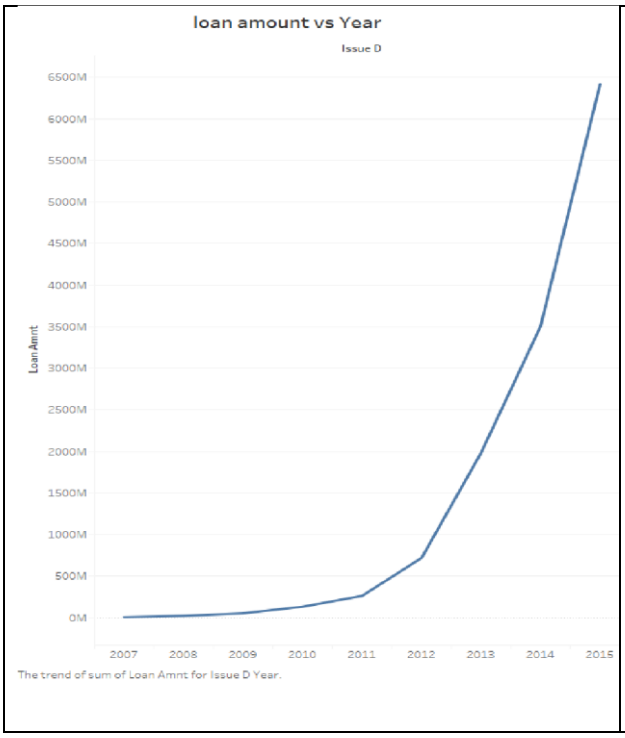
## Table of Contents

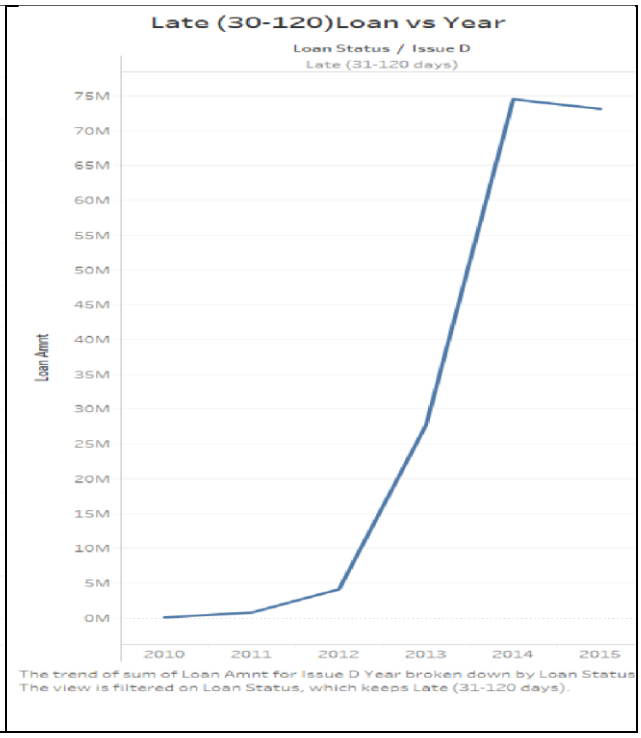


### Exploratory visualization of the data

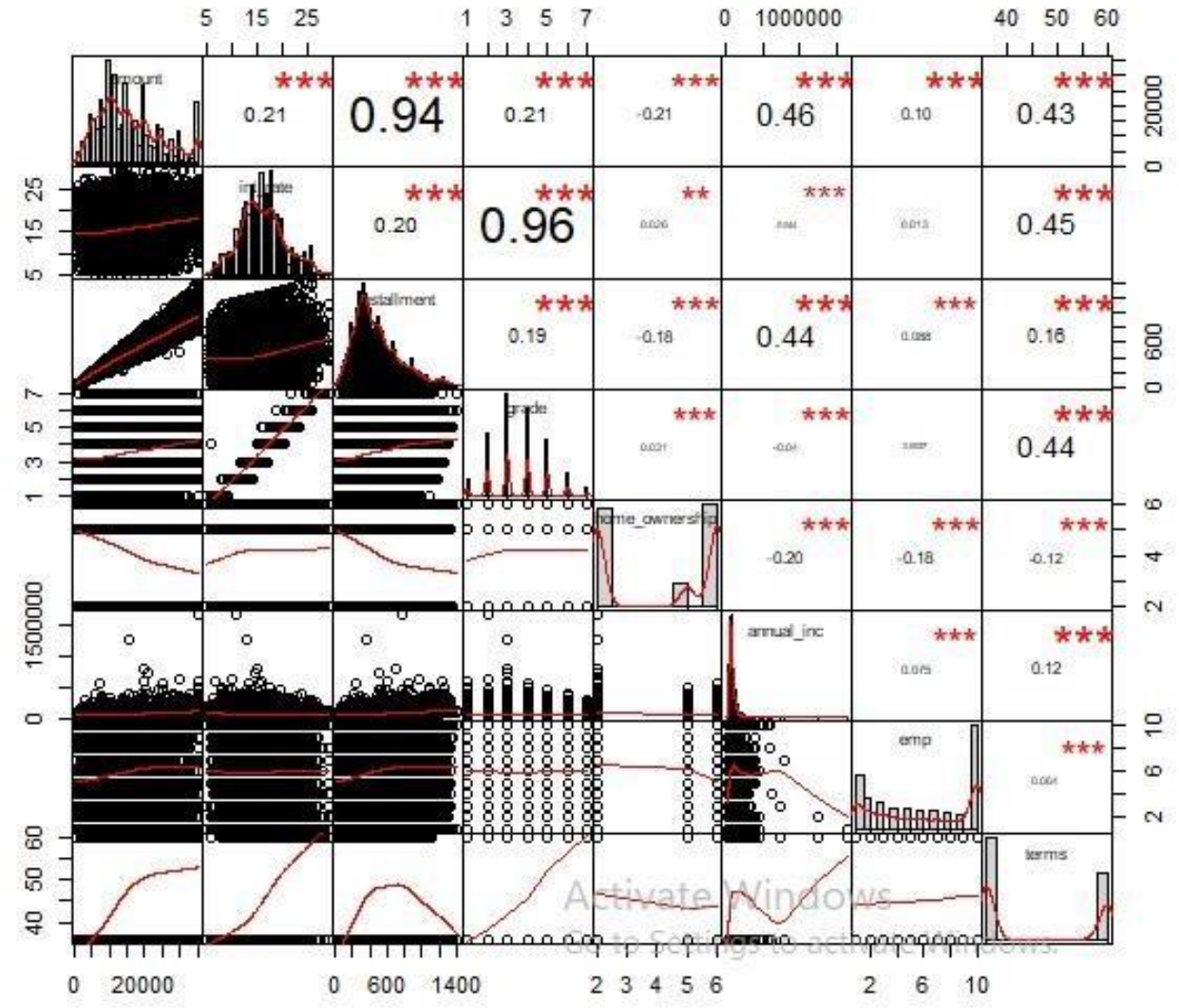
The goal of the report is to use visualize tools to find interested information from the loan club data set. The loan club data sets include 887380 objects and 148 columns. We are interested in loan amount, state, interest rate, installment, grade, home-ownership, annual income, employ length, term, issue day, loan status, title, purpose attributes

The loan amount graph below shows that the loan amount increases as the year increase. Surprisingly, the late loan amount which was extracted from the loan amount attribute filtered by status  attribute decreased at 2014 while the loan amount increased at 2014. The two line graph below clearly shows the difference between the loan amount and late loan amount with years change.



*Scatterplot Correlation Matrix of 8 attributes: late loan, interest rate, installation, grade, home ownership, annual income, employee length, terms*




The Scatterplot Correlation Matrix plot above shows the relationship between late loan amount, interest rate, installation, grade, home ownership, annual income, employee length, terms attributes. The plot clearly shows the histogram of each attribute in the diagonal boxes and its correlation with other attributes and the p- value level of the correlation coefficients. The plot shows there is a strong positive relationship (0.96 correlation coefficient and three stars p value) between late loan amount and installment.

**Selection of visualization techniques**

Scatterplot Correlation Matrix by R was used to show the relationships of multi-dimensional attributes as exploratory purpose. It shows the correlation coefficient of each pair of attributes as well as the p-value significant level. The diagonal line box shows the histogram of each attributes.

Geospatial map  by Tableau was used to show the distribution of the loan amount over each state. The size of the dot with color was used to represent the amount of the total loan amount in each state.

Choropleth map coordinated with color and log scale by R was used to show each state's late loan amount and late loan rate of each state to see their difference. Red color used in the map show the warning.

Tree map by Tableau was used to find the distribution of the subclasses in loan status attributes.

Heat map with Tableau was used to show the change patterns of loan amount and normalized loan amount  as the time change of month and year.

Word cloud by R was used to find the high frequency words in purpose and title attributes. The word frequencies scaled by log was used to smooth the size of the word and get better results.

Bar chart and Line chart by Tableau were used to show the relationships between attributes and subclasses.

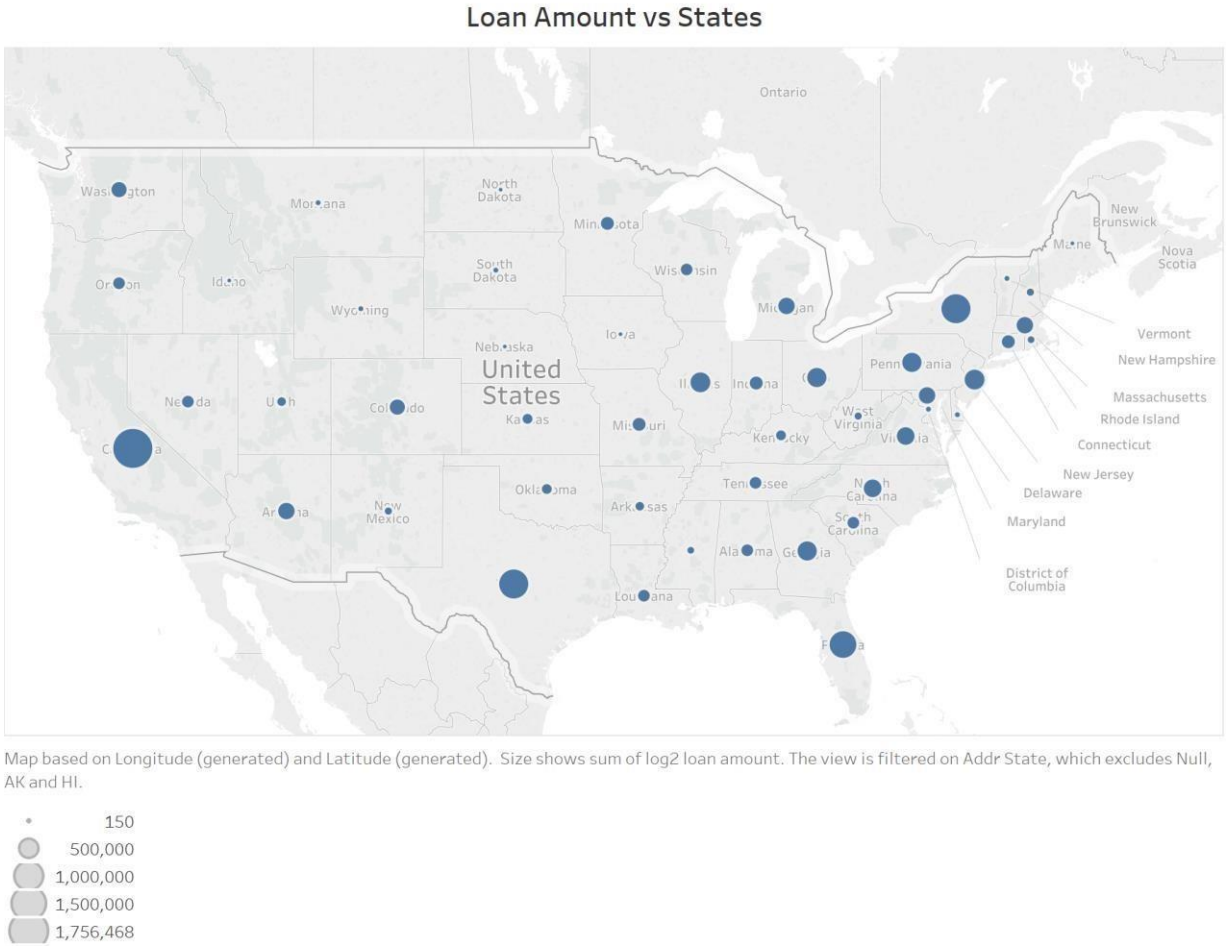
Level plot was used to show the relationship of bas loan issue day.

Rose plot were used to show the subclasses of loan statues.

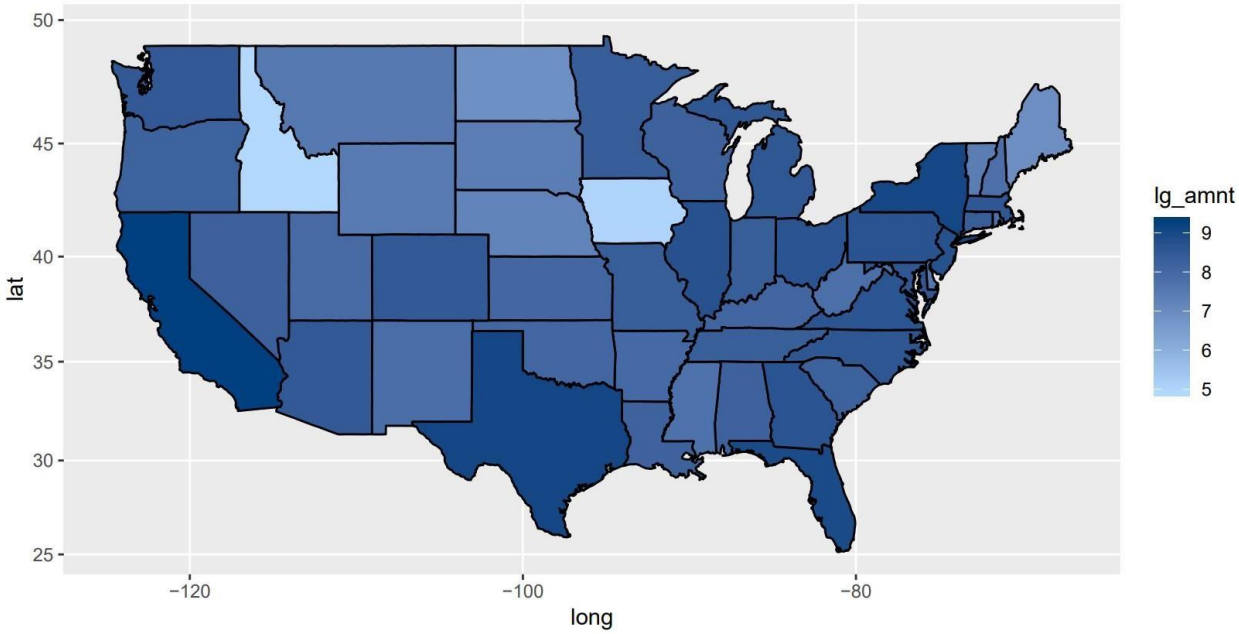
**Discussion and Results**

The map graph below shows the distribution of loan amount over states. The total loan amount in the graph was scaled by log  to get a better visualization result. The size of the blue dot shows the amount of the total loan amount of the state. The bigger the blue dot, the larger the loan amount. California has the biggest  blue dot, which means it has the highest total loan amount.

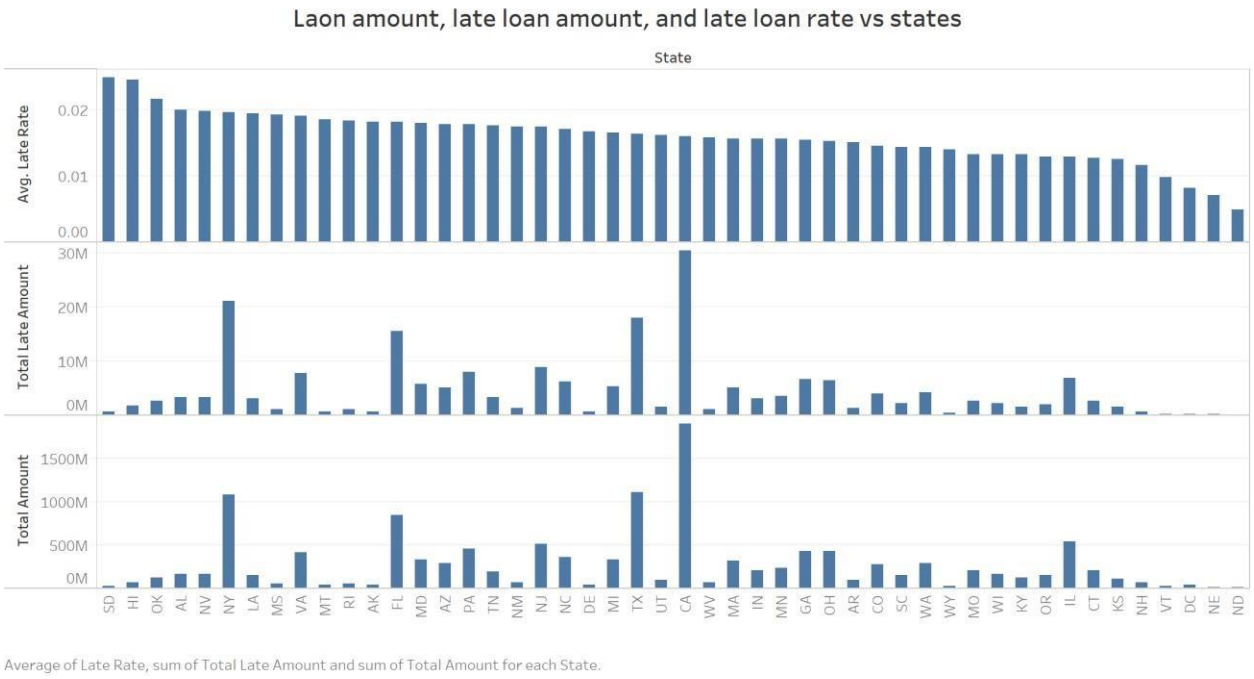
*Figure 1 Loan amount vs. states*



*Choropleth of Loan amount vs states*




The Choropleth above shows the loan amount of each state. The dark blue shows the high value of the loan amount.




The multi-dimensional bar chart shows that the loan amount , late loan amount, and late loan amount rate of each state. The state of CA has the highest total loan amount and total late

amount, but it does not have the  highest average late loan rate. SD has the highest average late loan rate but does have a very low total loan amount and total late loan amount.

*Figure 2 Loan status*


The tree map shows the loan amounts of subclasses of the loan status attribute. The blue area is the biggest, which means that Current is the highest, Fully Paid which is the second, and Charged Off the third, and late the fourth. We are interested in late debt.

*Figure 3 Moving average loan amount t Normalized by year average vs. year vs. month*

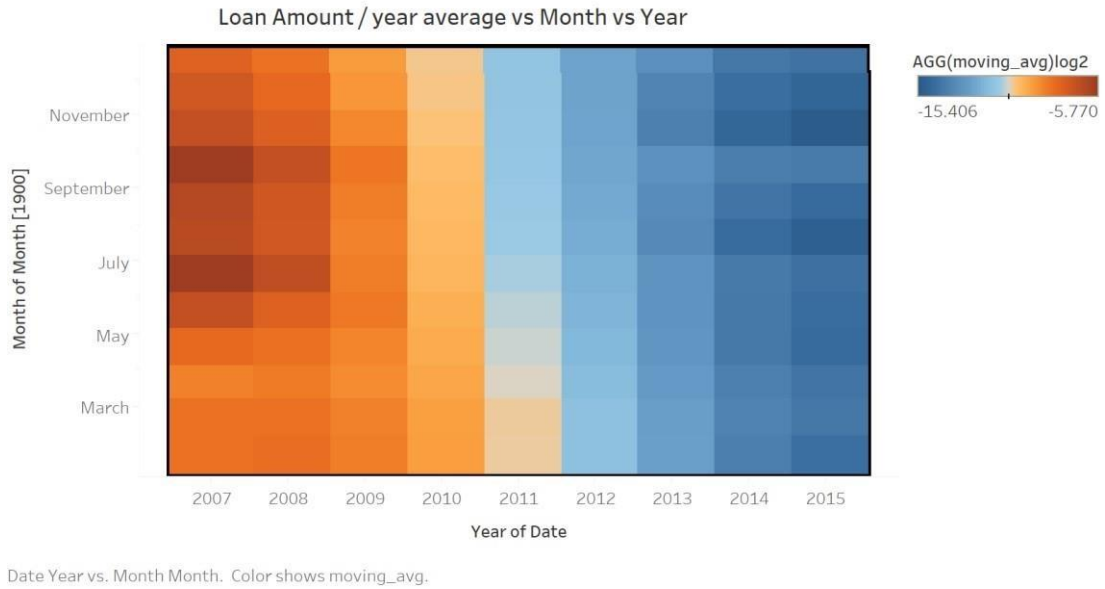
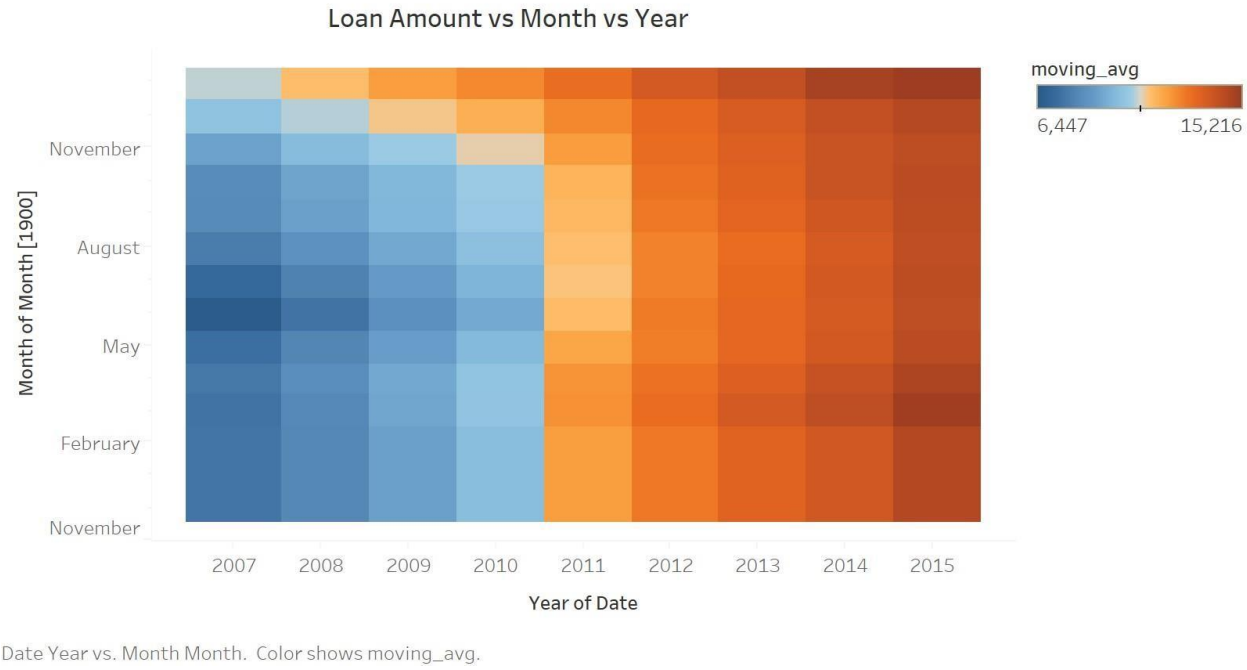



Figure 3 shows the change of the loan amount, which is divided by the average year amount through the year period 2007 to 2015 and the months during each year. The moving average is used to express the change. Figure 3 shows that July 2007 and October 2007 have the highest normalized loan amount. 2010 and 2011 have an average normalized loan amount in the years. And the first quarter of 2011 has the lower normalized loan amount in 2011. Normalized loan amount from 2012 to 2015, which is shown in blue is lower than the year from 2007 to 2010, which is shown in gold.

*Figure 4 Moving average loan amount vs year vs month*




In compared to Figure 3, the change of loan amount without normalized during the years from 2011 to 2015 is higher than that of from 2007 to 2010 because the colors switched the side in Figure 4. This means that the absolute loan amount is increasing while the normalized loan amount is decreasing as the year increasing. July 2007 has the lowest loan amount, and January 2015 has the highest loan amount.

*Figure 5 Late Loan vs states*

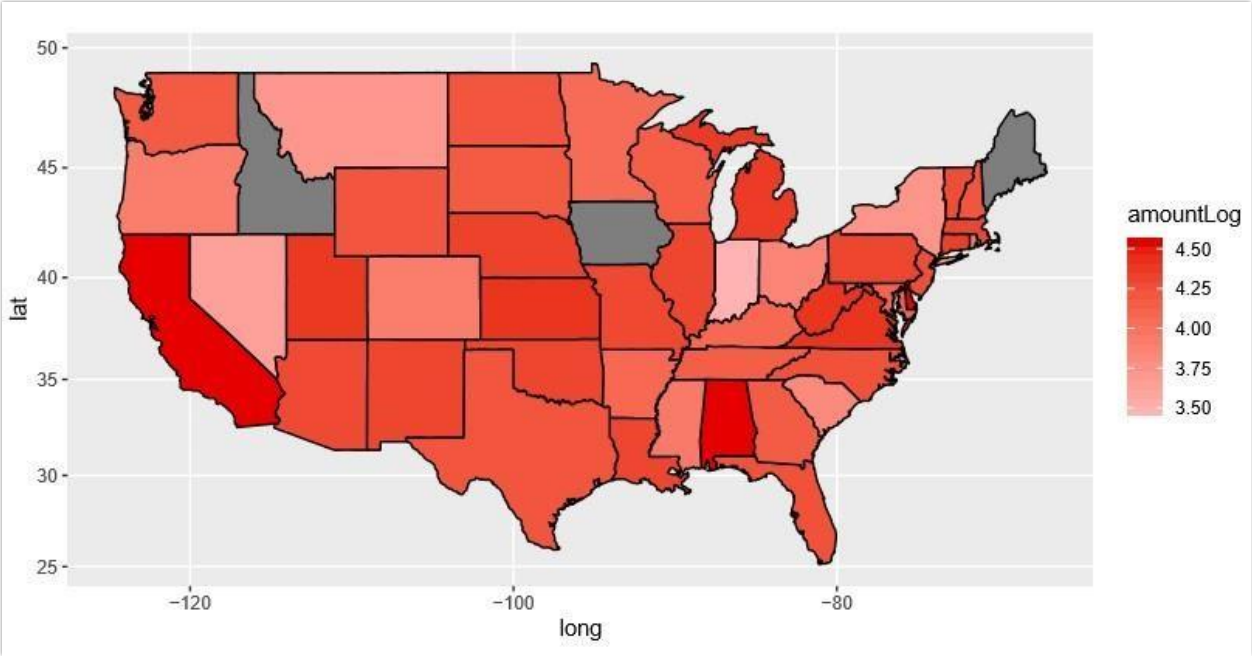



Figure 5 shows the distribution of the late loan amount with a log scale over the states. CA has the highest late loan amount. Red color means warning, and dark color means higher level.

*Figure 6 Late Loan Rate vs State*

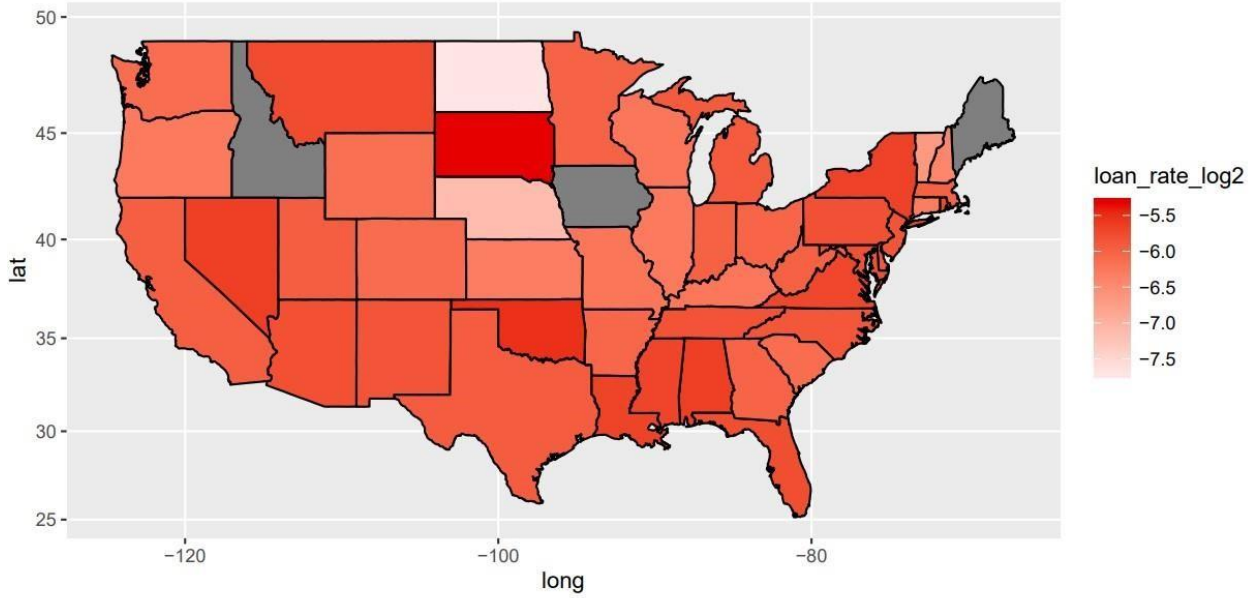


Figure 6 shows the late loan rates over the states. The difference between Figure 5 and Figure 6 show some states has a higher amount of late loan but has lower late loan rate such as CA.

*Figure 7 Late debt purpose word cloud with log2 frequency*

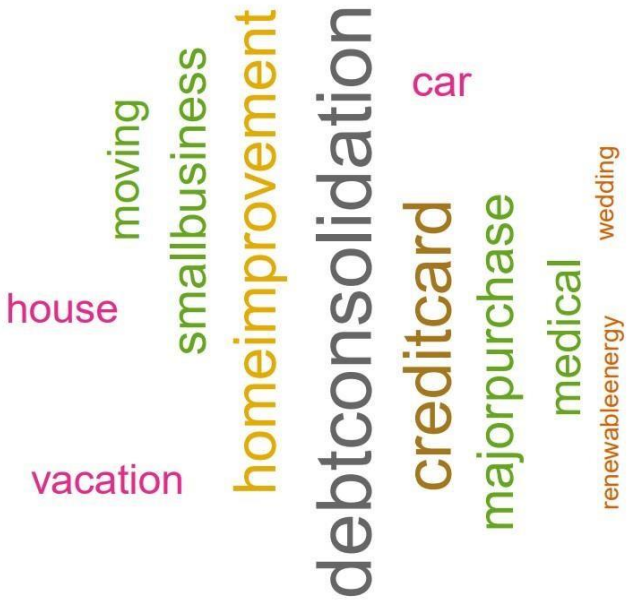



Figure 7 is created by R with the word frequency being log2 scaled. The purpose attribute of late loan word cloud shows that “debtconsolidation” has the highest frequency. The word s frequencies scaled by log smoothened the size of the words and got a better result.

*Figure 8 Late debt Title word cloud with log2 words frequency*

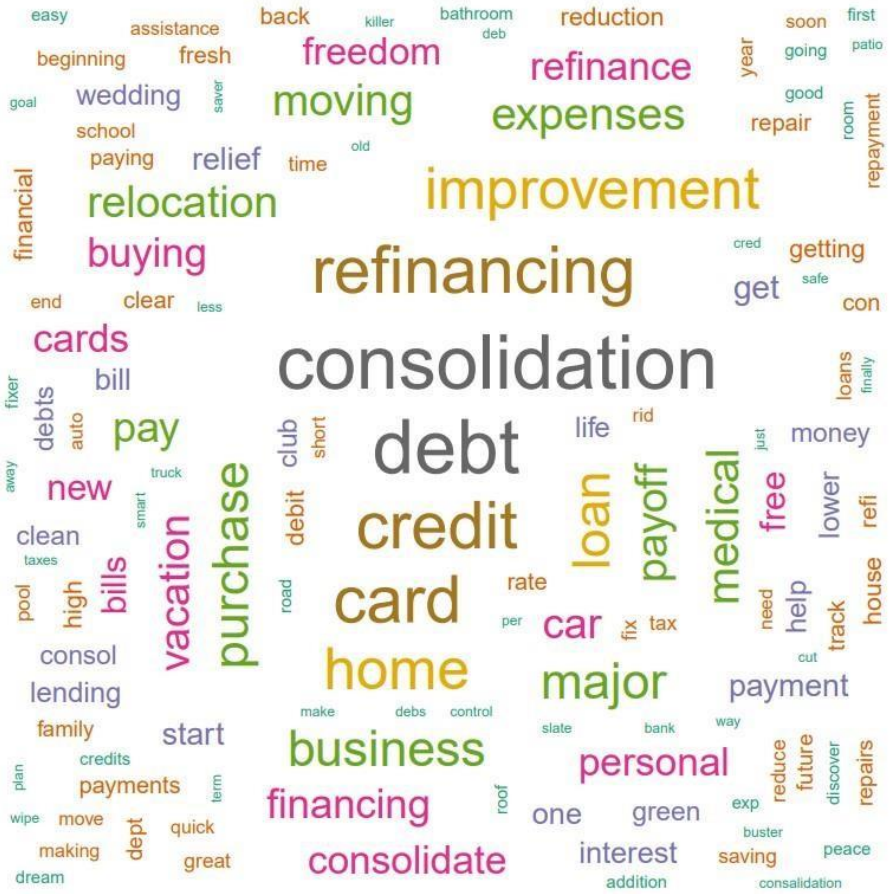
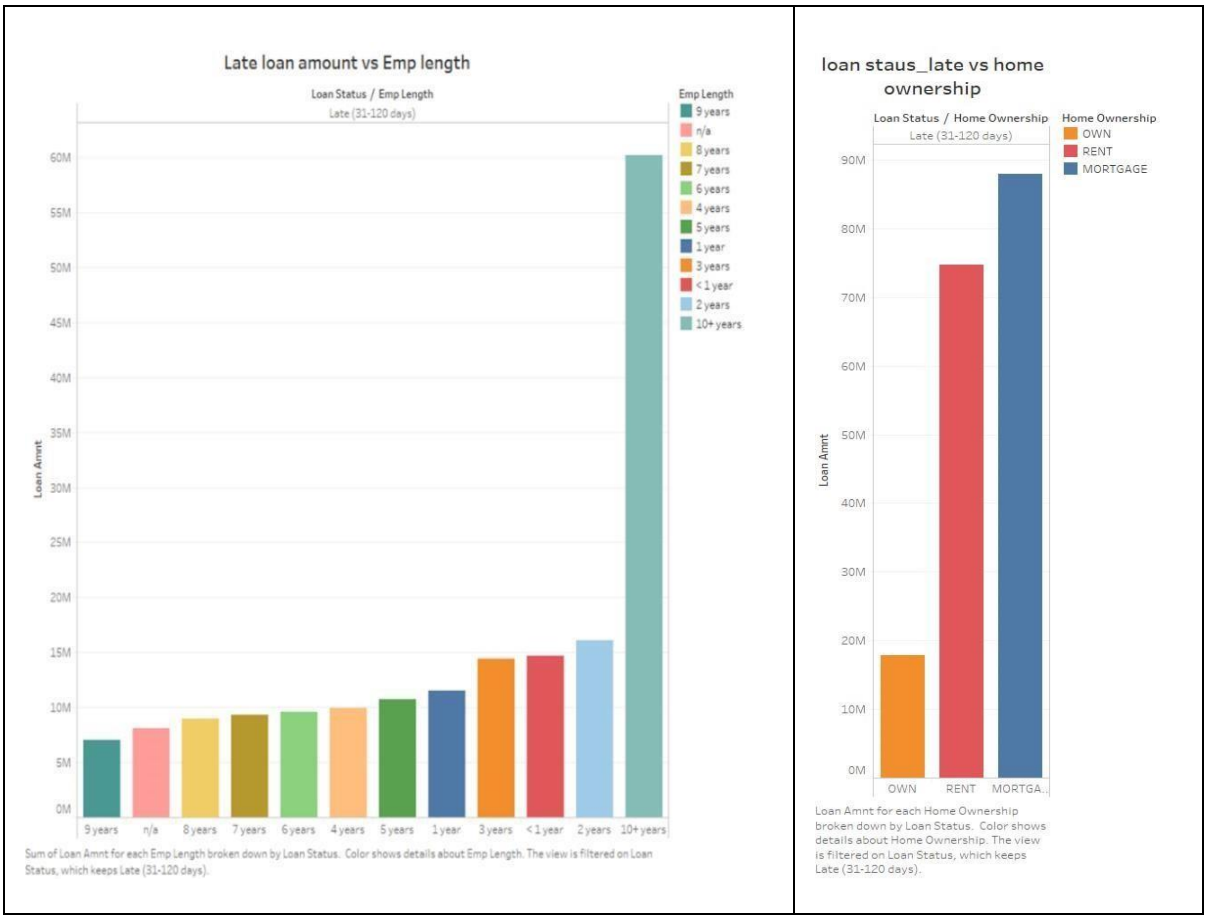
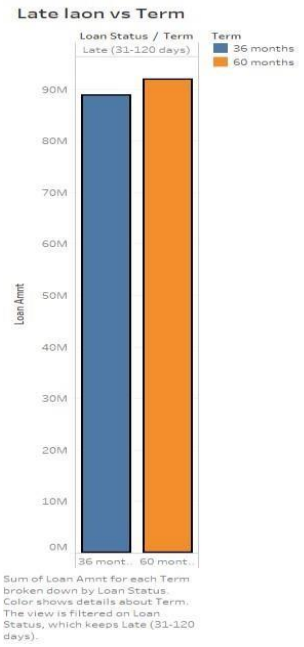
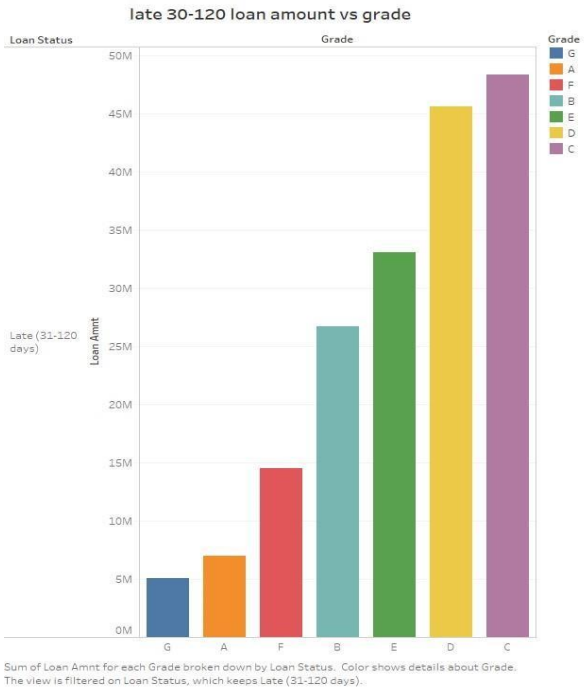
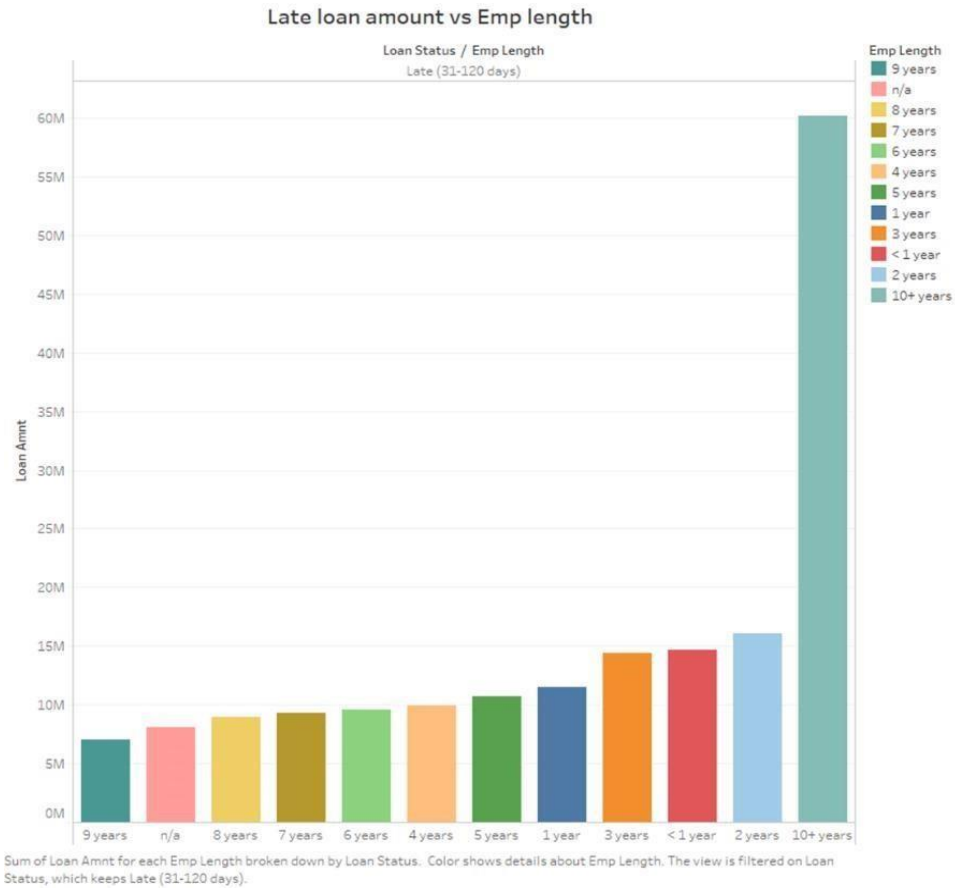



Figure 8, the title of the late loan word cloud, shows that consolidation debt has the highest frequency. The word frequencies were scaled by log2, which smoothened the size of the words and got a better result than the unscaled word frequencies.


Late loan amount vs Emp length
Loan Status / Emp Length
Late (31-120 days)
Emp Length
9 years
n/a
8 years
7 years
6 years
4 years
5 years
1 year
3 years
< 1 year
2 years
10+ years
Loan Amnt
60M
55M
50M
45M
40M
35M
30M
25M
20M
15M
10M
5M
0M
9 years
n/a
8 years
7 years
6 years
4 years
5 years
1 year
3 years
< 1 year
2 years
10+ years
Sum of Loan Amnt for each Emp Length broken down by Loan Status. Color shows details about Emp Length. The view is filtered on Loan Status, which keeps Late (31-120 days).
loan staus_late vs home ownership
Loan Status / Home Ownership
Late (31-120 days)
Home Ownership
OWN
RENT
MORTGAGE
Loan Amnt
90M
80M
70M
60M
50M
40M
30M
20M
10M
0M
OWN
RENT
MORTGA..
Loan Amnt for each Home Ownership broken down by Loan Status. Color shows details about Home Ownership. The view is filtered on Loan Status, which keeps Late (31-120 days).


Late laon vs Term
Loan Status / Term
Late (31-120 days)
Term
36 months
60 months
90M
80M
70M
60M
50M
40M
30M
20M
10M
0M
Loan Amnt
36 mont..
60 mont..
Sum of Loan Amnt for each Term broken down by Loan Status. Color shows details about Term. The view is filtered on Loan Status, which keeps Late (31-120 days).


late 30-120 loan amount vs grade
Loan Status
Grade
Grade
G
A
F
B
E
D
C
50M
45M
40M
35M
30M
25M
20M
15M
10M
5M
0M
Late (31-120 days)
Loan Amnt
G
A
F
B
E
D
C
Sum of Loan Amnt for each Grade broken down by Loan Status. Color shows details about Grade. The view is filtered on Loan Status, which keeps Late (31-120 days).




The bar chats above show that 10 years loan has the largest amount of late loan among the employee length. The home owner has the lowest late loan amount in the home ownership attribute. Grade G has the lowest late loan amount in the Grade attributes. There is not a bigger difference between the late loan amount of 36 months term and that of 60 months.

## Reference

Lending Club Loan Data download from

https://www.kaggle.com/wendykan/lendingclubloan-data/version/1

## Appendix

R code:

```
data<-read.csv("C:/Users/Yingping Li/Documents/csc465/presentation/loan.csv", header = T)
head(data) ds <-
data.frame(id=data$id,amount=data$loan_amnt,title=data$title,purpose=data$purpose,state=data
$addr_state,desc=data$desc,status=data$loan_status)


#a>---------- library("tm")
library("SnowballC") library("wordcloud")
library("RColorBrewer") ds = ds[grepl("Late",
dsBig$status),] ds1 = ds[grepl("Paid", ds$status),] ds2
= ds[grepl("Off", ds$status),] paidCharged =
rbind(ds1,ds2) # Read the text file from internet
purposes <- as.character(ds$purpose) lateTitle <-
as.character(ds$title) paidChargedTitle <-
as.character(paidCharged$title) # We make a "text
corpus" out of the text docs =
Corpus(VectorSource(purposes)) docs =
Corpus(VectorSource(lateTitle)) docs =
Corpus(VectorSource(paidChargedTitle)) head(docs)
inspect(docs)

# Replace any special characters with " "  toSpace =
content_transformer(function (x , pattern ) gsub(pattern, " ", x))
docs = tm_map(docs, toSpace, "/") docs = tm_map(docs, toSpace, "@") docs
= tm_map(docs, toSpace, "\\|")
```

```
# Convert the text to lower case docs = tm_map(docs,
content_transformer(tolower)) # Remove numbers
docs = tm_map(docs, removeNumbers) # Remove
english common stopwords docs = tm_map(docs,
removeWords, stopwords("english")) # Remove
punctuations docs <- tm_map(docs,
removePunctuation) # Eliminate extra white spaces
docs <- tm_map(docs, stripWhitespace)

# Build a term matrix dtm =
TermDocumentMatrix(docs) m =
as.matrix(dtm) v =
sort(rowSums(m),decreasing=TRUE) d =
data.frame(word = names(v),freq=v) d$logFreq
= log2(d$freq)

set.seed(1234) wordcloud(words = d$word, freq =
d$logFreq, min.freq = 1,          max.words=200,
random.order=FALSE, rot.per=0.35,
colors=brewer.pal(8, "Dark2")) wordcloud(words = d$word,
freq = d$freq, min.freq = 1,          max.words=200,
random.order=FALSE, rot.per=0.35,
colors=brewer.pal(8, "Dark2"))

#b)--------Loan Amount Total----------

library(maps) library(mapdata) library(ggplot2) library(mapproj) states_50 <-
read.csv("C:/Users/Yingping Li/Documents/csc465/presentation/states_50.csv", header  = T)
states_50$stateA <- tolower(states_50$state) states_map = map_data("state")
states_map$regionAbrv = states_50$stateAbrv[match(states_map$region, states_50$stateA)]

loan_map = states_map ls = aggregate(ds$amount,
by=list(state=ds$state), FUN=sum) loan_map$loan_amnt =
ls$x[match(loan_map$regionAbrv, ls$state)] loan_map$lg_amnt
= log10(loan_map$loan_amnt) loan_map$log2_amnt =
log2(loan_map$loan_amnt) library(plyr)

ggplot(loan_map, aes(x=long, y=lat, group=group, fill=lg_amnt)) +
```

```
geom_polygon(color="black") + scale_fill_gradient(low='#b3e0ff', high='#006bb3') +
coord_map("mercator")

#d) late loan dsBig
<-
data.frame(id=data$id,amount=data$loan_amnt,state=data$addr_state,status=as.character(data$lo
an_status),term=data$term,                int_rate=data$int_rate,installment=data$installment,
grade=data$grade, emp_length=data$emp_length,home_ownership=data$home_ownership
                ,annual_inc=data$annual_inc) ds = dsBig[grepl("Late", dsBig$status),] data_state_sum <-
aggregate(data$loan_amnt, by=list(state=data$addr_state), FUN=sum) states_50 <-
read.csv("C:/Users/Yingping Li/Documents/csc465/presentation/states_50.csv", header  = T)
states_50$stateA <- tolower(states_50$state) states_map = map_data("state")
states_map$regionAbrv = states_50$stateAbrv[match(states_map$region, states_50$stateA)]

loan_map_late = states_map ls = aggregate(ds$amount,
by=list(state=ds$state), FUN=sum) loan_map_late$loan_amnt = ls$x[match(loan_map$regionAbrv,
ls$state)]

loan_map_late$amount_sum = data_state_sum$x[match(loan_map_late$regionAbrv,
data_state_sum$state)]
loan_map_late$loan_rate = loan_map_late$loan_amnt/loan_map_late$amount_sum
loan_map_late$loan_rate_log2 = log2(loan_map_late$loan_rate) loan_map_late$lg_amnt
= log10(loan_map$loan_amnt) loan_map_late$late_rate = log2(loan_map$loan_amnt)
library(plyr)

ggplot(loan_map_late, aes(x=long, y=lat, group=group, fill=lg_amnt)) +
geom_polygon(color="black")+ scale_fill_gradient(low='#b3d9ff', high='#004080') +
coord_map("mercator")


ggplot(loan_map_late, aes(x=long, y=lat, group=group, fill=loan_rate)) +
geom_polygon(color="black") + scale_fill_gradient(low='#ffd699', high='#b36b00') +
coord_map("mercator")

ggplot(loan_map_late, aes(x=long, y=lat, group=group, fill=loan_rate_log2)) +
geom_polygon(color="black") + scale_fill_gradient(low='#ffd699', high='#b36b00') +
coord_map("mercator")
```

```
library(stringr) numextract <- function(string){
str_extract(string,
"\\-*\\d+\\.*\\d*")
}
```