Installing the packages:

```
library('choroplethr')
library('choroplethrMaps')
install.packages('tm')
install.packages('RColorBrewer')
install.packages('wordcloud')
install.packages("choroplethrMaps")
library(tm)
library(RColorBrewer)
library(wordcloud)
install.packages('ggplot2')
library(ggplot2)
library(wordcloud)
install.packages('dplyr')
library(dplyr)
```

Reading the csv file and selecting the columns:

```
loan <- read.csv('loan.csv')
loan = loan %>%
  select(loan_status , loan_amnt , funded_amnt, installment, int_rate, issue_d , grade ,
sub_grade, emp_title , purpose, addr_state, dti, emp_length , home_ownership ,
       annual_inc , term, )
str(loan)
loan
loan<- na.omit(loan)
```

Summary for the Dataset

```
summary(loan)
```

```
           loan_status        loan_amnt      funded_amnt     installment         int_rate         issue_d
Fully Paid        :1041952   Min.   :  500   Min.   :  500   Min.   :   4.93   Min.   : 5.31   Mar-2016:  61992
Current           : 919695   1st Qu.: 8000   1st Qu.: 8000   1st Qu.: 251.65   1st Qu.: 9.49   Oct-2015:  48631
Charged Off       : 261655   Median :12900   Median :12875   Median : 377.99   Median :12.62   May-2018:  46311
Late (31-120 days):  21897   Mean   :15047   Mean   :15042   Mean   : 445.81   Mean   :13.09   Oct-2018:  46305
In Grace Period   :   8952   3rd Qu.:20000   3rd Qu.:20000   3rd Qu.: 593.32   3rd Qu.:15.99   Aug-2018:  46079
Late (16-30 days) :   3737   Max.   :40000   Max.   :40000   Max.   :1719.83   Max.   :30.99   Jul-2015:  45962
(Other)           :   2780                                                                     (Other) :1965388
grade        sub_grade                emp_title                    purpose          addr_state          dti
A:433027   C1     : 145903                  : 166931   debt_consolidation:1277877   CA     : 314533   Min.   : -1.00
B:663557   B5     : 140288   Teacher        :  38824   credit_card       : 516971   NY     : 186389   1st Qu.: 11.89
C:650053   B4     : 139793   Manager        :  34298   home_improvement  : 150457   TX     : 186335   Median : 17.84
D:324424   B3     : 131514   Owner          :  21977   other             : 139440   FL     : 161991   Mean   : 18.82
E:135639   C2     : 131116   Registered Nurse:  15867   major_purchase    :  50445   IL     :  91173   3rd Qu.: 24.49
F: 41800   C3     : 129193   (Other)        :1982761   medical           :  27488   NJ     :  83132   Max.   :999.00
G: 12168   (Other):1442861   NA's           :     10   (Other)           :  97990   (Other):1237115   NA's   :1711
    emp_length      home_ownership      annual_inc             term
10+ years:748005   ANY     :    996   Min.   :        0   36 months:1609754
2 years  :203677   MORTGAGE:1111450   1st Qu.:    46000   60 months: 650914
< 1 year :189988   NONE    :     54   Median :    65000
3 years  :180753   OTHER   :    182   Mean   :    77992
1 year   :148403   OWN     : 253057   3rd Qu.:    93000
n/a      :146907   RENT    : 894929   Max.   :110000000
(Other)  :642935                      NA's   :4
```

Number of Rows and Columns

NROW(loan)

NCOL(loan)

```
> NROW(loan)
[1] 2260668
> NCOL(loan)
[1] 16
```

Column Names:

colnames(loan)

```
 [1] "loan_status"    "loan_amnt"      "funded_amnt"    "installment"    "int_rate"       "issue_d"
 [7] "grade"          "sub_grade"      "emp_title"      "purpose"        "addr_state"     "dti"
[13] "emp_length"     "home_ownership" "annual_inc"     "term"
```

Dimensions of the Dataset

dim(loan)

```
[1] 2260668      16
```

First and Last 6 rows of the Dataset

head(loan)

	loan_status	loan_amnt	funded_amnt	installment	int_rate	issue_d	grade	sub_grade	emp_title	purpose	addr_state	dti	emp_length	home_ownership	annual_inc	term
1	Current	2500	2500	84.92	13.56	Dec-2018	C	C1	Chef	debt_consolidation	NY	18.24	10+ years	RENT	55000	36 months
2	Current	30000	30000	777.23	18.94	Dec-2018	D	D2	Postmaster	debt_consolidation	LA	26.52	10+ years	MORTGAGE	90000	60 months
3	Current	5000	5000	180.69	17.97	Dec-2018	D	D1	Administrative	debt_consolidation	MI	10.51	6 years	MORTGAGE	59280	36 months
4	Current	4000	4000	146.51	18.94	Dec-2018	D	D2	IT Supervisor	debt_consolidation	WA	16.74	10+ years	MORTGAGE	92000	36 months
5	Current	30000	30000	731.78	16.14	Dec-2018	C	C4	Mechanic	debt_consolidation	MD	26.35	10+ years	MORTGAGE	57250	60 months
6	Current	5550	5550	192.45	15.02	Dec-2018	C	C3	Director COE	credit_card	IN	37.94	10+ years	MORTGAGE	152500	36 months

tail(loan)

	loan_status	loan_amnt	funded_amnt	installment	int_rate	issue_d	grade	sub_grade	emp_title	purpose	addr_state	dti	emp_length	home_ownership	annual_inc	term
2260663	Current	30000	30000	984.47	30.75	Oct-2017	F	F5	Sales	debt_consolidation	FL	22.13	6 years	MORTGAGE	100000	60 months
2260664	Current	12000	12000	279.72	14.08	Oct-2017	C	C3	house keeper	debt_consolidation	VT	20.88	10+ years	MORTGAGE	58000	60 months
2260665	Fully Paid	12000	12000	358.01	25.82	Oct-2017	E	E4	Skilled Labor	debt_consolidation	OR	19.28	< 1 year	MORTGAGE	30000	60 months
2260666	Current	10000	10000	332.10	11.99	Oct-2017	B	B5	Teacher	debt_consolidation	IL	12.96	10+ years	OWN	64000	36 months
2260667	Current	12000	12000	327.69	21.45	Oct-2017	D	D5		debt_consolidation	AK	30.82	n/a	RENT	60000	60 months
2260668	Current	16550	16550	451.94	21.45	Oct-2017	D	D5	BABYSITTER	credit_card	NY	18.40	3 years	RENT	60000	60 months

Data for top 4 states:

stateCA = loan[loan$addr_state== 'CA',]

stateCA

	loan_status	loan_amnt	funded_amnt	installment	int_rate	issue_d	grade	sub_grade	emp_title	purpose	addr_state	dti	emp_length	home_ownership
21	Current	10000	10000	339.65	13.56	Dec-2018	C	C1		credit_card	CA	10.62	< 1 year	MORTGAGE
57	Current	14400	14400	338.59	14.47	Dec-2018	C	C2	Director	debt_consolidation	CA	10.83	3 years	RENT
65	Current	20000	20000	679.29	13.56	Dec-2018	C	C1	Lead Transportation Security Officer	debt_consolidation	CA	9.06	10+ years	MORTGAGE
66	Current	29450	29450	925.44	8.19	Dec-2018	A	A4	General Manager	debt_consolidation	CA	10.99	10+ years	MORTGAGE
69	Current	15500	15500	505.40	10.72	Dec-2018	B	B2	Administrative Assistant	credit_card	CA	16.66	10+ years	RENT
80	Current	10250	10250	334.22	10.72	Dec-2018	B	B2	Teacher	debt_consolidation	CA	8.31	10+ years	MORTGAGE

stateNY = loan[loan$addr_state== 'NY',]

stateNY

	loan_status	loan_amnt	funded_amnt	installment	int_rate	issue_d	grade	sub_grade	emp_title	purpose	addr_state	dti	emp_length	home_ownership
1	Current	2500	2500	84.92	13.56	Dec-2018	C	C1	Chef	debt_consolidation	NY	18.24	10+ years	RENT
38	Current	10000	10000	376.19	20.89	Dec-2018	D	D4	Foreman shop	credit_card	NY	11.21	7 years	RENT
70	Current	12000	12000	397.43	11.80	Dec-2018	B	B4	Transportation Manager	debt_consolidation	NY	29.70	10+ years	RENT
73	Current	3000	3000	108.42	17.97	Dec-2018	D	D1	IT Assistant	debt_consolidation	NY	30.03	10+ years	RENT
78	Current	3000	3000	104.03	15.02	Dec-2018	C	C3	Banker	home_improvement	NY	21.36	4 years	MORTGAGE
106	Current	6000	6000	206.44	14.47	Dec-2018	C	C2	Manager	credit_card	NY	10.65	10+ years	RENT
107	Current	5000	5000	168.43	12.98	Dec-2018	B	B5	Consultant	debt_consolidation	NY	2.39	2 years	RENT

stateTX = loan[loan$addr_state== 'TX',]

stateTX

	loan_status	loan_amnt	funded_amnt	installment	int_rate	issue_d	grade	sub_grade	emp_title	purpose	addr_state	dti	emp_length	home_ownership	annual_inc
26	Current	15000	15000	352.69	14.47	Dec-2018	C	C2		debt_consolidation	TX	41.60	n/a	MORTGAGE	30000
39	Current	6000	6000	183.79	6.46	Dec-2018	A	A1	Substance Abuse Counselor	credit_card	TX	12.87	1 year	MORTGAGE	40000
67	Current	34625	34625	844.60	16.14	Dec-2018	C	C4	Vice President	debt_consolidation	TX	27.88	10+ years	MORTGAGE	137500
72	Current	30000	30000	1056.79	16.14	Dec-2018	C	C4	Structural Engineer	debt_consolidation	TX	28.21	7 years	MORTGAGE	116000
83	Current	24000	24000	890.95	19.92	Dec-2018	D	D3	overnight supervisor	credit_card	TX	67.68	7 years	RENT	30000
125	Current	3000	3000	97.82	10.72	Dec-2018	B	B2		credit_card	TX	24.20	n/a	MORTGAGE	30000

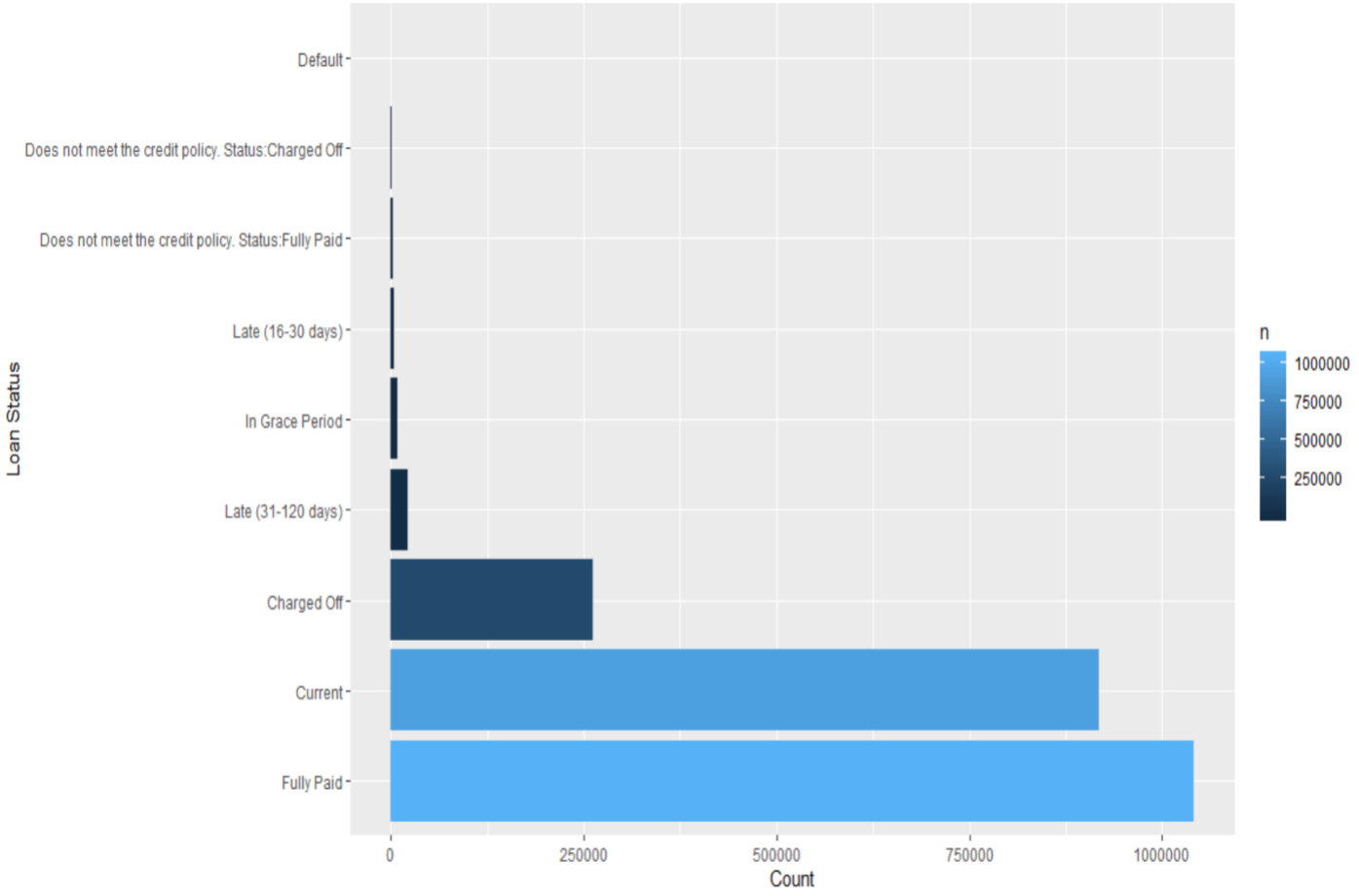
stateFL = loan[loan$addr_state== 'FL',]

stateFL

	loan_status	loan_amnt	funded_amnt	installment	int_rate	issue_d	grade	sub_grade	emp_title	purpose	addr_state	dti	emp_length	home_ownership	annual_inc
9	Current	5000	5000	180.69	17.97	Dec-2018	D	D1	Legal Assistant III	debt_consolidation	FL	21.16	10+ years	MORTGAGE	53580
110	Current	22650	22650	738.54	10.72	Dec-2018	B	B2	President	home_improvement	FL	17.06	7 years	MORTGAGE	145000
119	Current	5525	5525	191.59	15.02	Dec-2018	C	C3	Vice President	car	FL	3.25	10+ years	RENT	118000
120	Current	2525	2525	89.92	16.91	Dec-2018	C	C5	Tech Driver	credit_card	FL	15.89	1 year	RENT	21000
157	Current	12000	12000	422.72	16.14	Dec-2018	C	C4	Sales Account Manager	credit_card	FL	18.96	4 years	RENT	70000
159	Current	4000	4000	142.44	16.91	Dec-2018	C	C5	Vacation Advisor	home_improvement	FL	9.64	5 years	RENT	54000
160	Current	32000	32000	727.78	12.98	Dec-2018	B	B5	Manager	debt_consolidation	FL	20.17	4 years	MORTGAGE	70000

Horizontal stacked bar graph for Total Number for loan_status labels

```
loan %>%
  count(loan_status) %>%
  ggplot(aes(x = reorder(loan_status , desc(n)) , y = n , fill = n)) +
  geom_col() +
  coord_flip() +
  labs(x = 'Loan Status' , y = 'Count')
```

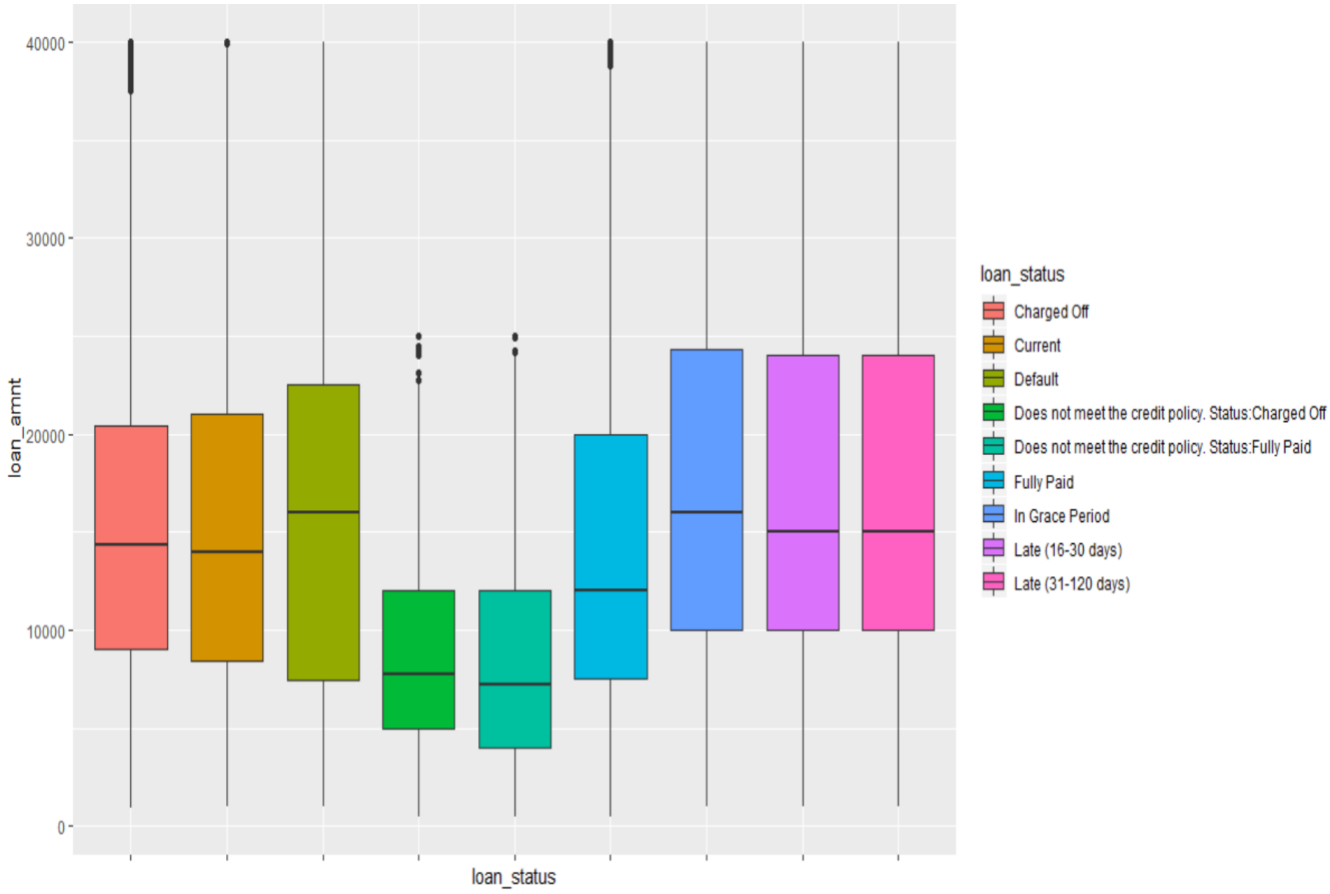


Word-Cloud for Loan-Purpose

```
text <- Corpus(VectorSource(loan$purpose))
text <- tm_map(text, tolower) # might simplify the vector
text <- tm_map(text, removePunctuation)
tdm <- TermDocumentMatrix(text) # turn it into a term document matrix
m <- as.matrix(tdm)
freq <- sort(rowSums(m), decreasing = TRUE)
wordcloud(words = names(freq), freq = freq, min.freq = 1, random.order = FALSE,
        col=brewer.pal(8, "Paired"), max.words=10000000, rot.per=.30)
```

Box-Plot for Loan Amount vs Loan Status

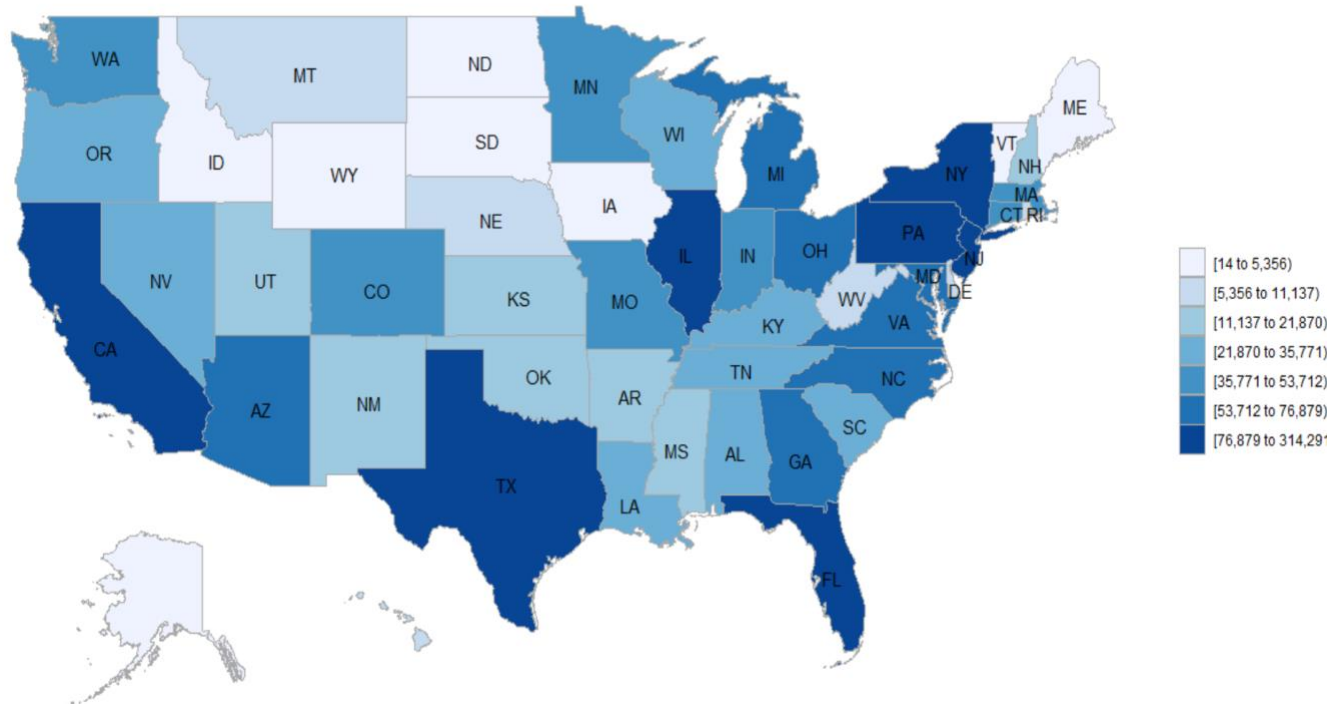
```
box_status <- ggplot(loan, aes(loan_status, loan_amnt))
box_status + geom_boxplot(aes(fill = loan_status)) +
  theme(axis.text.x = element_blank()) +
  labs(list(
    title = "Loan amount by status",
    x = "Status",
    y = "Amount")) + scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```



Heat Map for states

```
loan$region <- loan$addr_state
loan$region <- as.factor(loan$region)
levels(loan$region)<- c("alaska", "alabama","arkansas", "arizona",
"california","colorado","connecticut","district of
columbia","delaware","florida","georgia","hawaii","iowa","idaho","illinois","indiana","kansas","
kentucky","louisiana","massachusetts","maryland","maine","michigan","minnesota","missouri"
,"mississippi","montana","north carolina","north dakota","nebraska","new hampshire","new
jersey","new mexico","nevada","new york","ohio","oklahoma","oregon","pennsylvania","rhode
island","south carolina","south
dakota","tennessee","texas","utah","virginia","vermont","washington","wisconsin","west
virginia","wyoming")
state_by_volume <-
  loan %>% group_by(region) %>%
  summarise(value = n())
state_choropleth(state_by_volume, title = "Volume by State")
```

Volume by State



calculating the covariance between variables

```
measure <-structure(list(v1 = 1:25, v2 = c(2500L, 30000L, 5000L, 4000L, 30000L, 5550L,
                                           2000L, 6000L, 5000L, 6000L, 5500L, 28000L, 11200L, 6500L, 22000L,
3500L, 7000L, 25000L, 16000L,
                                           13000L,10000L,13000L,9600L,3500L,16000L),
                    v3 = c(13.56, 18.94, 17.97, 18.94, 16.14, 15.02, 17.97, 13.56, 17.97, 14.47, 22.35,
11.31, 8.19, 17.97, 12.98, 16.14, 12.98, 16.91, 20.89, 14.47,13.56,14.47,23.4,20.89,26.31),
                    v4 = c(18.24,326.52, 10.51, 16.74, 26.35, 37.94, 2.4, 30.1, 21.16, 17.43, 15.94,
22.01, 23.6, 28.78, 11.19, 13.63,15.2, 6.26, 27.57, 26.16,10.62,10.58,23.01,9.09,33.62),
                    v5=c(55000L, 90000L, 59280L, 92000L,
57250L,152500L,51000L,65000L,53580L,300000L,50000L,70000L,65000L,154000L,65000L,8000
0L,102500L,23878L,120000L,75000L,65000L,55000L,65000L,40000L,33000L)), .Names = c("V1",
"V2", "V3","V4","V5"), class = "data.frame", row.names = c(NA, -25L))
measure <- measure[,-1]
names(measure) <- c("loan_amnt", "int_rate", "dti","annual_inc")
x <- dist(scale(measure[, c("loan_amnt", "int_rate", "annual_inc")],
          center = FALSE))
as.dist(round(as.matrix(x), 2)[1:12, 1:12])
```

```
       1    2    3    4    5    6    7    8    9   10   11
2  1.93
3  0.31 1.73
4  0.49 1.77 0.34
5  1.88 0.36 1.70 1.81
6  1.00 1.79 0.95 0.65 1.92
7  0.26 1.95 0.22 0.44 1.91 1.06
8  0.26 1.68 0.27 0.43 1.64 0.88 0.40
9  0.30 1.74 0.06 0.39 1.70 1.00 0.21 0.28
10 2.46 2.67 2.41 2.10 2.92 1.47 2.51 2.35 2.47
11 0.54 1.72 0.27 0.47 1.71 1.11 0.35 0.52 0.25 2.54
12 1.75 0.50 1.61 1.70 0.33 1.75 1.82 1.50 1.62 2.75 1.67
```

x <- measure[, c("loan_amnt", "int_rate", "dti","annual_inc")]
x

```
   loan_amnt int_rate    dti annual_inc
1       2500    13.56  18.24      55000
2      30000    18.94 326.52      90000
3       5000    17.97  10.51      59280
4       4000    18.94  16.74      92000
5      30000    16.14  26.35      57250
6       5550    15.02  37.94     152500
7       2000    17.97   2.40      51000
8       6000    13.56  30.10      65000
9       5000    17.97  21.16      53580
10      6000    14.47  17.43     300000
11      5500    22.35  15.94      50000
12     28000    11.31  22.01      70000
13     11200     8.19  23.60      65000
14      6500    17.97  28.78     154000
15     22000    12.98  11.19      65000
16      3500    16.14  13.63      80000
17      7000    12.98  15.20     102500
18     25000    16.91   6.26      23878
19     16000    20.89  27.57     120000
20     13000    14.47  26.16      75000
21     10000    13.56  10.62      65000
22     13000    14.47  10.58      55000
23      9600    23.40  23.01      65000
24      3500    20.89   9.09      40000
25     16000    26.31  33.62      33000
```

cm <- colMeans(x) #calculating the means of the columns
S <- cov(x)
S

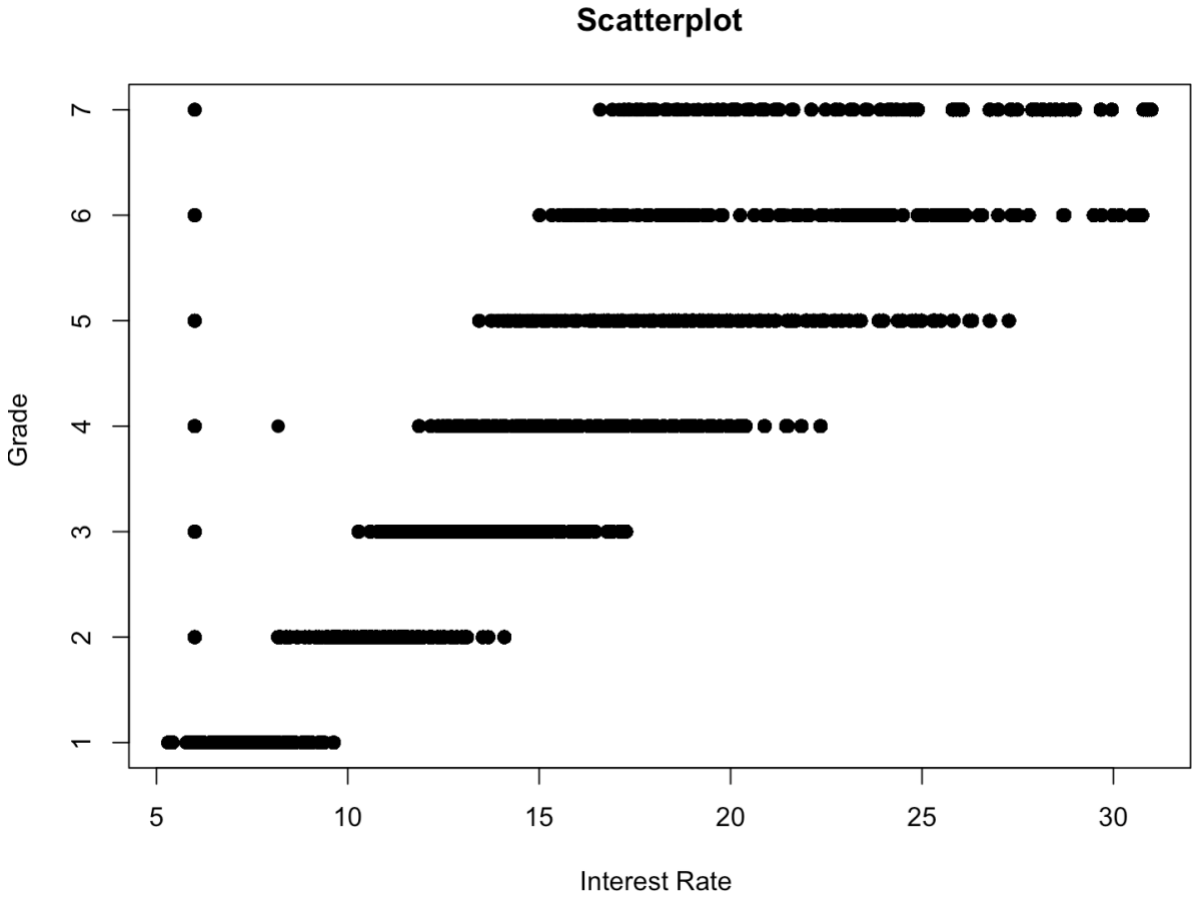
```
              loan_amnt     int_rate          dti    annual_inc
loan_amnt   79885150.00  -3095.01000 246398.53750  -89485991.33
int_rate       -3095.01     16.57929     31.90208     -37211.88
dti           246398.54     31.90208   3859.38808     234743.00
annual_inc -89485991.33 -37211.87863 234743.00050 3094428067.43
```

```
d <- apply(x, MARGIN = 1, function(x)t(x - cm) %*% solve(S) %*% (x - cm))
d
```

```
 [1]  2.8145026 22.6141713  0.8347310  0.9425539  5.6640908  1.9819368  1.6995918
 [8]  1.7032559  1.0185573 15.8652761  2.4525805  5.5980117  5.0747127  2.0582156
[15]  2.7264311  0.9343421  1.1899459  4.1740563  3.0143459  0.3528678  0.8812065
[22]  0.6957913  2.8435209  2.2832068  6.5820974
```

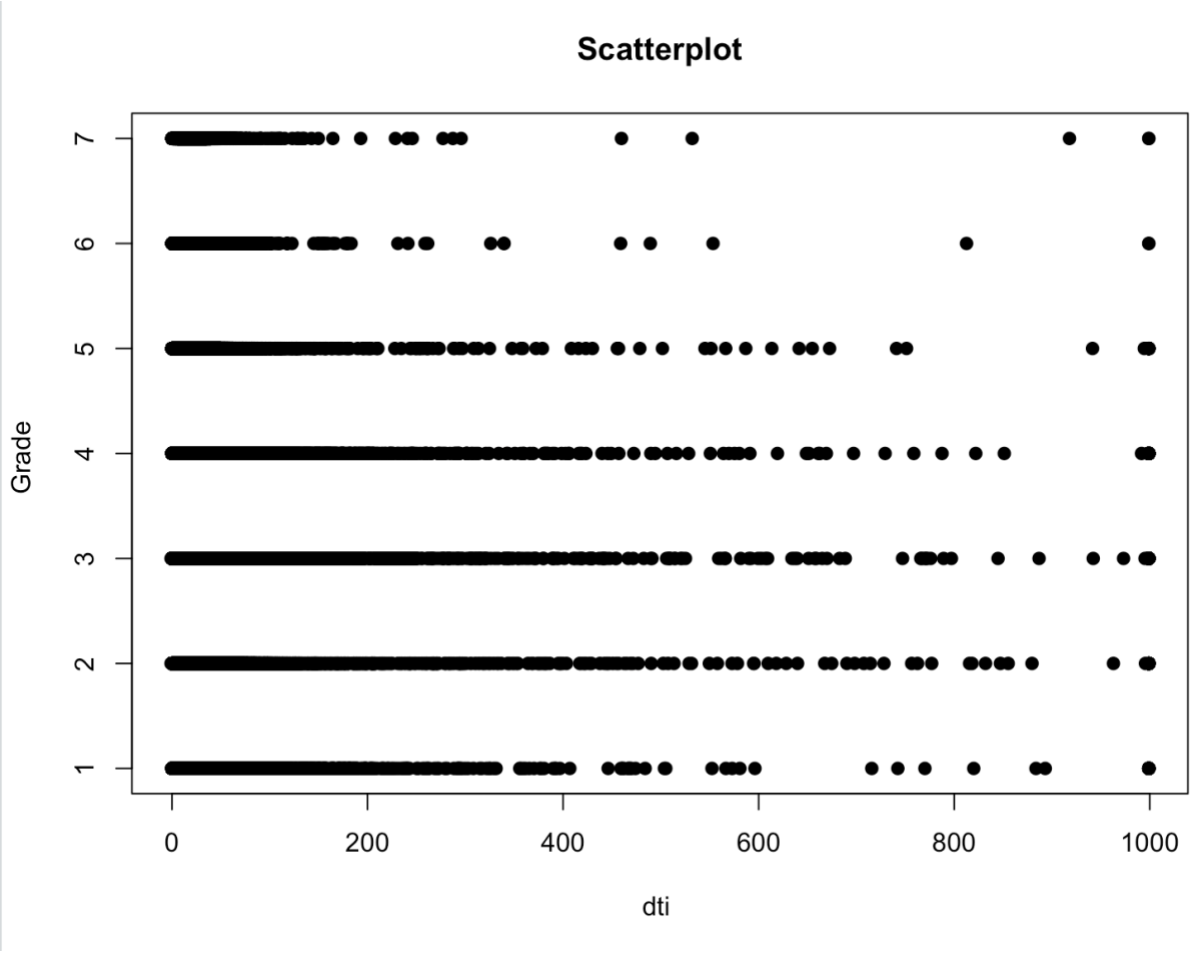
Scatterplot: Interest Rate vs Grade

```
plot(int_rate, grade, main="Scatterplot",
xlab="Interest Rate ", ylab="Grade ", pch=19)
```



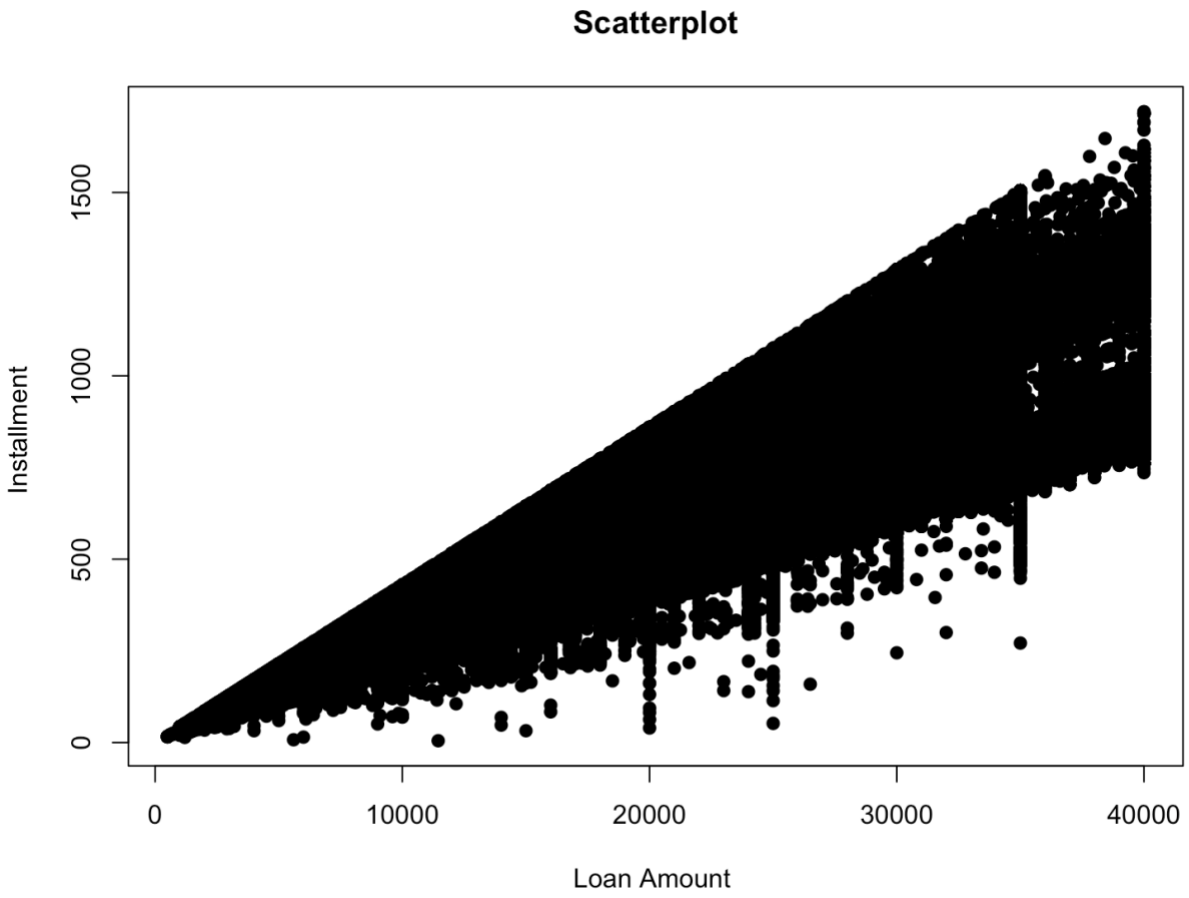
Scatterplot: dti vs Grade

```
plot(dti, grade, main="Scatterplot",
xlab=”dti ", ylab="Grade ", pch=19)
```



Scatterplot: Loan Amount vs installment

plot(loan_amnt, installment, main="Scatterplot",
xlab=" Loan Amount ", ylab="Installment ", pch=19)


Scatterplot
Installment
0
500
1000
1500
0
10000
20000
30000
40000
Loan Amount