LoanRiskAnalysis

Liang Tan

Read data

```
loan <- read.csv("loan.csv", stringsAsFactors = FALSE)
loanT <- loan
head(loan)
```

```
##        id member_id loan_amnt funded_amnt funded_amnt_inv       term
## 1 1077501   1296599      5000        5000            4975  36 months
## 2 1077430   1314167      2500        2500            2500  60 months
## 3 1077175   1313524      2400        2400            2400  36 months
## 4 1076863   1277178     10000       10000           10000  36 months
## 5 1075358   1311748      3000        3000            3000  60 months
## 6 1075269   1311441      5000        5000            5000  36 months
##   int_rate installment grade sub_grade                 emp_title emp_length
## 1    10.65      162.87     B        B2                              10+ years
## 2    15.27       59.83     C        C4                     Ryder    < 1 year
## 3    15.96       84.33     C        C5                              10+ years
## 4    13.49      339.31     C        C1      AIR RESOURCES BOARD     10+ years
## 5    12.69       67.79     B        B5 University Medical Group       1 year
## 6     7.90      156.46     A        A4     Veolia Transportaton       3 years
##   home_ownership annual_inc verification_status  issue_d loan_status
## 1           RENT      24000            Verified Dec-2011  Fully Paid
## 2           RENT      30000     Source Verified Dec-2011 Charged Off
## 3           RENT      12252        Not Verified Dec-2011  Fully Paid
## 4           RENT      49200     Source Verified Dec-2011  Fully Paid
## 5           RENT      80000     Source Verified Dec-2011     Current
## 6           RENT      36000     Source Verified Dec-2011  Fully Paid
##   pymnt_plan
## 1          n
## 2          n
## 3          n
## 4          n
## 5          n
## 6          n
##                                                                  url
## 1 https://www.lendingclub.com/browse/loanDetail.action?loan_id=1077501
## 2 https://www.lendingclub.com/browse/loanDetail.action?loan_id=1077430
## 3 https://www.lendingclub.com/browse/loanDetail.action?loan_id=1077175
## 4 https://www.lendingclub.com/browse/loanDetail.action?loan_id=1076863
## 5 https://www.lendingclub.com/browse/loanDetail.action?loan_id=1075358
## 6 https://www.lendingclub.com/browse/loanDetail.action?loan_id=1075269
##
## 1
## 2   Borrower added on 12/22/11 > I plan to use this money to finance the motorcycle i am looking at.
## 3
## 4
## 5
## 6
```

```
##          purpose                                   title zip_code addr_state
## 1    credit_card                                Computer    860xx         AZ
## 2            car                                    bike    309xx         GA
## 3 small_business                    real estate business    606xx         IL
## 4          other                                personel    917xx         CA
## 5          other                                Personal    972xx         OR
## 6        wedding My wedding loan I promise to pay back    852xx         AZ
##     dti delinq_2yrs earliest_cr_line inq_last_6mths mths_since_last_delinq
## 1 27.65           0         Jan-1985              1                      NA
## 2  1.00           0         Apr-1999              5                      NA
## 3  8.72           0         Nov-2001              2                      NA
## 4 20.00           0         Feb-1996              1                      35
## 5 17.94           0         Jan-1996              0                      38
## 6 11.20           0         Nov-2004              3                      NA
##   mths_since_last_record open_acc pub_rec revol_bal revol_util total_acc
## 1                     NA        3       0     13648       83.7         9
## 2                     NA        3       0      1687        9.4         4
## 3                     NA        2       0      2956       98.5        10
## 4                     NA       10       0      5598       21.0        37
## 5                     NA       15       0     27783       53.9        38
## 6                     NA        9       0      7963       28.3        12
##   initial_list_status out_prncp out_prncp_inv total_pymnt total_pymnt_inv
## 1                   f       0.0           0.0    5861.071         5831.78
## 2                   f       0.0           0.0    1008.710         1008.71
## 3                   f       0.0           0.0    3003.654         3003.65
## 4                   f       0.0           0.0   12226.302        12226.30
## 5                   f     766.9         766.9    3242.170         3242.17
## 6                   f       0.0           0.0    5631.378         5631.38
##   total_rec_prncp total_rec_int total_rec_late_fee recoveries
## 1         5000.00        861.07               0.00       0.00
## 2          456.46        435.17               0.00     117.08
## 3         2400.00        603.65               0.00       0.00
## 4        10000.00       2209.33              16.97       0.00
## 5         2233.10       1009.07               0.00       0.00
## 6         5000.00        631.38               0.00       0.00
##   collection_recovery_fee last_pymnt_d last_pymnt_amnt next_pymnt_d
## 1                    0.00     Jan-2015          171.62
## 2                    1.11     Apr-2013          119.66
## 3                    0.00     Jun-2014          649.91
## 4                    0.00     Jan-2015          357.48
## 5                    0.00     Jan-2016           67.79     Feb-2016
## 6                    0.00     Jan-2015          161.03
##   last_credit_pull_d collections_12_mths_ex_med
## 1           Jan-2016                         0
## 2           Sep-2013                         0
## 3           Jan-2016                         0
## 4           Jan-2015                         0
## 5           Jan-2016                         0
## 6           Sep-2015                         0
##   mths_since_last_major_derog policy_code application_type
## 1                          NA           1       INDIVIDUAL
## 2                          NA           1       INDIVIDUAL
## 3                          NA           1       INDIVIDUAL
## 4                          NA           1       INDIVIDUAL
```

```
## 5                          NA           1       INDIVIDUAL
## 6                          NA           1       INDIVIDUAL
##   annual_inc_joint dti_joint verification_status_joint acc_now_delinq
## 1               NA        NA                                       0
## 2               NA        NA                                       0
## 3               NA        NA                                       0
## 4               NA        NA                                       0
## 5               NA        NA                                       0
## 6               NA        NA                                       0
##   tot_coll_amt tot_cur_bal open_acc_6m open_il_6m open_il_12m open_il_24m
## 1           NA          NA          NA         NA          NA          NA
## 2           NA          NA          NA         NA          NA          NA
## 3           NA          NA          NA         NA          NA          NA
## 4           NA          NA          NA         NA          NA          NA
## 5           NA          NA          NA         NA          NA          NA
## 6           NA          NA          NA         NA          NA          NA
##   mths_since_rcnt_il total_bal_il il_util open_rv_12m open_rv_24m
## 1                 NA           NA      NA          NA          NA
## 2                 NA           NA      NA          NA          NA
## 3                 NA           NA      NA          NA          NA
## 4                 NA           NA      NA          NA          NA
## 5                 NA           NA      NA          NA          NA
## 6                 NA           NA      NA          NA          NA
##   max_bal_bc all_util total_rev_hi_lim inq_fi total_cu_tl inq_last_12m
## 1         NA       NA               NA     NA          NA           NA
## 2         NA       NA               NA     NA          NA           NA
## 3         NA       NA               NA     NA          NA           NA
## 4         NA       NA               NA     NA          NA           NA
## 5         NA       NA               NA     NA          NA           NA
## 6         NA       NA               NA     NA          NA           NA
```

Check dimension

```
print(dim(loan))
```

```
## [1] 887379     74
```

Check data format

```
str(loan)
```

```
## 'data.frame':    887379 obs. of  74 variables:
##  $ id                          : int  1077501 1077430 1077175 1076863 1075358 1075269 1069639 1072053
##  $ member_id                   : int  1296599 1314167 1313524 1277178 1311748 1311441 1304742 1288686
##  $ loan_amnt                   : num  5000 2500 2400 10000 3000 ...
##  $ funded_amnt                 : num  5000 2500 2400 10000 3000 ...
##  $ funded_amnt_inv             : num  4975 2500 2400 10000 3000 ...
##  $ term                        : chr  " 36 months" " 60 months" " 36 months" " 36 months" ...
##  $ int_rate                    : num  10.7 15.3 16 13.5 12.7 ...
##  $ installment                 : num  162.9 59.8 84.3 339.3 67.8 ...
##  $ grade                       : chr  "B" "C" "C" "C" ...
##  $ sub_grade                   : chr  "B2" "C4" "C5" "C1" ...
##  $ emp_title                   : chr  "" "Ryder" "" "AIR RESOURCES BOARD" ...
##  $ emp_length                  : chr  "10+ years" "< 1 year" "10+ years" "10+ years" ...
##  $ home_ownership              : chr  "RENT" "RENT" "RENT" "RENT" ...
##  $ annual_inc                  : num  24000 30000 12252 49200 80000 ...
```

```
##  $ verification_status       : chr  "Verified" "Source Verified" "Not Verified" "Source Verified" .
##  $ issue_d                   : chr  "Dec-2011" "Dec-2011" "Dec-2011" "Dec-2011" ...
##  $ loan_status               : chr  "Fully Paid" "Charged Off" "Fully Paid" "Fully Paid" ...
##  $ pymnt_plan                : chr  "n" "n" "n" "n" ...
##  $ url                       : chr  "https://www.lendingclub.com/browse/loanDetail.action?loan_id=1
##  $ desc                      : chr  "  Borrower added on 12/22/11 > I need to upgrade my business t
##  $ purpose                   : chr  "credit_card" "car" "small_business" "other" ...
##  $ title                     : chr  "Computer" "bike" "real estate business" "personel" ...
##  $ zip_code                  : chr  "860xx" "309xx" "606xx" "917xx" ...
##  $ addr_state                : chr  "AZ" "GA" "IL" "CA" ...
##  $ dti                       : num  27.65 1 8.72 20 17.94 ...
##  $ delinq_2yrs               : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ earliest_cr_line          : chr  "Jan-1985" "Apr-1999" "Nov-2001" "Feb-1996" ...
##  $ inq_last_6mths            : num  1 5 2 1 0 3 1 2 2 0 ...
##  $ mths_since_last_delinq    : num  NA NA NA 35 38 NA NA NA NA NA ...
##  $ mths_since_last_record    : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ open_acc                  : num  3 3 2 10 15 9 7 4 11 2 ...
##  $ pub_rec                   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ revol_bal                 : num  13648 1687 2956 5598 27783 ...
##  $ revol_util                : num  83.7 9.4 98.5 21 53.9 28.3 85.6 87.5 32.6 36.5 ...
##  $ total_acc                 : num  9 4 10 37 38 12 11 4 13 3 ...
##  $ initial_list_status       : chr  "f" "f" "f" "f" ...
##  $ out_prncp                 : num  0 0 0 0 767 ...
##  $ out_prncp_inv             : num  0 0 0 0 767 ...
##  $ total_pymnt               : num  5861 1009 3004 12226 3242 ...
##  $ total_pymnt_inv           : num  5832 1009 3004 12226 3242 ...
##  $ total_rec_prncp           : num  5000 456 2400 10000 2233 ...
##  $ total_rec_int             : num  861 435 604 2209 1009 ...
##  $ total_rec_late_fee        : num  0 0 0 17 0 ...
##  $ recoveries                : num  0 117 0 0 0 ...
##  $ collection_recovery_fee   : num  0 1.11 0 0 0 0 0 0 2.09 2.52 ...
##  $ last_pymnt_d              : chr  "Jan-2015" "Apr-2013" "Jun-2014" "Jan-2015" ...
##  $ last_pymnt_amnt           : num  171.6 119.7 649.9 357.5 67.8 ...
##  $ next_pymnt_d              : chr  "" "" "" "" ...
##  $ last_credit_pull_d        : chr  "Jan-2016" "Sep-2013" "Jan-2016" "Jan-2015" ...
##  $ collections_12_mths_ex_med : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ mths_since_last_major_derog: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ policy_code               : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ application_type          : chr  "INDIVIDUAL" "INDIVIDUAL" "INDIVIDUAL" "INDIVIDUAL" ...
##  $ annual_inc_joint          : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ dti_joint                 : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ verification_status_joint  : chr  "" "" "" "" ...
##  $ acc_now_delinq            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ tot_coll_amt              : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ tot_cur_bal               : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ open_acc_6m               : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ open_il_6m                : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ open_il_12m               : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ open_il_24m               : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ mths_since_rcnt_il        : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ total_bal_il              : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ il_util                   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ open_rv_12m               : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ open_rv_24m               : num  NA NA NA NA NA NA NA NA NA NA ...
```

```
##  $ max_bal_bc                : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ all_util                  : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ total_rev_hi_lim          : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ inq_fi                    : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ total_cu_tl               : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ inq_last_12m              : num  NA NA NA NA NA NA NA NA NA NA ...
```

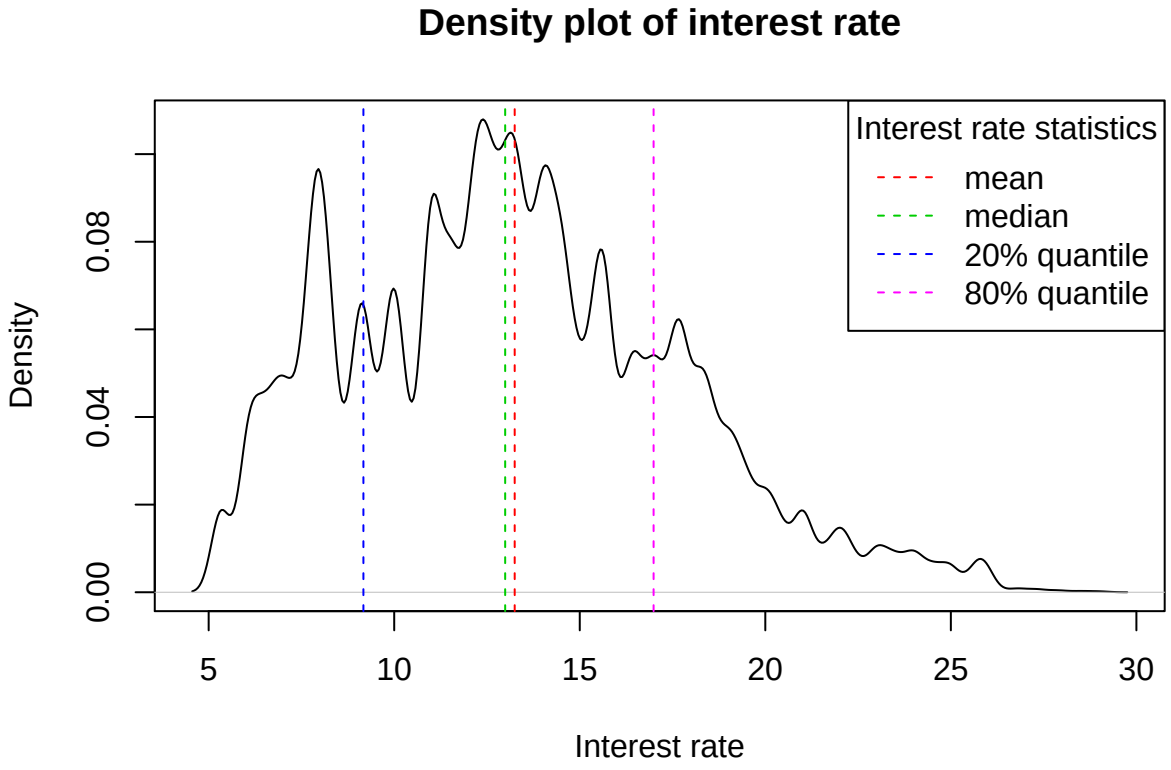
Calculate the number of na values for each column.

```
# seperate columns with over 80% NA values
num.NA <- sort(sapply(loan, function(x) sum(is.na(x))), decreasing = TRUE)
remain.col = names(num.NA)[(num.NA < 0.8 * dim(loan)[1])]
delete.col = names(num.NA)[(num.NA >= 0.8 * dim(loan)[1])]
delete.col
```

```
##  [1] "dti_joint"             "annual_inc_joint"
##  [3] "il_util"               "mths_since_rcnt_il"
##  [5] "open_acc_6m"           "open_il_6m"
##  [7] "open_il_12m"           "open_il_24m"
##  [9] "total_bal_il"          "open_rv_12m"
## [11] "open_rv_24m"           "max_bal_bc"
## [13] "all_util"              "inq_fi"
## [15] "total_cu_tl"           "inq_last_12m"
## [17] "mths_since_last_record"
```

EDA_part_1 (What factor will influence the interest rate?)

```
{
    plot(density((loan$int_rate)), main = "Density plot of interest rate", xlab = "Interest rate")
    abline(v = mean(loan$int_rate), lty = 2, col = 2)
    abline(v = median(loan$int_rate), lty = 2, col = 3)
    abline(v = quantile(loan$int_rate, 0.2), lty = 2, col = 4)
    abline(v = quantile(loan$int_rate, 0.8), lty = 2, col = 6)
    legend("topright", c("mean", "median", "20% quantile", "80% quantile"),
        col = c(2, 3, 4, 6), lty = 2, title = "Interest rate statistics")
}
```



The distribution is a bit right skew. In the future, if we want to build model with interest rate then it is better to use square root to adjust the skewness. Next I want to explore the correlation between interest rate with other numeric variables. However, we know some features are with high number of NA values. So I want to remove those features with over 80% NA values temperarily.

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

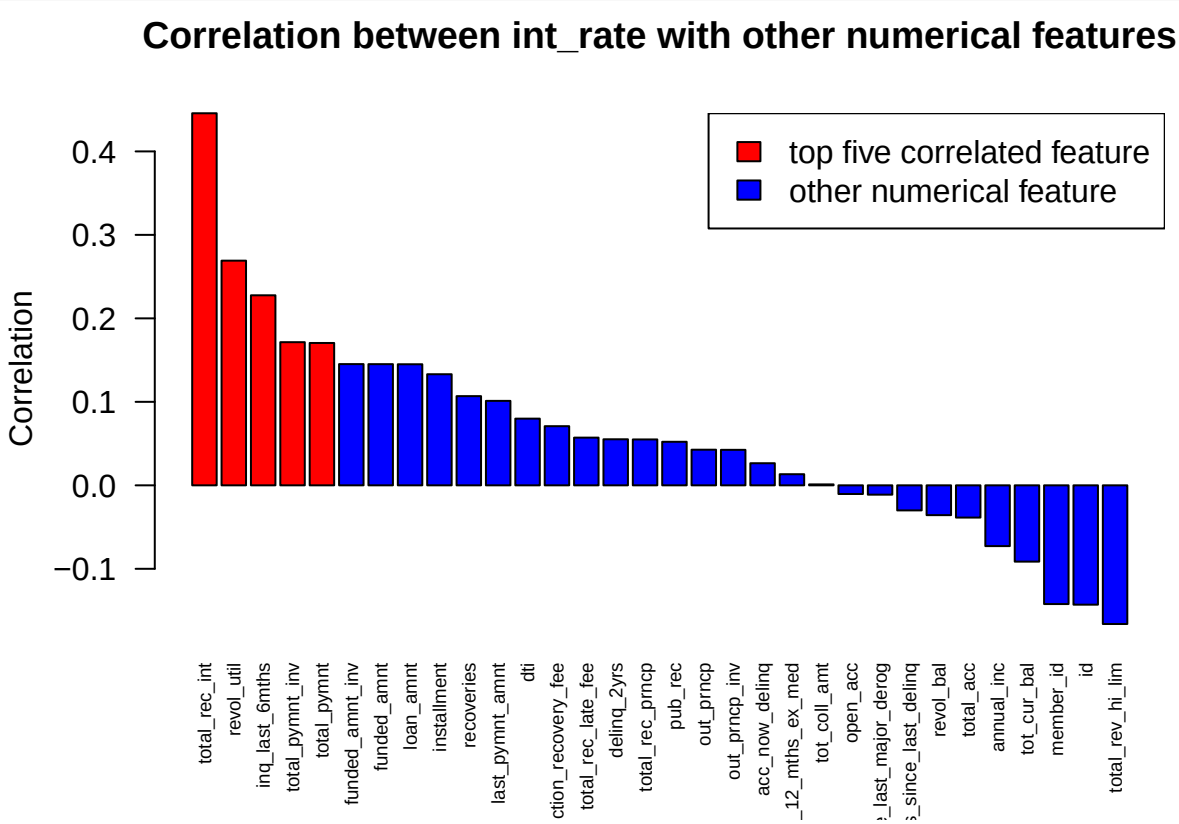
```
loan <- loan[, remain.col]
# select numerical features' name
num.feature <- names(loan[, sapply(loan, is.numeric)])
# select char features' name
char.feature <- names(loan[, sapply(loan, is.character)])
# calculate the correlationship between int_rate and other numerical
# features
correlation <- cor(loan$int_rate, loan[, num.feature], use = "pairwise.complete.obs")
```

```
## Warning in cor(loan$int_rate, loan[, num.feature], use =
## "pairwise.complete.obs"): the standard deviation is zero
```

```
# sort the value
correlation <- correlation[, order(correlation[1, ], decreasing = TRUE)]
# remove correlation with itself and correlation with police_code
correlation <- correlation[2:33]
```

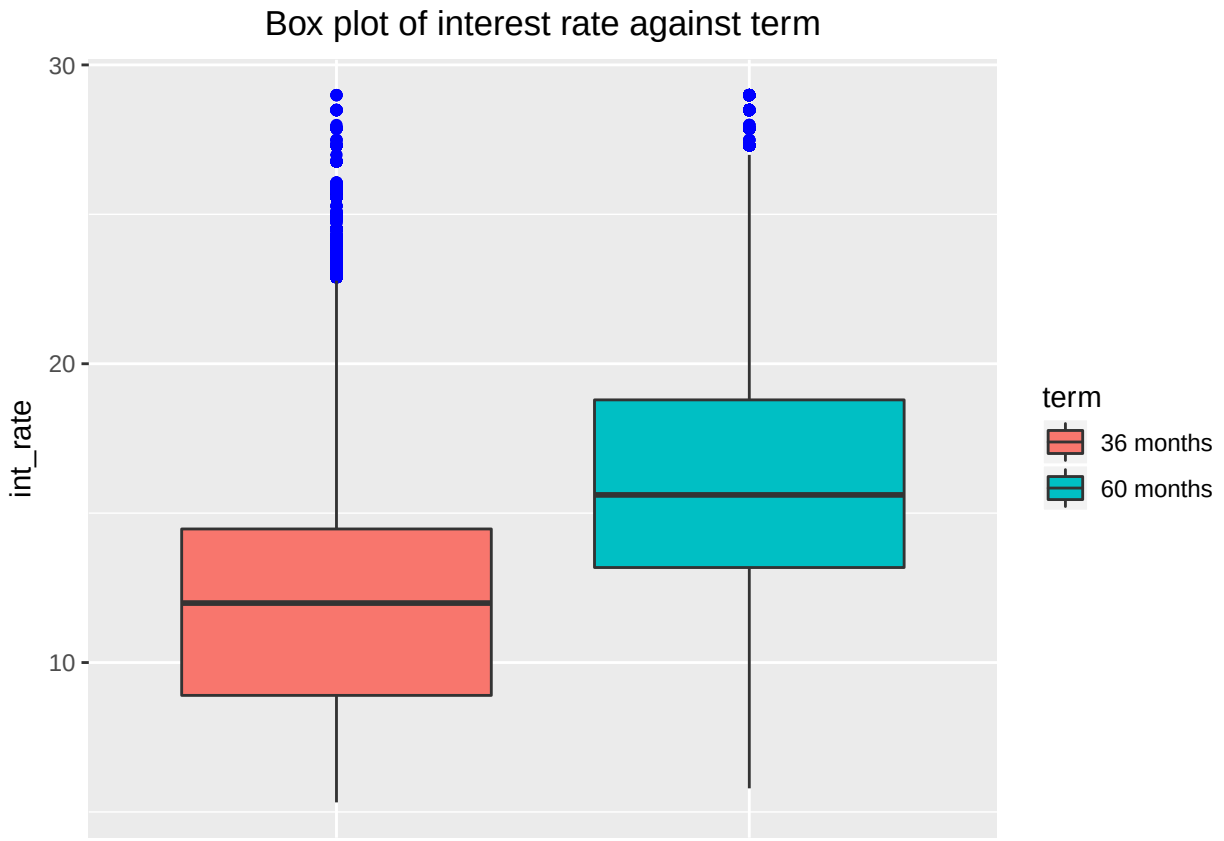
```
{
    barplot(correlation, main = "Correlation between int_rate with other numerical features",
        ylab = "Correlation", las = 2, cex.names = 0.6, col = ifelse(correlation >
            0.15, "red", "blue"))
    legend("topright", leg = c("top five correlated feature", "other numerical feature"),
        fill = c("red", "blue"))
```

```
}
```



Based on correlation calculation, the top five predictive numerical features for int_rate are "total_rec_int", "revol_util", "inq_last_6mths", "total_pymnt_inv" and "total_pymnt". Next step is to explore the top five influencial category features.

```
library(ggplot2)
ggplot(data = loan, aes(term, int_rate, fill = term)) + geom_boxplot(outlier.color = "blue") +
    labs(title = "Box plot of interest rate against term") + theme(axis.text.x = element_blank(),
    axis.title.x = element_blank(), axis.ticks.x = element_blank(), plot.title = element_text(hjust = 0
```

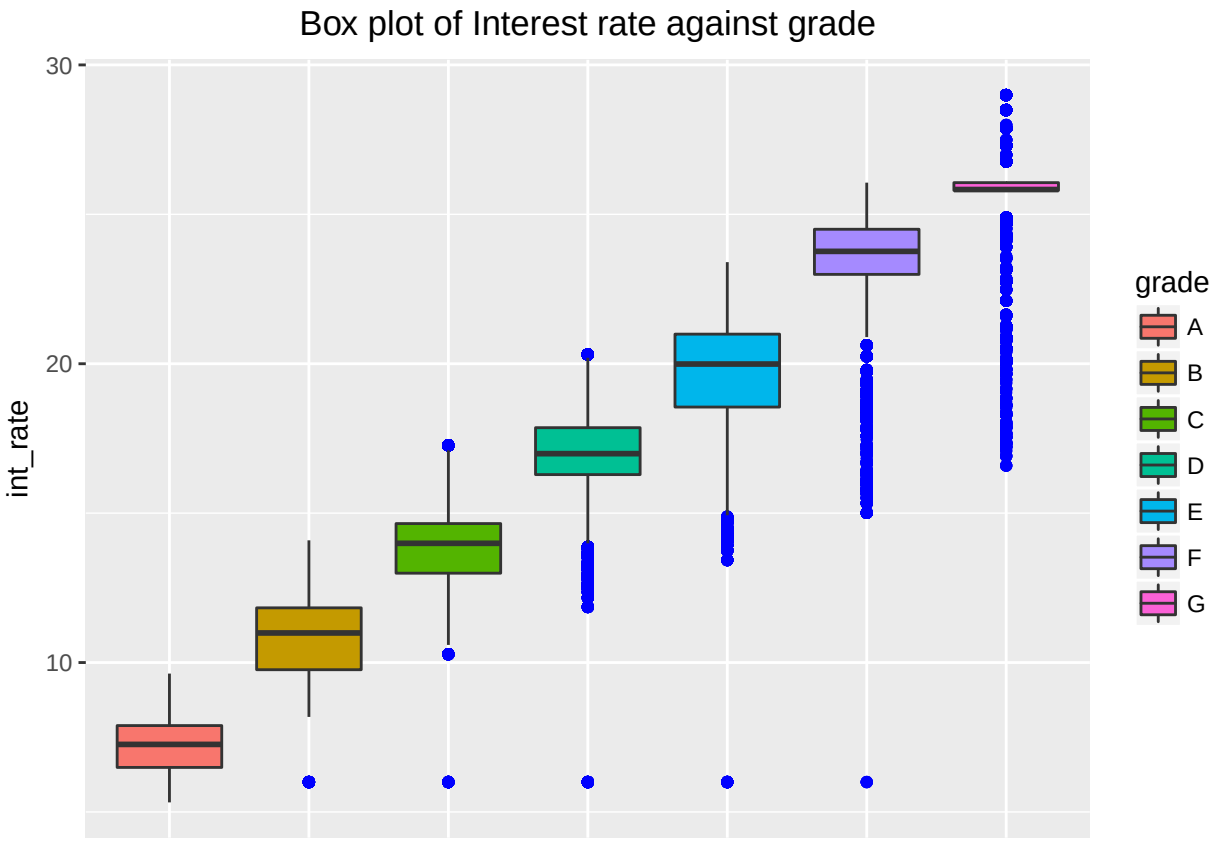



```
t.test(subset(loan, term == " 36 months")$int_rate, subset(loan, term == " 60 months")$int_rate,
    conf.level = 0.95, mu = 0, alternative = "two.sided", paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(loan, term == " 36 months")$int_rate and subset(loan, term == " 60 months")$int_rate
## t = -431.12, df = 467040, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.111525 -4.074310
## sample estimates:
## mean of x mean of y
##  12.01868  16.11160
```

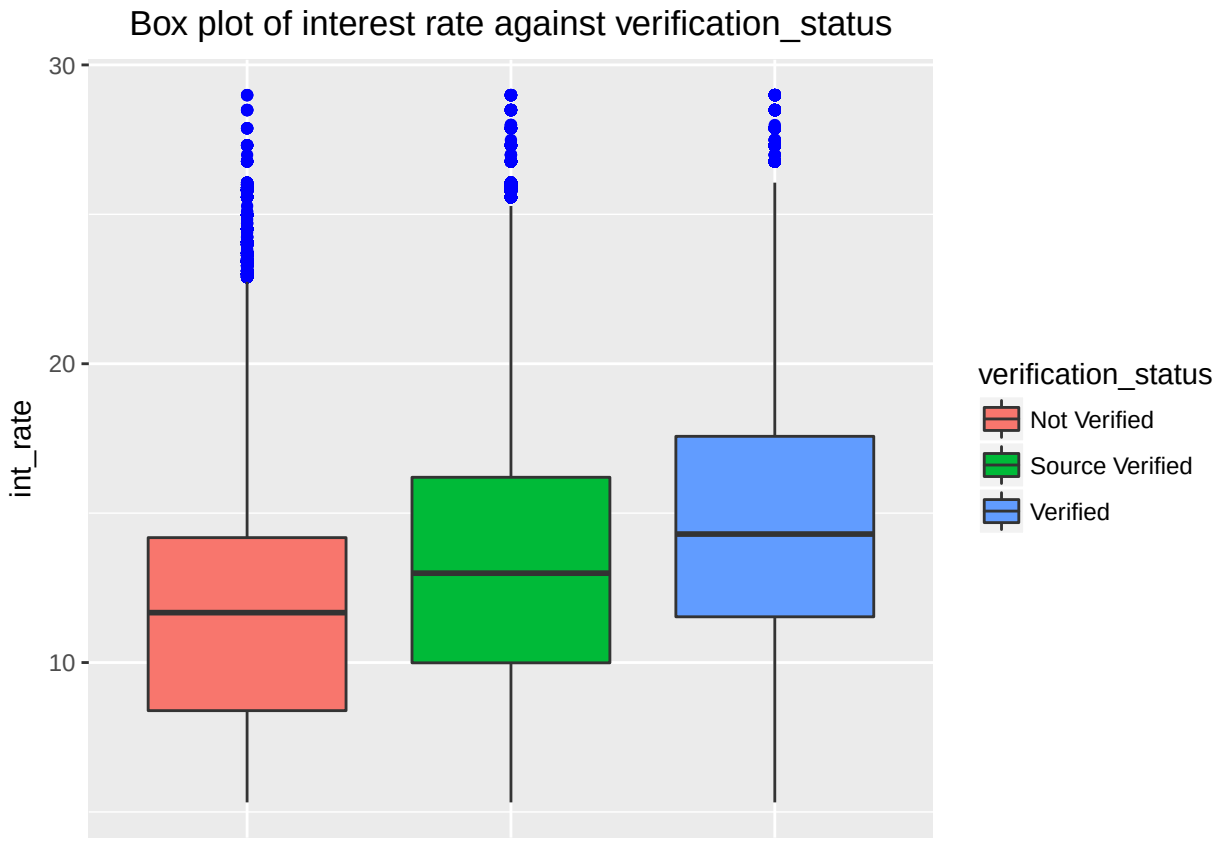
There is a significant difference between different term. Therefore, term can be used a predictor for interest rate.

```
ggplot(data = loan, aes(grade, int_rate, fill = grade)) + geom_boxplot(outlier.color = "blue") +
    labs(title = "Box plot of Interest rate against grade") + theme(axis.text.x = element_blank(),
    axis.title.x = element_blank(), axis.ticks.x = element_blank(), plot.title = element_text(hjust = 0
```



There is a clear linear trend between interest rate and grade. Therefore, grade can be used a predictor for interest rate. However, based on the description of grade. It is assigned by the Lending Club. Therefore, I probably don't have this feature in advance. If a client is a return user and I can definitely use this information.

```
ggplot(data = loan, aes(verification_status, int_rate, fill = verification_status)) +
    geom_boxplot(outlier.color = "blue") + labs(title = "Box plot of interest rate against verification_
    theme(axis.text.x = element_blank(), axis.title.x = element_blank(), axis.ticks.x = element_blank()
        plot.title = element_text(hjust = 0.5))
```




```
t.test(subset(loan, verification_status == "Verified")$int_rate, subset(loan,
    verification_status == "Source Verified")$int_rate, conf.level = 0.95, mu = 0,
    alternative = "two.sided", paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(loan, verification_status == "Verified")$int_rate and subset(loan, verification_status
## t = 117.2, df = 605530, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.291005 1.334919
## sample estimates:
## mean of x mean of y
##  14.57173  13.25877
```

```
t.test(subset(loan, verification_status == "Not Verified")$int_rate, subset(loan,
    verification_status == "Source Verified")$int_rate, conf.level = 0.95, mu = 0,
    alternative = "two.sided", paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(loan, verification_status == "Not Verified")$int_rate and subset(loan, verification_sta
## t = -138.82, df = 590720, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.493490 -1.451905
```

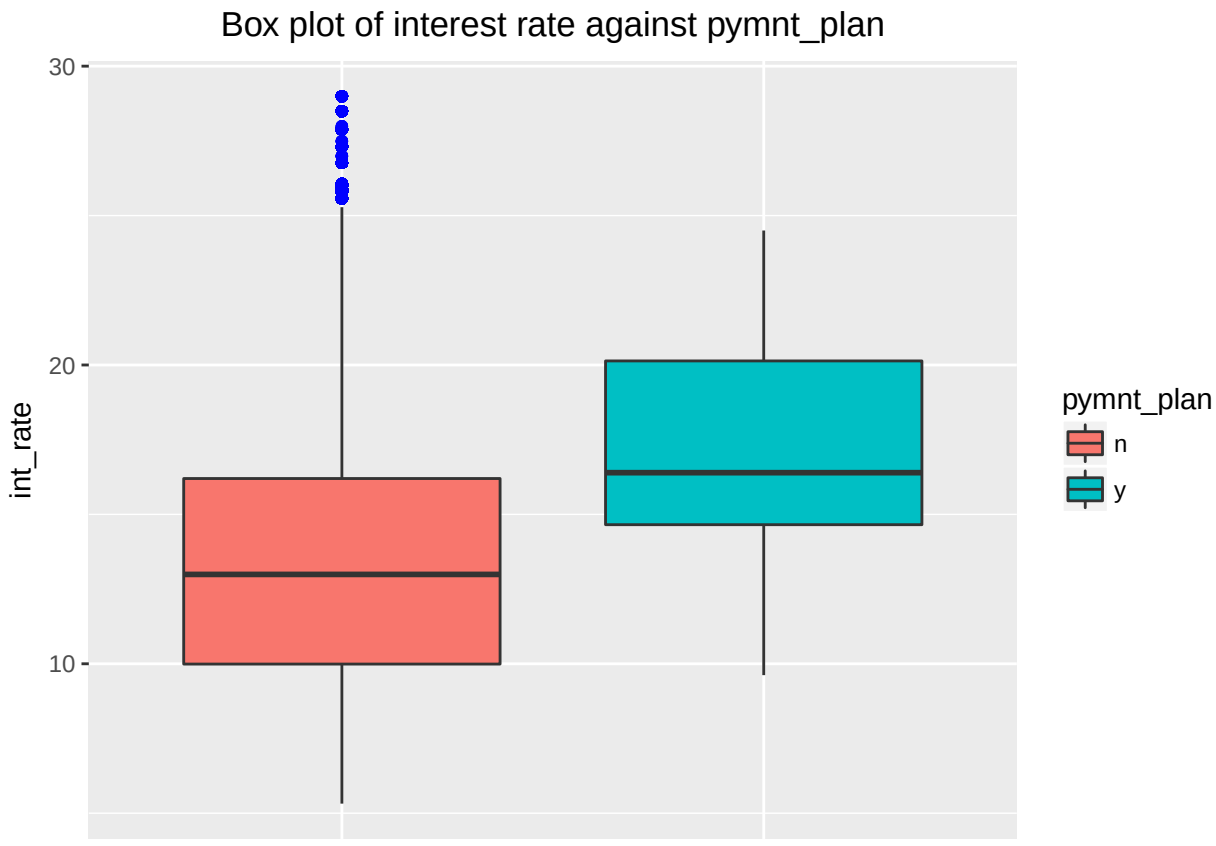
```
## sample estimates:
## mean of x mean of y
##  11.78607  13.25877
```

```
t.test(subset(loan, verification_status == "Not Verified")$int_rate, subset(loan,
    verification_status == "Verified")$int_rate, conf.level = 0.95, mu = 0,
    alternative = "two.sided", paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(loan, verification_status == "Not Verified")$int_rate and subset(loan, verification_sta
## t = -249.62, df = 555780, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.807532 -2.763787
## sample estimates:
## mean of x mean of y
##  11.78607  14.57173
```

There is a significant difference between different verification_status Therefore, verification_status can be used a predictor for interest rate.

```
ggplot(data = loan, aes(pymnt_plan, int_rate, fill = pymnt_plan)) + geom_boxplot(outlier.color = "blue"
    labs(title = "Box plot of interest rate against pymnt_plan") + theme(axis.text.x = element_blank(),
    axis.title.x = element_blank(), axis.ticks.x = element_blank(), plot.title = element_text(hjust = 0
```



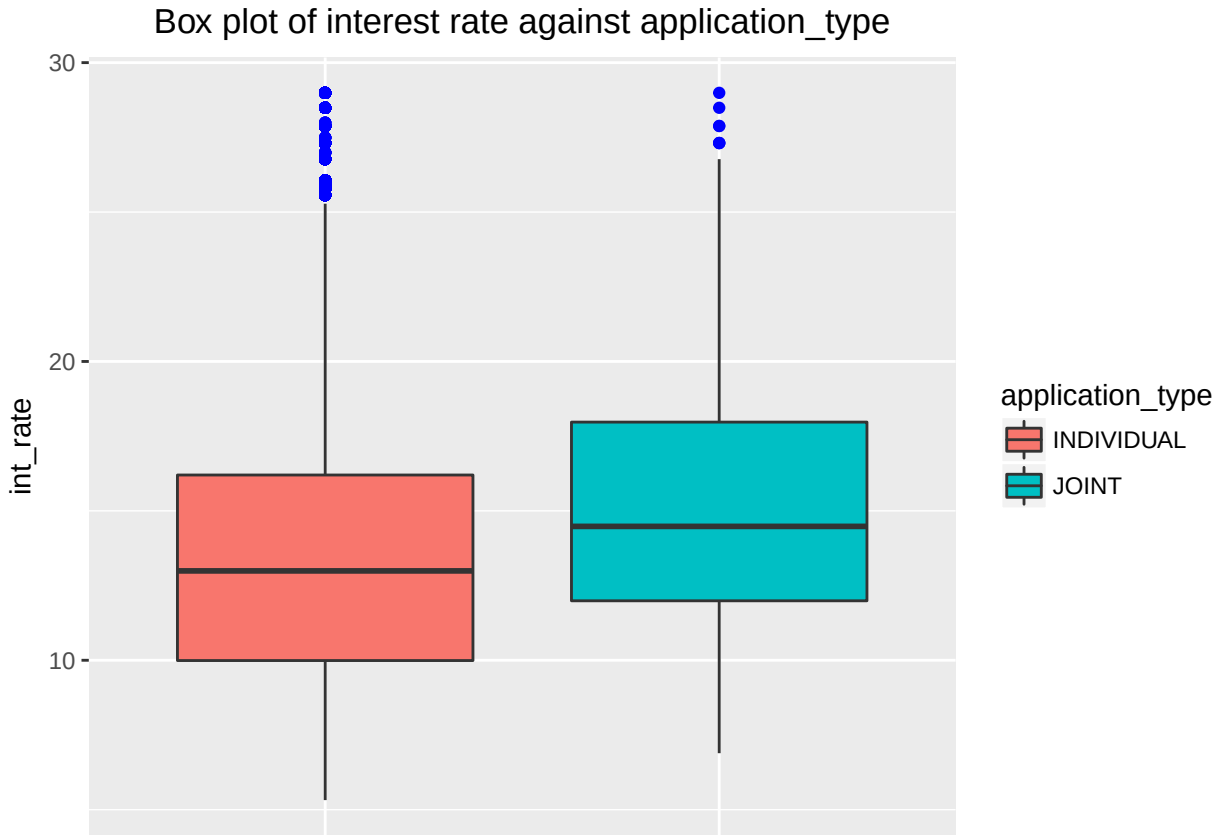

```
t.test(subset(loan, pymnt_plan == "n")$int_rate, subset(loan, pymnt_plan ==
    "y")$int_rate, conf.level = 0.95, mu = 0, alternative = "two.sided", paired = FALSE,
```

```
    var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(loan, pymnt_plan == "n")$int_rate and subset(loan, pymnt_plan == "y")$int_rate
## t = -2.7241, df = 9.0002, p-value = 0.02345
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.3313511 -0.6792598
## sample estimates:
## mean of x mean of y
##  13.24669  17.25200
```

There is a significant difference between different pymnt_plan. Therefore, pymnt_plan. can be used a predictor for interest rate.

```
ggplot(data = loan, aes(application_type, int_rate, fill = application_type)) +
    geom_boxplot(outlier.color = "blue") + labs(title = "Box plot of interest rate against application_
    theme(axis.text.x = element_blank(), axis.title.x = element_blank(), axis.ticks.x = element_blank()
        plot.title = element_text(hjust = 0.5))
```




```
t.test(subset(loan, application_type == "INDIVIDUAL")$int_rate, subset(loan,
    application_type == "JOINT")$int_rate, conf.level = 0.95, mu = 0, alternative = "two.sided",
    paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
```

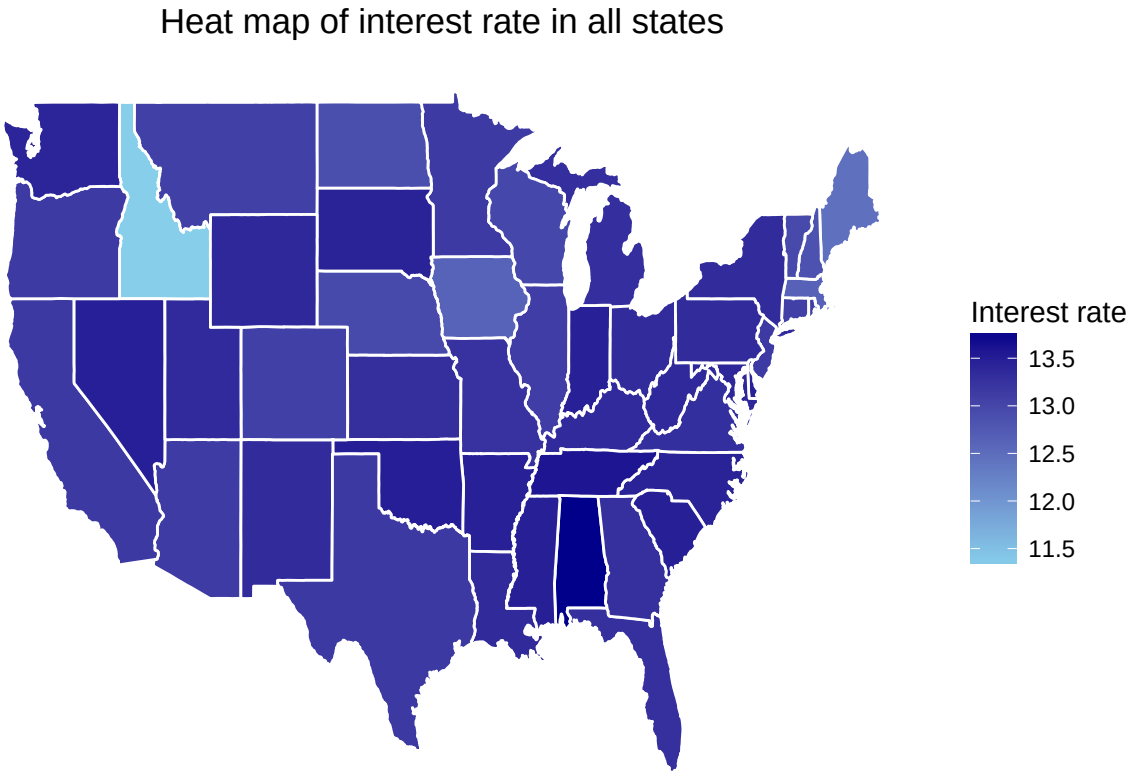
```
##
## data:  subset(loan, application_type == "INDIVIDUAL")$int_rate and subset(loan, application_type ==
## t = -10.139, df = 510.61, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.299387 -1.552913
## sample estimates:
## mean of x mean of y
##  13.24563  15.17178
```

There is a significant difference between different application_type. Therefore, application_type can be used a predictor for interest rate. In conclustion, these five category variables are influenciable: "term", "grade", "verification_status", "pymnt_plan" and "application_type". Besides I am also curious about how interest rate vary with space and time.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
suppressPackageStartupMessages(library(maps))
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")

all_states <- map_data("state")
state_by_rate <- loan %>% group_by(region) %>% summarise(value = mean(int_rate,
    na.rm = TRUE))
state_by_rate$region <- as.character(state_by_rate$region)

Total <- merge(all_states, state_by_rate, by = "region")

p <- ggplot()
p <- p + geom_polygon(data = Total, aes(x = long, y = lat, group = group, fill = Total$value),
    colour = "white") + scale_fill_continuous(low = "skyblue", high = "darkblue",
    guide = "colorbar")
P1 <- p + theme_bw() + labs(fill = "Interest rate", title = "Heat map of interest rate in all states",
    x = "", y = "")
P1 + scale_y_continuous(breaks = c()) + scale_x_continuous(breaks = c()) + theme(panel.border = element_
    plot.title = element_text(hjust = 0.5))
```

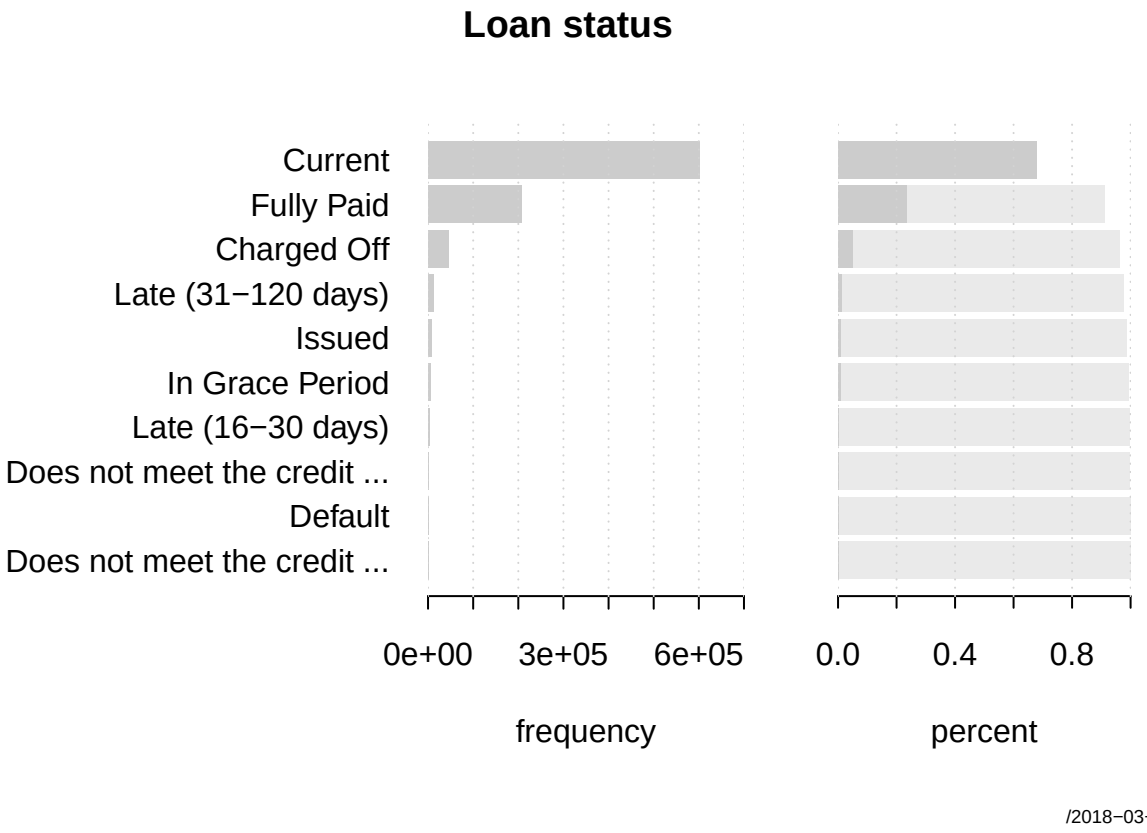


EDA_part_2 (What are the distribution of loan status?)

```
library(DescTools)
Desc(loan$loan_status, plotit = TRUE, main = "Loan status")
```

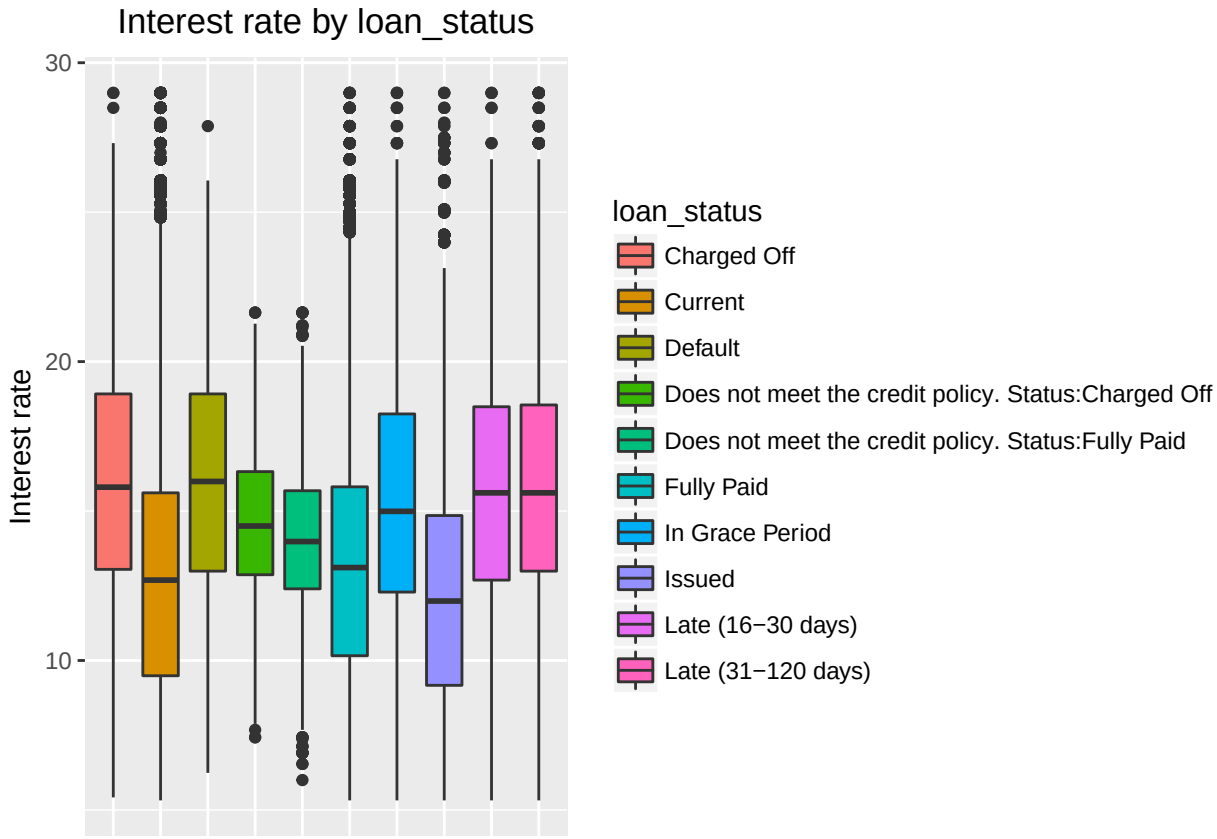
```
## ------------------------------------------------------------------------
## Loan status
##
##    length       n     NAs  unique  levels   dupes
##   887'379 887'379       0      10      10       y
##           100.0%    0.0%
##
##                                                    level     freq   perc  cumfreq  cumperc
## 1                                                Current  601'779  67.8%  601'779    67.8%
## 2                                             Fully Paid  207'723  23.4%  809'502    91.2%
## 3                                            Charged Off   45'248   5.1%  854'750    96.3%
## 4                                    Late (31-120 days)   11'591   1.3%  866'341    97.6%
## 5                                                Issued    8'460   1.0%  874'801    98.6%
## 6                                     In Grace Period    6'253   0.7%  881'054    99.3%
## 7                                    Late (16-30 days)    2'357   0.3%  883'411    99.6%
## 8    Does not meet the credit policy. Status:Fully Paid    1'988   0.2%  885'399    99.8%
## 9                                                Default    1'219   0.1%  886'618    99.9%
## 10  Does not meet the credit policy. Status:Charged Off      761   0.1%  887'379   100.0%
```

```
## Warning in as.POSIXlt.POSIXct(Sys.time()): unknown timezone 'default/
## America/Los_Angeles'
```



See how does loan_status affact interest rate.

```
ggplot(data = loan, aes(loan_status, int_rate), las = 2) + geom_boxplot(aes(fill = loan_status)) +
    labs(list(title = "Interest rate by loan_status", x = "Loan_status", y = "Interest rate")) +
    theme(axis.text.x = element_blank(), axis.title.x = element_blank(), axis.ticks.x = element_blank()
        plot.title = element_text(hjust = 0.5))
```



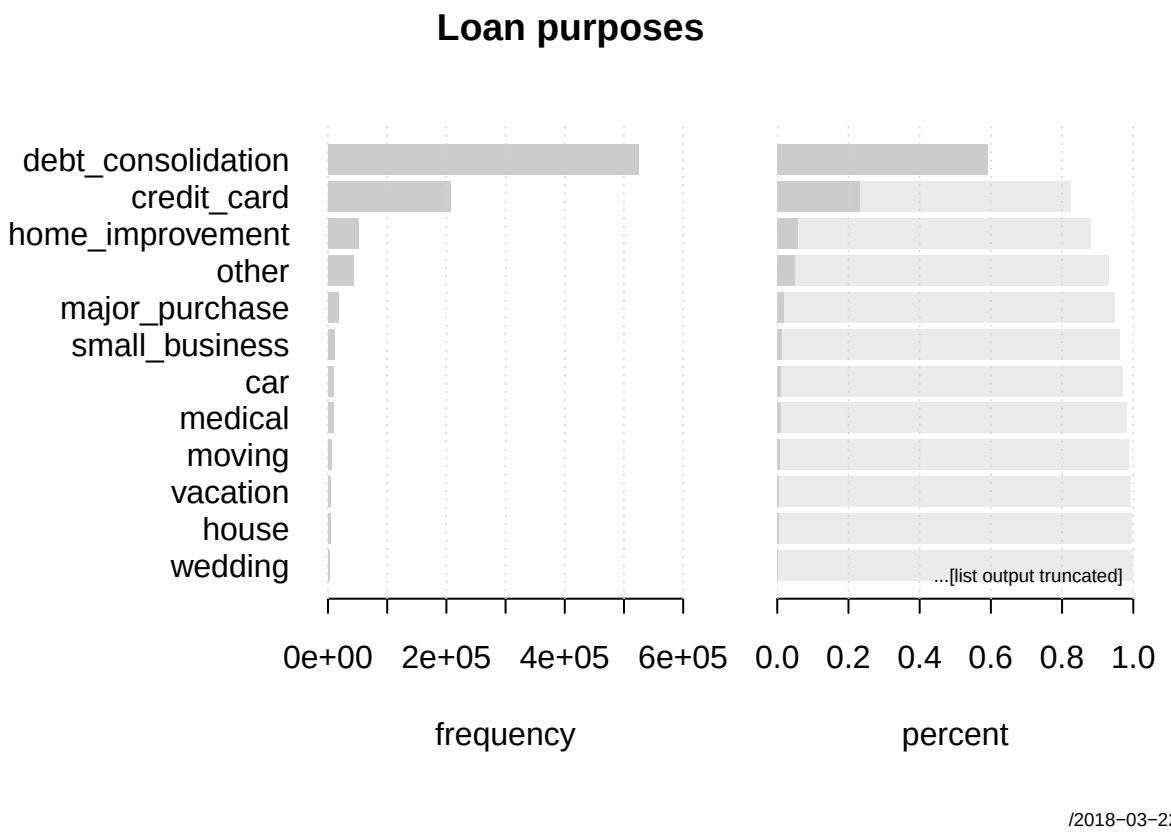
The majority status is "Current". The finished loan can be grouped into "Fully Paid" and "Charged off" or "Late payment".

EDA_part_3 (What are the purpose of applying a loan with Lending Club ?)

```
Desc(loan$purpose, main = "Loan purposes", plotit = TRUE)
```

```
## ------------------------------------------------------------------------
## Loan purposes
##
##    length       n     NAs  unique  levels   dupes
##   887'379 887'379       0      14      14       y
##           100.0%    0.0%
##
##                  level     freq   perc  cumfreq  cumperc
## 1   debt_consolidation  524'215  59.1%  524'215    59.1%
## 2          credit_card  206'182  23.2%  730'397    82.3%
## 3     home_improvement   51'829   5.8%  782'226    88.2%
## 4                other   42'894   4.8%  825'120    93.0%
## 5       major_purchase   17'277   1.9%  842'397    94.9%
## 6       small_business   10'377   1.2%  852'774    96.1%
## 7                  car    8'863   1.0%  861'637    97.1%
## 8              medical    8'540   1.0%  870'177    98.1%
## 9               moving    5'414   0.6%  875'591    98.7%
```

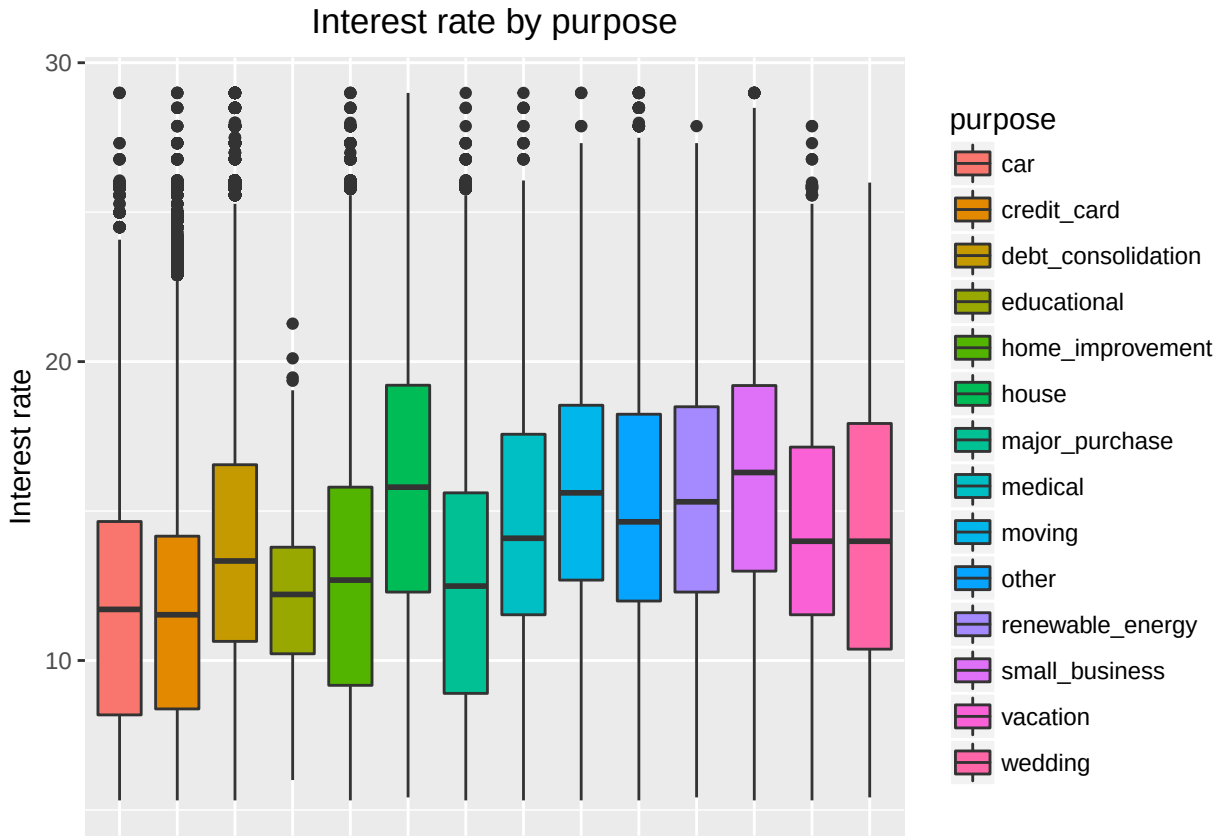
```
## 10            vacation    4'736   0.5%  880'327    99.2%
## 11               house    3'707   0.4%  884'034    99.6%
## 12             wedding    2'347   0.3%  886'381    99.9%
## ... etc.
##  [list output truncated]
```



/2018−03−22

I am also curious about how interest rate vary with purpose.

```
ggplot(data = loan, aes(purpose, int_rate), las = 2) + geom_boxplot(aes(fill = purpose)) +
    labs(list(title = "Interest rate by purpose", x = "purpose of Loan", y = "Interest rate")) +
    theme(axis.text.x = element_blank(), axis.title.x = element_blank(), axis.ticks.x = element_blank()
        plot.title = element_text(hjust = 0.5))
```


Interest rate by purpose
Interest rate
30
20
10
purpose
car
credit_card
debt_consolidation
educational
home_improvement
house
major_purchase
medical
moving
other
renewable_energy
small_business
vacation
wedding