R Notebook

Code

Reading data

Hide

```
library('dplyr',quietly = TRUE)
library('ggplot2',quietly = TRUE)
library('gridExtra',quietly = TRUE)
library('grid',quietly = TRUE)
library("rgdal",quietly = TRUE)
library("choroplethrMaps",quietly = TRUE)
library("choroplethr",quietly = TRUE)
library('ClustOfVar', quietly = TRUE)
library('corrplot', quietly = TRUE)
library('cluster', quietly = TRUE)

#data = read.csv("C:/Users/mua377n/Downloads/archive/Loan_status_2007-2020Q3.gzip")
#data = data[-1]
print(dim(data))
```

```
[1] 2925493     142
```

Hide

```
colnames(data)
```

```
  [1] "X"
  [2] "id"
  [3] "loan_amnt"
  [4] "funded_amnt"
  [5] "funded_amnt_inv"
  [6] "term"
  [7] "int_rate"
  [8] "installment"
  [9] "grade"
 [10] "sub_grade"
 [11] "emp_title"
 [12] "emp_length"
 [13] "home_ownership"
 [14] "annual_inc"
 [15] "verification_status"
 [16] "issue_d"
 [17] "loan_status"
 [18] "pymnt_plan"
 [19] "url"
 [20] "purpose"
 [21] "title"
 [22] "zip_code"
 [23] "addr_state"
 [24] "dti"
 [25] "delinq_2yrs"
```

```
[26] "earliest_cr_line"
[27] "fico_range_low"
[28] "fico_range_high"
[29] "inq_last_6mths"
[30] "mths_since_last_delinq"
[31] "mths_since_last_record"
[32] "open_acc"
[33] "pub_rec"
[34] "revol_bal"
[35] "revol_util"
[36] "total_acc"
[37] "initial_list_status"
[38] "out_prncp"
[39] "out_prncp_inv"
[40] "total_pymnt"
[41] "total_pymnt_inv"
[42] "total_rec_prncp"
[43] "total_rec_int"
[44] "total_rec_late_fee"
[45] "recoveries"
[46] "collection_recovery_fee"
[47] "last_pymnt_d"
[48] "last_pymnt_amnt"
[49] "next_pymnt_d"
[50] "last_credit_pull_d"
[51] "last_fico_range_high"
[52] "last_fico_range_low"
[53] "collections_12_mths_ex_med"
[54] "mths_since_last_major_derog"
[55] "policy_code"
[56] "application_type"
[57] "annual_inc_joint"
[58] "dti_joint"
[59] "verification_status_joint"
[60] "acc_now_delinq"
[61] "tot_coll_amt"
[62] "tot_cur_bal"
[63] "open_acc_6m"
[64] "open_act_il"
[65] "open_il_12m"
[66] "open_il_24m"
[67] "mths_since_rcnt_il"
[68] "total_bal_il"
[69] "il_util"
[70] "open_rv_12m"
[71] "open_rv_24m"
[72] "max_bal_bc"
[73] "all_util"
[74] "total_rev_hi_lim"
[75] "inq_fi"
[76] "total_cu_tl"
[77] "inq_last_12m"
[78] "acc_open_past_24mths"
[79] "avg_cur_bal"
[80] "bc_open_to_buy"
[81] "bc_util"
[82] "chargeoff_within_12_mths"
```

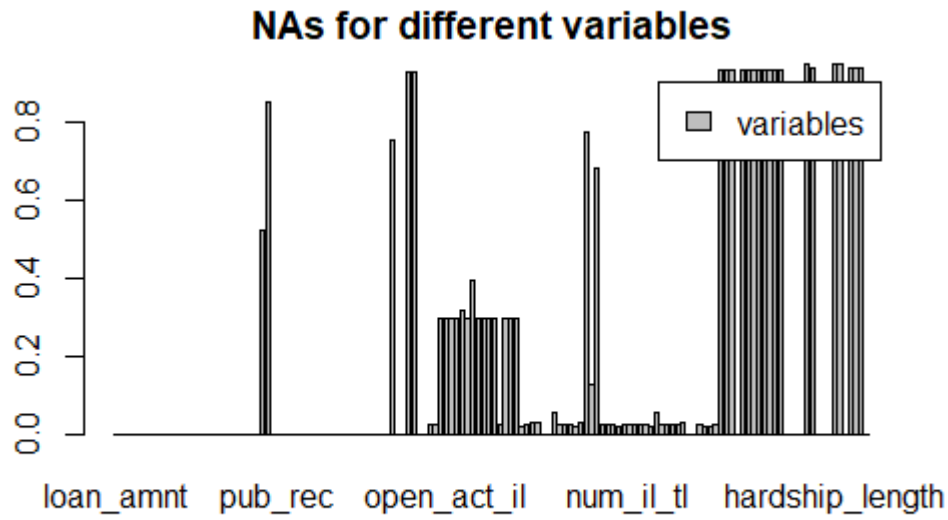
```
 [83] "delinq_amnt"
 [84] "mo_sin_old_il_acct"
 [85] "mo_sin_old_rev_tl_op"
 [86] "mo_sin_rcnt_rev_tl_op"
 [87] "mo_sin_rcnt_tl"
 [88] "mort_acc"
 [89] "mths_since_recent_bc"
 [90] "mths_since_recent_bc_dlq"
 [91] "mths_since_recent_inq"
 [92] "mths_since_recent_revol_delinq"
 [93] "num_accts_ever_120_pd"
 [94] "num_actv_bc_tl"
 [95] "num_actv_rev_tl"
 [96] "num_bc_sats"
 [97] "num_bc_tl"
 [98] "num_il_tl"
 [99] "num_op_rev_tl"
[100] "num_rev_accts"
[101] "num_rev_tl_bal_gt_0"
[102] "num_sats"
[103] "num_tl_120dpd_2m"
[104] "num_tl_30dpd"
[105] "num_tl_90g_dpd_24m"
[106] "num_tl_op_past_12m"
[107] "pct_tl_nvr_dlq"
[108] "percent_bc_gt_75"
[109] "pub_rec_bankruptcies"
[110] "tax_liens"
[111] "tot_hi_cred_lim"
[112] "total_bal_ex_mort"
[113] "total_bc_limit"
[114] "total_il_high_credit_limit"
[115] "revol_bal_joint"
[116] "sec_app_fico_range_low"
[117] "sec_app_fico_range_high"
[118] "sec_app_earliest_cr_line"
[119] "sec_app_inq_last_6mths"
[120] "sec_app_mort_acc"
[121] "sec_app_open_acc"
[122] "sec_app_revol_util"
[123] "sec_app_open_act_il"
[124] "sec_app_num_rev_accts"
[125] "sec_app_chargeoff_within_12_mths"
[126] "sec_app_collections_12_mths_ex_med"
[127] "hardship_flag"
[128] "hardship_type"
[129] "hardship_reason"
[130] "hardship_status"
[131] "deferral_term"
[132] "hardship_amount"
[133] "hardship_start_date"
[134] "hardship_end_date"
[135] "payment_plan_start_date"
[136] "hardship_length"
[137] "hardship_dpd"
[138] "hardship_loan_status"
[139] "orig_projected_additional_accrued_interest"
```

```
[140] "hardship_payoff_balance_amount"
[141] "hardship_last_payment_amount"
[142] "debt_settlement_flag"
```

missing values plot for different variables

Hide

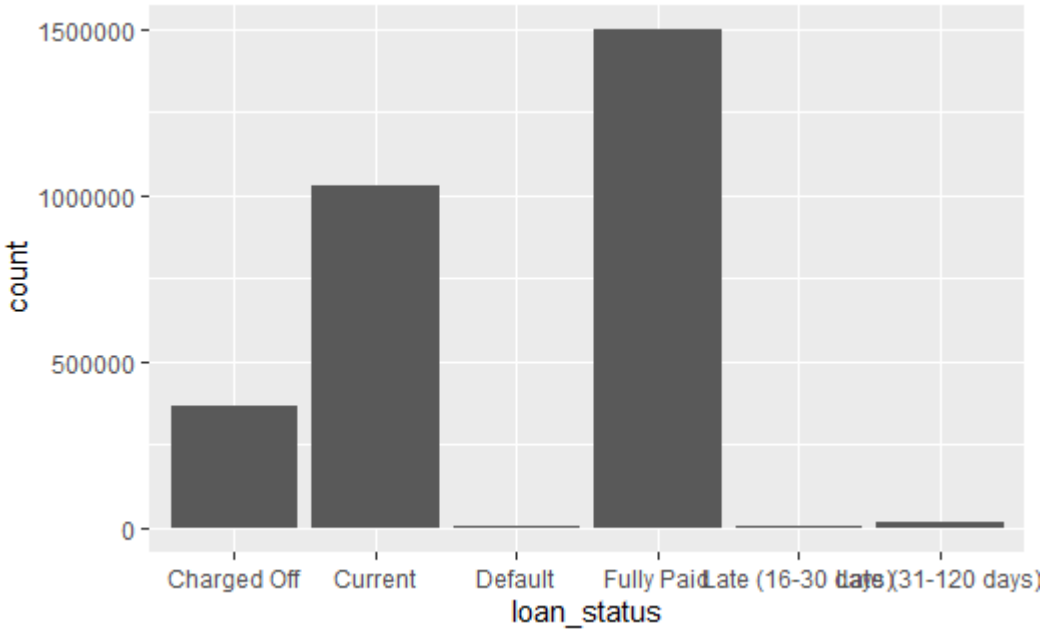
```
missing_values = apply(is.na(data),2, sum )/nrow(data)
barplot(missing_values,legend.text = 'variables', main='NAs for different variable
s')
```



treating loan_status

Hide

```
loan_status_keep = c(unique(data$loan_status)[1],
                     unique(data$loan_status)[2],
                     unique(data$loan_status)[6],
                     unique(data$loan_status)[7],
                     unique(data$loan_status)[9],
                     unique(data$loan_status)[10])
data_new = data[data$loan_status %in% loan_status_keep,]
remove(data)
ggplot(data = data_new)+geom_bar(mapping = aes(x = loan_status))
```



dropping temporal variables

Hide

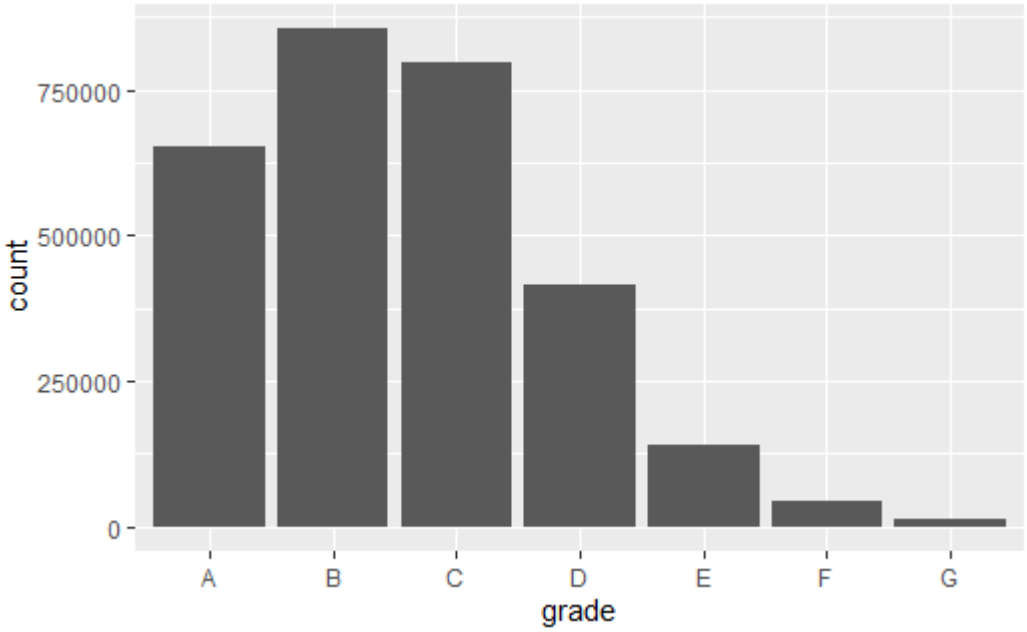
```
drop_colums = c('last_credit_pull_d','last_fico_range_high','last_fico_range_low',
'total_pymnt',
                        'total_pymnt_inv','recoveries','collection_recovery_fee',
'out_prncp','out_prncp_inv',
                        'total_rec_prncp','total_rec_int','last_pymnt_d','last_pym
nt_amnt','next_pymnt_d',
                        'total_rec_late_fee','hardship_flag','hardship_amount',
                        'orig_projected_additional_accrued_interest','hardship_pay
off_balance_amount',
                        'hardship_last_payment_amount','debt_settlement_flag','har
dship_type','hardship_reason',
                        'hardship_status','deferral_term','hardship_start_date','h
ardship_end_date','payment_plan_start_date',
                        'hardship_length','hardship_dpd','hardship_loan_status','i
nstallment','pymnt_plan',                        'acc_now_delinq','url','earliest_
cr_line','last_fico_range_low','last_fico_range_high','policy_code','zip_code',
                      'sec_app_fico_range_low','sec_app_fico_range_high','sec_app_
earliest_cr_line','fico_range_low', 'fico_range_high',
                     'annual_inc_joint','dti_joint','verification_status_joint',
'revol_bal_joint','sec_app_inq_last_6mths',
                     'sec_app_mort_acc','sec_app_open_acc','sec_app_revol_util','s
ec_app_open_act_il',
                      'sec_app_num_rev_accts','sec_app_chargeoff_within_12_mths',
'sec_app_collections_12_mths_ex_med')

data_new = data_new[,!names(data_new)%in% drop_colums]
# data_new = data_new %>% select (-one_of(drop_colums)
#stripping % from charater type and converting to numeric
data_new$int_rate = as.numeric(sub("%", "", data_new$int_rate))
data_new$revol_util = as.numeric(sub("%", "", data_new$revol_util))
data_new$id = as.integer(data_new$id)
```

visualizing grade

Hide

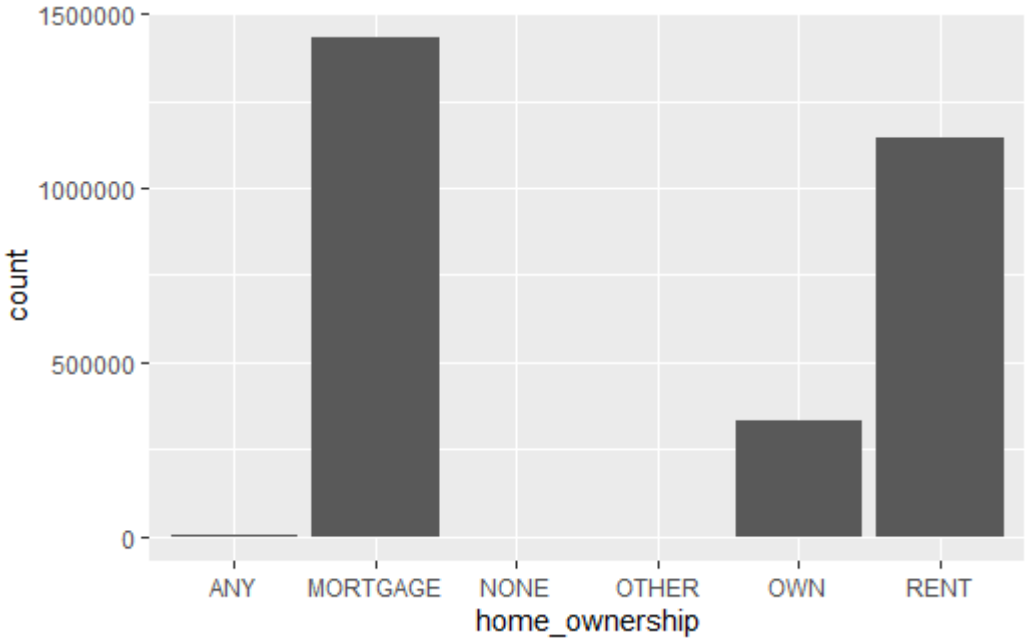
```
ggplot(data = data_new)+geom_bar(mapping = aes(x = grade))
```



visualizing home ownership

Hide

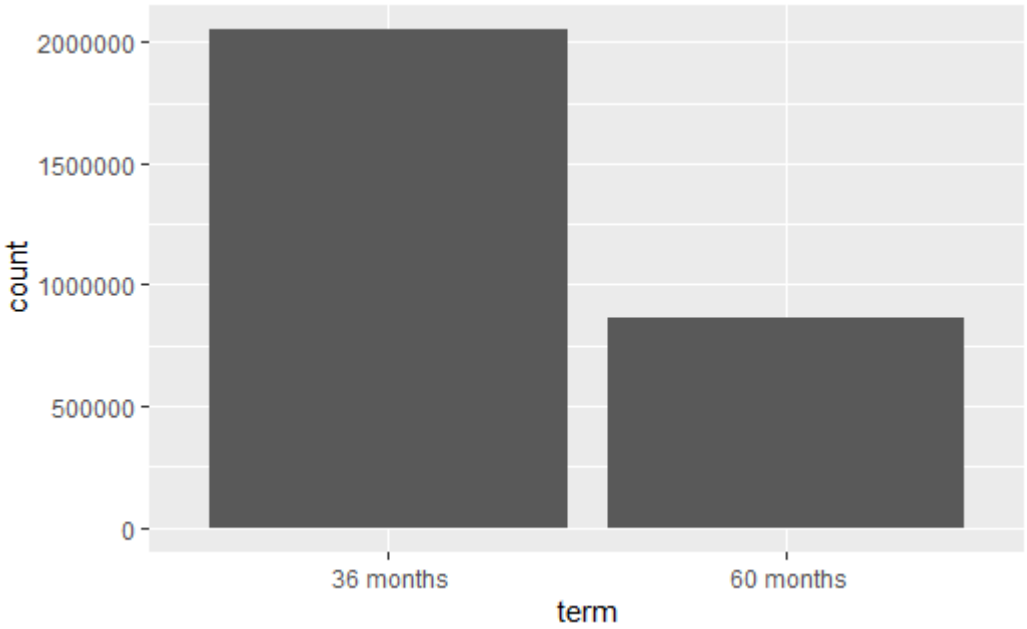
```
ggplot(data = data_new)+geom_bar(mapping = aes(x = home_ownership))
```



visualizing term

Hide

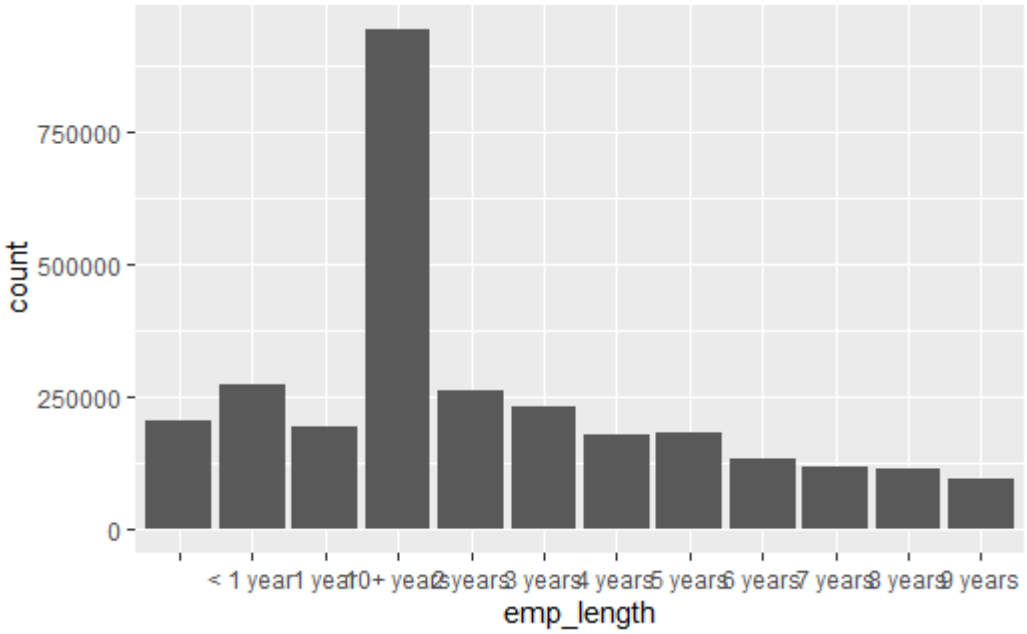
```
ggplot(data = data_new)+geom_bar(mapping = aes(x = term))
```



visualizing emp_length

Hide

```
ggplot(data = data_new)+geom_bar(mapping = aes(x = emp_length))
```



creating target Y variable coFlag

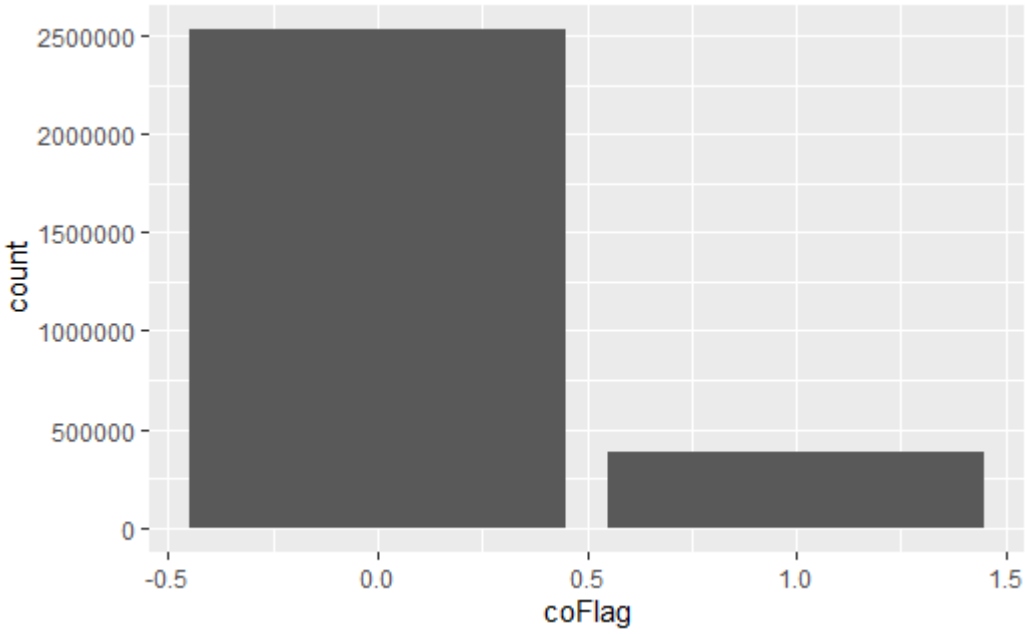
Hide

```
data_new$coFlag = ifelse (data_new$loan_status %in% c(loan_status_keep[1], loan_st
atus_keep[5], loan_status_keep[3]), 0, 1)
```

visualizing Y variable

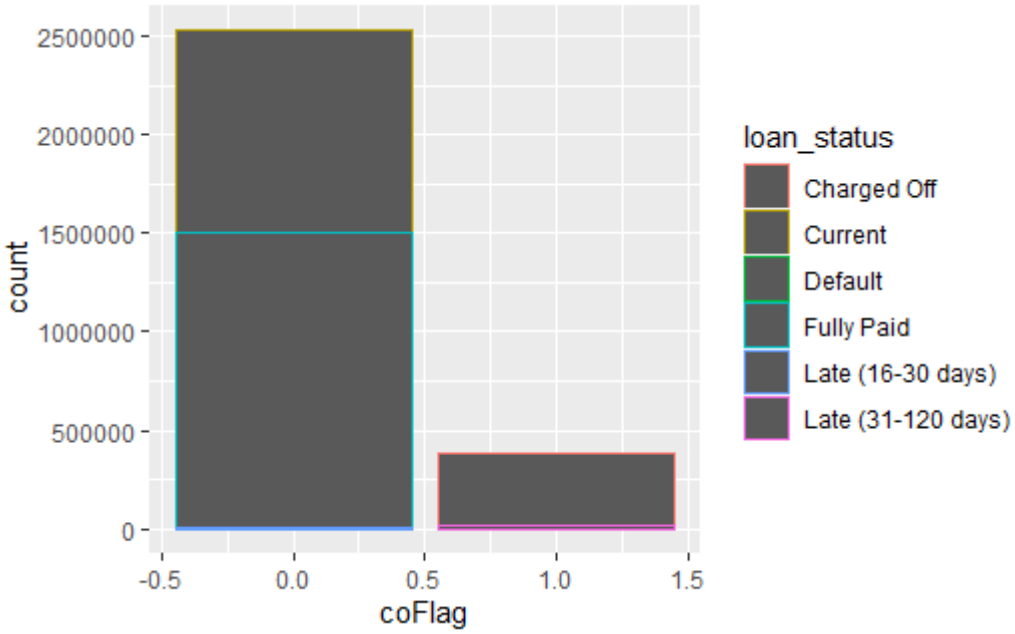
Hide

```
ggplot(data = data_new)+geom_bar(mapping = aes(x = coFlag))
```



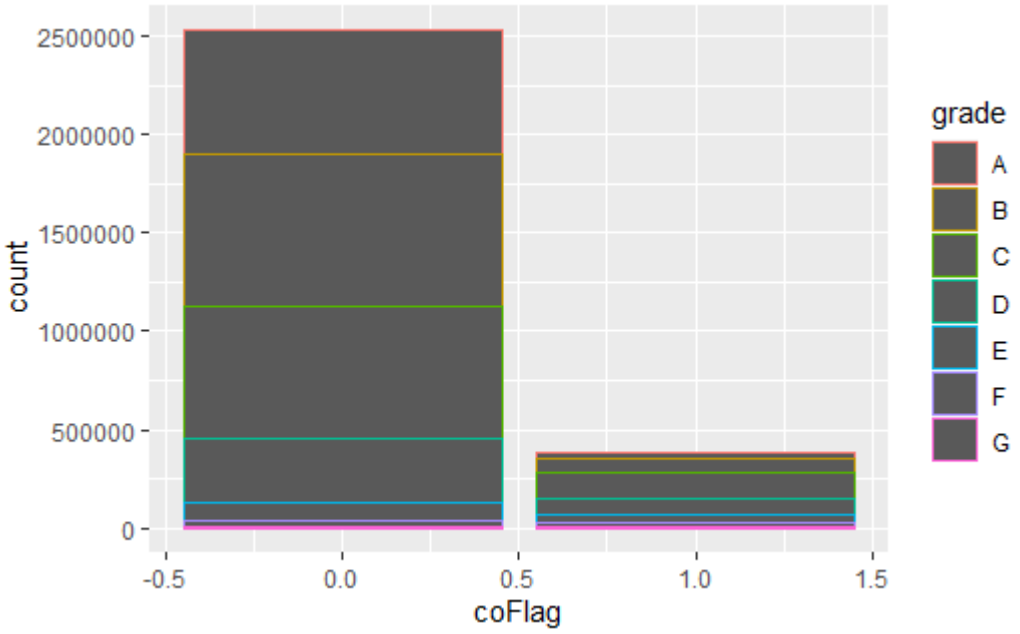
Hide

```
ggplot(data_new, mapping= aes(coFlag))+geom_bar(mapping = aes(color= loan_status))
```



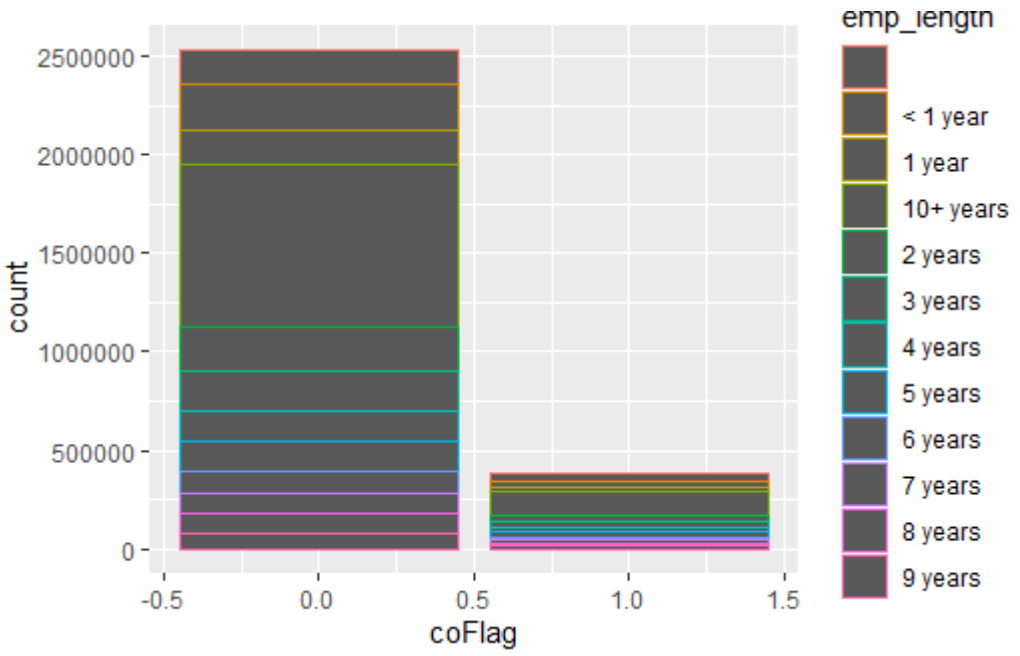
Hide

```
ggplot(data_new, mapping= aes(coFlag))+geom_bar(mapping = aes(color = grade))
```



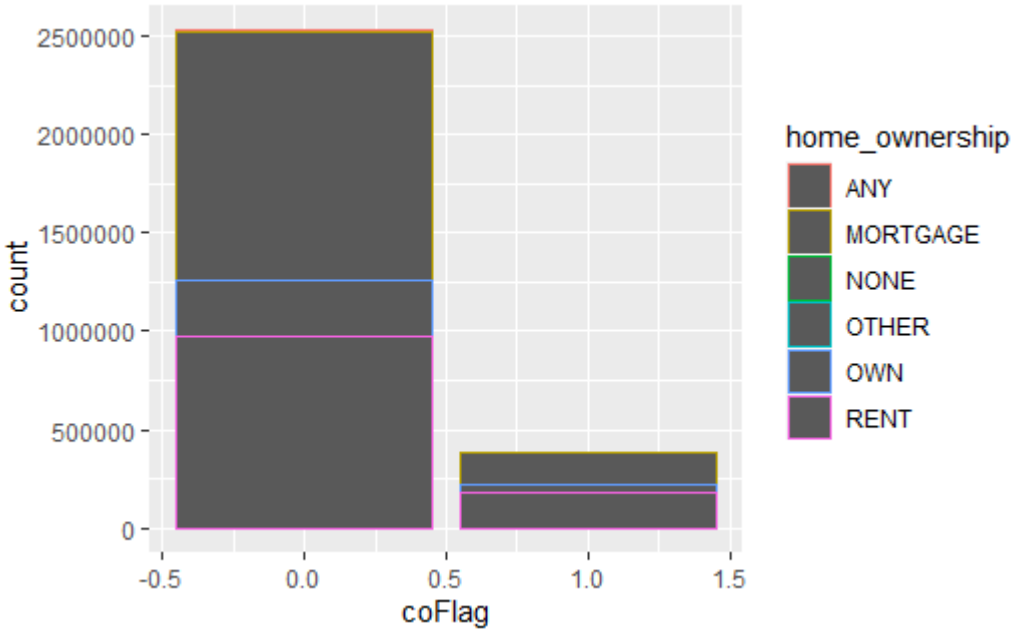
Hide

```
ggplot(data_new, mapping= aes(coFlag))+geom_bar(mapping = aes(color = emp_length))
```



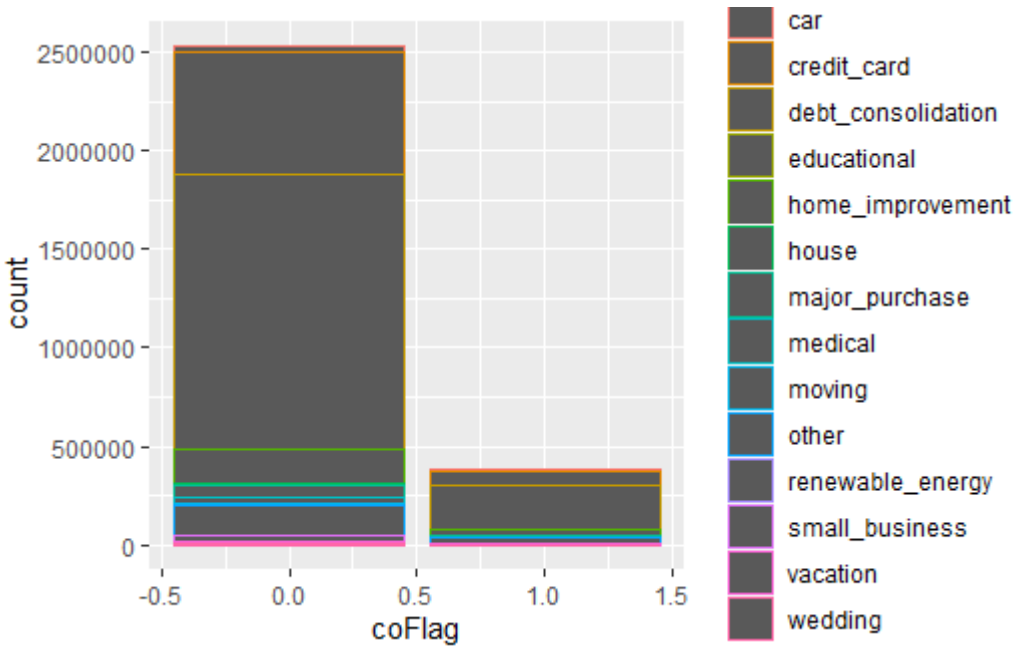
Hide

```
ggplot(data_new, mapping= aes(coFlag))+geom_bar(mapping = aes(color = home_ownersh
ip))
```



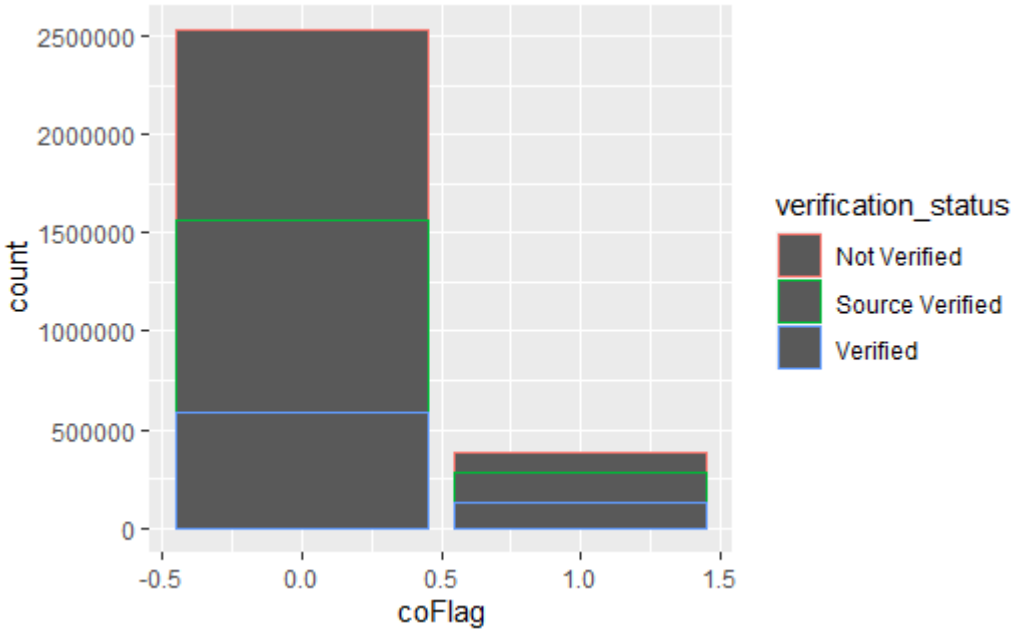
Hide

```
ggplot(data_new, mapping= aes(coFlag))+geom_bar(mapping = aes(color = purpose))
```



Hide

```
ggplot(data_new, mapping= aes(coFlag))+geom_bar(mapping = aes(color = verification
_status))
```



seperating continuous and categorical variables

Hide

```
all_vars <- unlist(lapply(data_new,class))

#continuous variable
cont_var = all_vars[all_vars=='integer' | all_vars == 'numeric']
cont_var = cont_var[c(1:71)]

#categorical variable
cat_var = all_vars[all_vars=='character']
```

Hide

```
data_cont_var = data_new[,names(data_new) %in% names(cont_var)]
summary(data_cont_var)
```

```
       id              loan_amnt      funded_amnt
 Min.   :    54734   Min.   :  500   Min.   :  500
 1st Qu.: 59110046   1st Qu.: 8000   1st Qu.: 8000
 Median :107381331   Median :13000   Median :13000
 Mean   : 97736221   Mean   :15356   Mean   :15352
 3rd Qu.:143124291   3rd Qu.:20000   3rd Qu.:20000
 Max.   :170998310   Max.   :40000   Max.   :40000

 funded_amnt_inv    int_rate       annual_inc
 Min.   :    0   Min.   : 5.31   Min.   :        0
 1st Qu.: 8000   1st Qu.: 9.17   1st Qu.:    47000
 Median :13000   Median :12.49   Median :    66000
 Mean   :15339   Mean   :13.04   Mean   :    79918
 3rd Qu.:20000   3rd Qu.:15.99   3rd Qu.:    95000
 Max.   :40000   Max.   :30.99   Max.   :110000000

      dti          delinq_2yrs      inq_last_6mths
 Min.   : -1.00   Min.   : 0.0000   Min.   :0.000
 1st Qu.: 12.08   1st Qu.: 0.0000   1st Qu.:0.000
```

```
Median : 18.10   Median : 0.0000   Median :0.000
Mean   : 19.30   Mean   : 0.2895   Mean   :0.551
3rd Qu.: 24.88   3rd Qu.: 0.0000   3rd Qu.:1.000
Max.   :999.00   Max.   :58.0000   Max.   :8.000
NA's   :3087                       NA's   :1
mths_since_last_delinq mths_since_last_record
Min.   :  0            Min.   :  0.0
1st Qu.: 17            1st Qu.: 57.0
Median : 32            Median : 77.0
Mean   : 35            Mean   : 74.8
3rd Qu.: 51            3rd Qu.: 96.0
Max.   :226            Max.   :129.0
NA's   :1529387        NA's   :2485522
   open_acc         pub_rec          revol_bal
Min.   :  0.00   Min.   : 0.0000   Min.   :      0
1st Qu.:  8.00   1st Qu.: 0.0000   1st Qu.:   5996
Median : 11.00   Median : 0.0000   Median :  11495
Mean   : 11.68   Mean   : 0.1765   Mean   :  16953
3rd Qu.: 15.00   3rd Qu.: 0.0000   3rd Qu.:  20641
Max.   :104.00   Max.   :86.0000   Max.   :2904836

  revol_util       total_acc      collections_12_mths_ex_med
Min.   :  0.00   Min.   :  2.00   Min.   : 0.00000
1st Qu.: 29.80   1st Qu.: 15.00   1st Qu.: 0.00000
Median : 48.60   Median : 22.00   Median : 0.00000
Mean   : 48.98   Mean   : 24.02   Mean   : 0.01769
3rd Qu.: 68.10   3rd Qu.: 31.00   3rd Qu.: 0.00000
Max.   :892.30   Max.   :176.00   Max.   :20.00000
NA's   :2599                      NA's   :56
mths_since_last_major_derog  tot_coll_amt
Min.   :  0.0                Min.   :      0
1st Qu.: 27.0                1st Qu.:      0
Median : 45.0                Median :      0
Mean   : 44.6                Mean   :    217
3rd Qu.: 62.0                3rd Qu.:      0
Max.   :226.0                Max.   :9152545
NA's   :2190900              NA's   :67527
 tot_cur_bal       open_acc_6m      open_act_il
Min.   :      0   Min.   : 0.0     Min.   : 0.0
1st Qu.:  29572   1st Qu.: 0.0     1st Qu.: 1.0
Median :  80537   Median : 1.0     Median : 2.0
Mean   : 145326   Mean   : 0.9     Mean   : 2.8
3rd Qu.: 217597   3rd Qu.: 1.0     3rd Qu.: 3.0
Max.   :9971659   Max.   :18.0     Max.   :78.0
NA's   :67527     NA's   :863116   NA's   :863115
 open_il_12m      open_il_24m     mths_since_rcnt_il
Min.   : 0.0     Min.   : 0.0     Min.   :  0.0
1st Qu.: 0.0     1st Qu.: 0.0     1st Qu.:  7.0
Median : 0.0     Median : 1.0     Median : 13.0
Mean   : 0.7     Mean   : 1.6     Mean   : 20.4
3rd Qu.: 1.0     3rd Qu.: 2.0     3rd Qu.: 23.0
Max.   :25.0     Max.   :51.0     Max.   :511.0
NA's   :863115   NA's   :863115   NA's   :923151
 total_bal_il        il_util         open_rv_12m
Min.   :      0   Min.   :   0      Min.   : 0.0
1st Qu.:   9136   1st Qu.:  55      1st Qu.: 0.0
Median :  23934   Median :  72      Median : 1.0
```

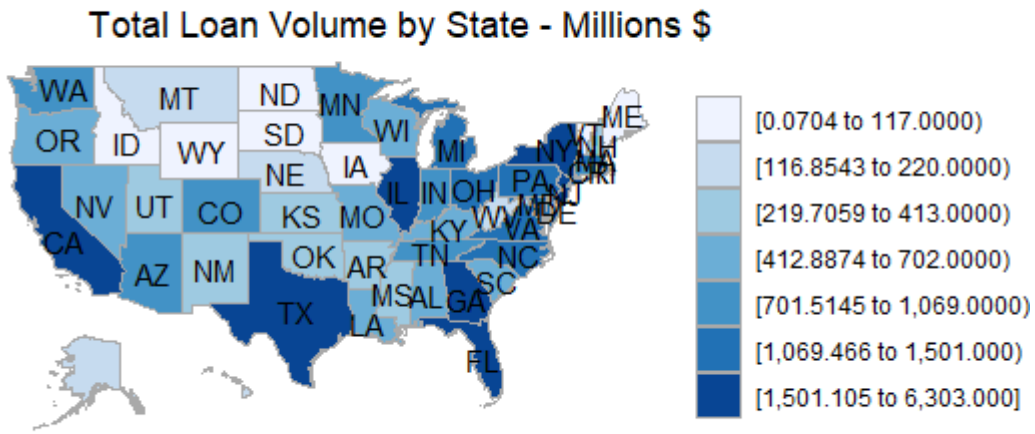
```
Mean   :  36818   Mean   :  69      Mean   : 1.3
3rd Qu.:  47638   3rd Qu.:  85      3rd Qu.: 2.0
Max.   :1990176   Max.   :1000      Max.   :28.0
NA's   :863115    NA's   :1152502   NA's   :863115
 open_rv_24m       max_bal_bc         all_util
Min.   : 0.0     Min.   :      0   Min.   :  0.0
1st Qu.: 1.0     1st Qu.:   2345   1st Qu.: 42.0
Median : 2.0     Median :   4568   Median : 57.0
Mean   : 2.6     Mean   :   6019   Mean   : 56.3
3rd Qu.: 4.0     3rd Qu.:   7913   3rd Qu.: 71.0
Max.   :60.0     Max.   :1170668   Max.   :239.0
NA's   :863115   NA's   :863115    NA's   :863468
total_rev_hi_lim      inq_fi        total_cu_tl
Min.   :      0   Min.   : 0.0     Min.   :  0.0
1st Qu.:  15400   1st Qu.: 0.0     1st Qu.:  0.0
Median :  26800   Median : 1.0     Median :  0.0
Mean   :  36335   Mean   : 1.1     Mean   :  1.5
3rd Qu.:  45600   3rd Qu.: 2.0     3rd Qu.:  2.0
Max.   :9999999   Max.   :48.0     Max.   :111.0
NA's   :67527     NA's   :863115   NA's   :863116
 inq_last_12m    acc_open_past_24mths  avg_cur_bal
Min.   : 0       Min.   : 0.00        Min.   :     0
1st Qu.: 0       1st Qu.: 2.00        1st Qu.:  3113
Median : 1       Median : 4.00        Median :  7406
Mean   : 2       Mean   : 4.49        Mean   : 13709
3rd Qu.: 3       3rd Qu.: 6.00        3rd Qu.: 18987
Max.   :67       Max.   :64.00        Max.   :958084
NA's   :863116   NA's   :47281        NA's   :67637
bc_open_to_buy      bc_util       chargeoff_within_12_mths
Min.   :     0   Min.   :  0.00   Min.   : 0.00000
1st Qu.:  2000   1st Qu.: 32.80   1st Qu.: 0.00000
Median :  6262   Median : 57.40   Median : 0.00000
Mean   : 12681   Mean   : 55.93   Mean   : 0.00801
3rd Qu.: 16128   3rd Qu.: 81.20   3rd Qu.: 0.00000
Max.   :781431   Max.   :339.60   Max.   :10.00000
NA's   :79746    NA's   :81212    NA's   :56
 delinq_amnt        mo_sin_old_il_acct mo_sin_old_rev_tl_op
Min.   :     0.00   Min.   :  0.0      Min.   :  1.0
1st Qu.:     0.00   1st Qu.: 93.0      1st Qu.:112.0
Median :     0.00   Median :130.0      Median :162.0
Mean   :     9.93   Mean   :125.9      Mean   :179.7
3rd Qu.:     0.00   3rd Qu.:155.0      3rd Qu.:230.0
Max.   :249925.00   Max.   :999.0      Max.   :999.0
                    NA's   :152559     NA's   :67528
mo_sin_rcnt_rev_tl_op mo_sin_rcnt_tl      mort_acc
Min.   :  0.00        Min.   :  0.00   Min.   : 0.00
1st Qu.:  4.00        1st Qu.:  3.00   1st Qu.: 0.00
Median :  9.00        Median :  6.00   Median : 1.00
Mean   : 14.46        Mean   :  8.34   Mean   : 1.51
3rd Qu.: 18.00        3rd Qu.: 11.00   3rd Qu.: 2.00
Max.   :564.00        Max.   :382.00   Max.   :94.00
NA's   :67528         NA's   :67527    NA's   :47281
mths_since_recent_bc mths_since_recent_bc_dlq
Min.   :  0.00       Min.   :  0.0
1st Qu.:  6.00       1st Qu.: 21.0
Median : 14.00       Median : 37.0
Mean   : 25.07       Mean   : 39.3
```

```
3rd Qu.: 30.00       3rd Qu.: 57.0
Max.   :675.00       Max.   :202.0
NA's   :77791        NA's   :2262938
mths_since_recent_inq mths_since_recent_revol_delinq
Min.   : 0.0          Min.   :  0.0
1st Qu.: 2.0          1st Qu.: 18.0
Median : 6.0          Median : 33.0
Mean   : 7.1          Mean   : 36.1
3rd Qu.:11.0          3rd Qu.: 51.0
Max.   :25.0          Max.   :212.0
NA's   :367844        NA's   :1984190
num_accts_ever_120_pd num_actv_bc_tl  num_actv_rev_tl
Min.   : 0.00         Min.   : 0.00   Min.   : 0.00
1st Qu.: 0.00         1st Qu.: 2.00   1st Qu.: 3.00
Median : 0.00         Median : 3.00   Median : 5.00
Mean   : 0.49         Mean   : 3.71   Mean   : 5.61
3rd Qu.: 0.00         3rd Qu.: 5.00   3rd Qu.: 7.00
Max.   :58.00         Max.   :50.00   Max.   :72.00
NA's   :67527         NA's   :67527   NA's   :67527
 num_bc_sats      num_bc_tl       num_il_tl
Min.   : 0.00   Min.   : 0.00   Min.   :  0.00
1st Qu.: 3.00   1st Qu.: 4.00   1st Qu.:  3.00
Median : 4.00   Median : 7.00   Median :  7.00
Mean   : 4.86   Mean   : 7.62   Mean   :  8.53
3rd Qu.: 6.00   3rd Qu.:10.00   3rd Qu.: 11.00
Max.   :77.00   Max.   :89.00   Max.   :159.00
NA's   :55841   NA's   :67527   NA's   :67527
num_op_rev_tl   num_rev_accts    num_rev_tl_bal_gt_0
Min.   : 0.00   Min.   :  0.00   Min.   : 0.00
1st Qu.: 5.00   1st Qu.:  8.00   1st Qu.: 3.00
Median : 7.00   Median : 12.00   Median : 5.00
Mean   : 8.27   Mean   : 13.78   Mean   : 5.56
3rd Qu.:11.00   3rd Qu.: 18.00   3rd Qu.: 7.00
Max.   :97.00   Max.   :151.00   Max.   :65.00
NA's   :67527   NA's   :67528    NA's   :67527
   num_sats      num_tl_120dpd_2m  num_tl_30dpd
Min.   :  0.00   Min.   :0        Min.   :0
1st Qu.:  8.00   1st Qu.:0        1st Qu.:0
Median : 11.00   Median :0        Median :0
Mean   : 11.69   Mean   :0        Mean   :0
3rd Qu.: 15.00   3rd Qu.:0        3rd Qu.:0
Max.   :104.00   Max.   :7        Max.   :4
NA's   :55841    NA's   :158568   NA's   :67527
num_tl_90g_dpd_24m num_tl_op_past_12m pct_tl_nvr_dlq
Min.   : 0.00      Min.   : 0.00      Min.   :  0.0
1st Qu.: 0.00      1st Qu.: 1.00      1st Qu.: 91.7
Median : 0.00      Median : 2.00      Median :100.0
Mean   : 0.08      Mean   : 2.07      Mean   : 94.3
3rd Qu.: 0.00      3rd Qu.: 3.00      3rd Qu.:100.0
Max.   :58.00      Max.   :32.00      Max.   :100.0
NA's   :67527      NA's   :67527      NA's   :67682
percent_bc_gt_75 pub_rec_bankruptcies   tax_liens
Min.   :  0.00   Min.   : 0.0000      Min.   : 0.00000
1st Qu.:  0.00   1st Qu.: 0.0000      1st Qu.: 0.00000
Median : 33.30   Median : 0.0000      Median : 0.00000
Mean   : 40.08   Mean   : 0.1228      Mean   : 0.03625
3rd Qu.: 66.70   3rd Qu.: 0.0000      3rd Qu.: 0.00000
```

```
 Max.   :100.00   Max.   :12.0000      Max.   :85.00000
 NA's   :80227    NA's   :697          NA's   :39
 tot_hi_cred_lim   total_bal_ex_mort total_bc_limit
 Min.   :      0   Min.   :      0   Min.   :      0
 1st Qu.:  52573   1st Qu.:  21334   1st Qu.:   8900
 Median : 118074   Median :  38828   Median :  17500
 Mean   : 183184   Mean   :  52635   Mean   :  24764
 3rd Qu.: 264682   3rd Qu.:  66422   3rd Qu.:  32500
 Max.   :9999999   Max.   :3408095   Max.   :1569000
 NA's   :67527     NA's   :47281     NA's   :47281
```

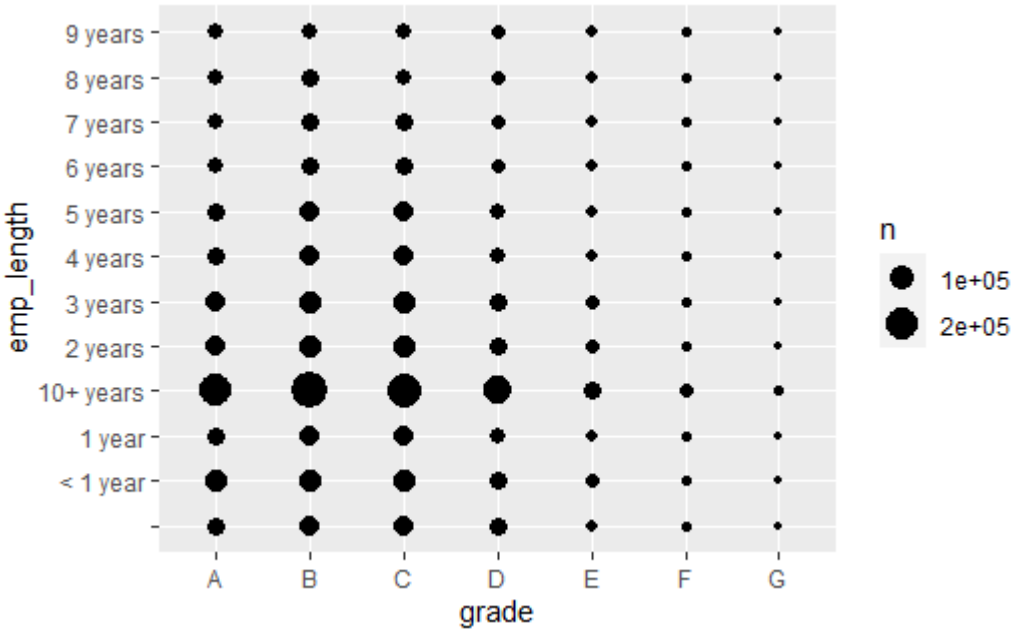
Hide

```
loan_by_state <- data_new %>%
                 group_by(addr_state) %>%
                 summarize(`Total Loans ($)` = sum(loan_amnt)/1e6) %>%
                 arrange(desc(`Total Loans ($)`))
colnames(loan_by_state) <- c("region","value")
top4_states <- round(100*sum(loan_by_state$value[1:4])/sum(loan_by_state$value),1)
data("state.regions")
loan_by_state$region <- sapply(loan_by_state$region, function(state_code) {
  inx <- grep(pattern = state_code, x = state.regions$abb)
  state.regions$region[inx]
})
print(state_choropleth(loan_by_state, title = "          Total Loan Volume by Sta
te - Millions $"))
```



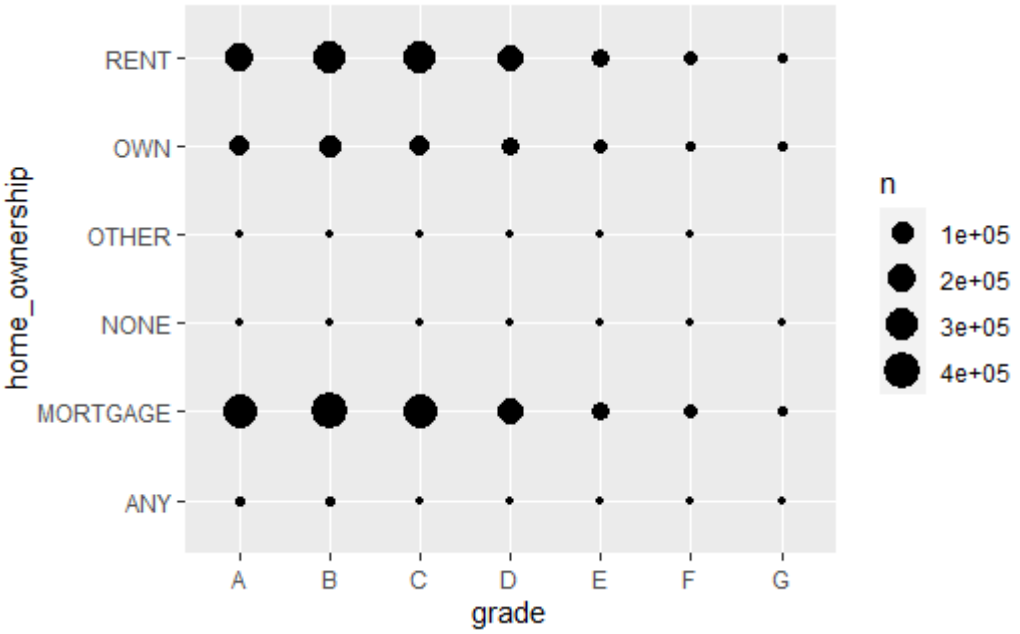
Hide

```
ggplot(data = data_new) + geom_count(mapping = aes(x = grade, y = emp_length))
```



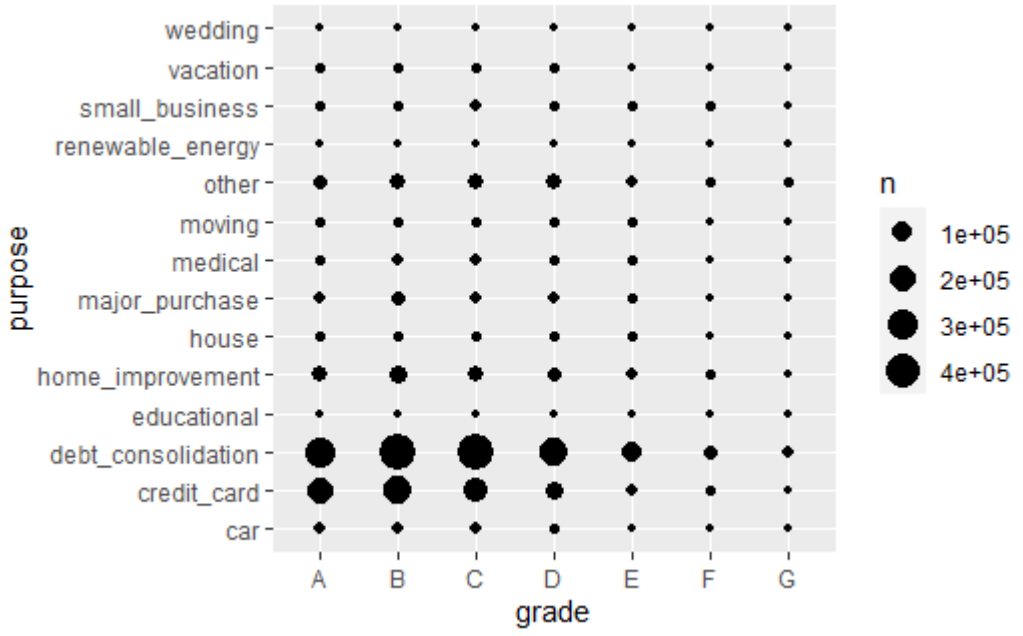
Hide

```
ggplot(data = data_new) + geom_count(mapping = aes(x = grade, y = home_ownership))
```



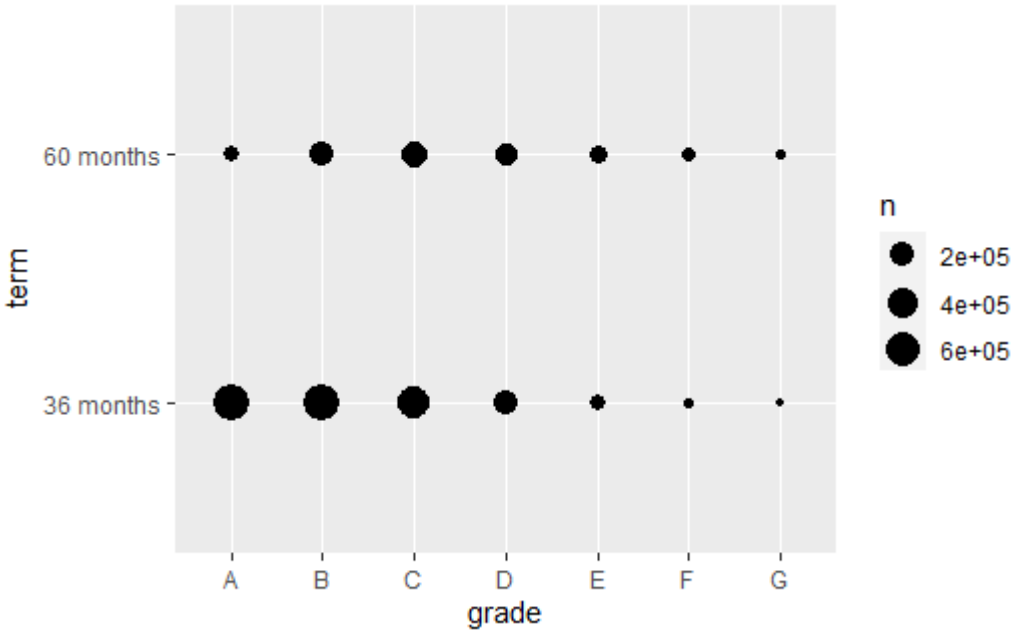
Hide

```
ggplot(data = data_new) + geom_count(mapping = aes(x = grade, y = purpose))
```



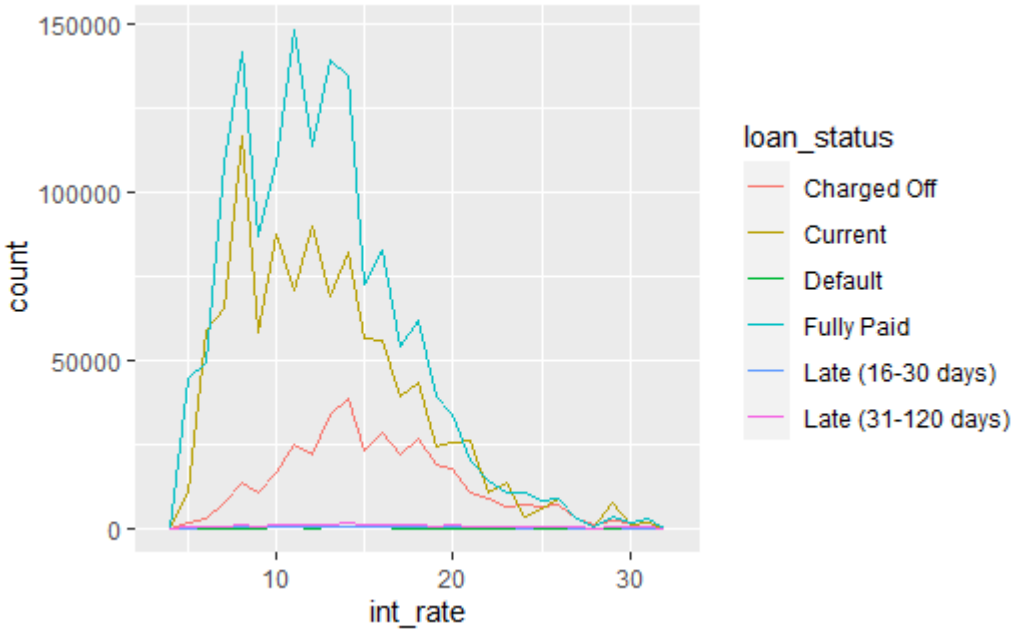
Hide

```
ggplot(data = data_new) + geom_count(mapping = aes(x = grade, y = term))
```



Hide

```
ggplot(data=data_new, mapping = aes(int_rate)) + geom_freqpoly(mapping = aes(color
= loan_status), binwidth =1)
```



correlation heatmap

Hide

```
m = data_new %>% mutate_if(is.character,as.factor)
m = m%>% mutate_if (is.factor, as.numeric)
m = m[,-1]
m = m[,-86]
m = cor(m)
print(m[c(1:17),c(1:17)])
```

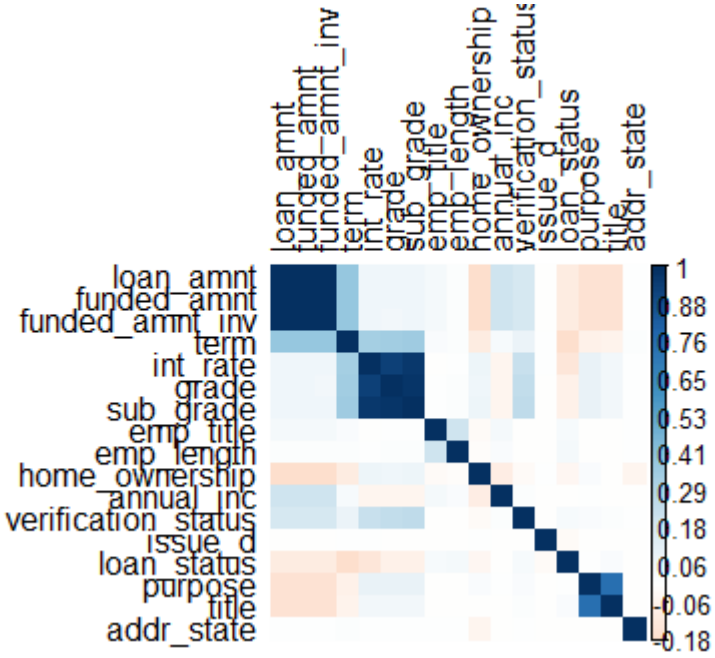
```
                      loan_amnt   funded_amnt
loan_amnt            1.0000000000  0.9998279337
funded_amnt          0.9998279337  1.0000000000
funded_amnt_inv      0.9994348058  0.9996425547
term                 0.3897116576  0.3894272962
int_rate             0.0615752556  0.0615821881
grade                0.0601622139  0.0600192255
sub_grade            0.0611919581  0.0610259709
emp_title            0.0494077365  0.0494075224
emp_length           0.0197373631  0.0196831665
home_ownership      -0.1763615042 -0.1763036601
annual_inc           0.2001104550  0.2000893419
verification_status  0.1745241820  0.1742639720
issue_d             -0.0007216838 -0.0007479279
loan_status         -0.1010566924 -0.1012436614
purpose             -0.1500151191 -0.1502233991
title               -0.1529339176 -0.1532377531
addr_state           0.0033956457  0.0034295410
                    funded_amnt_inv         term     int_rate
loan_amnt               0.999434806  0.389711658  0.061575256
funded_amnt             0.999642555  0.389427296  0.061582188
funded_amnt_inv         1.000000000  0.389664357  0.061684994
term                    0.389664357  1.000000000  0.338248840
int_rate                0.061684994  0.338248840  1.000000000
grade                   0.059685224  0.342609067  0.938696054
sub_grade               0.060688130  0.352013140  0.963573423
```

```
emp_title               0.049319407  0.026570191 -0.005943743
emp_length              0.019640373  0.016751570  0.003248659
home_ownership         -0.176316453 -0.109914980  0.071658572
annual_inc              0.200019767  0.035007698 -0.051091023
verification_status     0.174217441  0.088378056  0.222961468
issue_d                -0.000484331  0.008796418  0.003306612
loan_status            -0.101538836 -0.171121268 -0.130624341
purpose                -0.150668908 -0.070374629  0.093628952
title                  -0.153955431 -0.068201185  0.059362159
addr_state              0.003476023  0.013820570  0.002556297
                           grade    sub_grade    emp_title
loan_amnt             0.060162214  0.061191958  0.049407736
funded_amnt           0.060019226  0.061025971  0.049407522
funded_amnt_inv       0.059685224  0.060688130  0.049319407
term                  0.342609067  0.352013140  0.026570191
int_rate              0.938696054  0.963573423 -0.005943743
grade                 1.000000000  0.973055776  0.004307997
sub_grade             0.973055776  1.000000000  0.004561063
emp_title             0.004307997  0.004561063  1.000000000
emp_length            0.012263922  0.012868153  0.202912262
home_ownership        0.068254618  0.073196980 -0.028922884
annual_inc           -0.053843871 -0.056658459  0.040956913
verification_status   0.246145094  0.258762004 -0.003788841
issue_d              -0.005075624 -0.005494671 -0.004014341
loan_status          -0.074885616 -0.079180078  0.036596343
purpose               0.097236889  0.098499952 -0.006801566
title                 0.056466802  0.056894314 -0.003525159
addr_state            0.002685015  0.003293821  0.009656219
                     emp_length home_ownership   annual_inc
loan_amnt            0.019737363  -0.1763615042  0.200110455
funded_amnt          0.019683167  -0.1763036601  0.200089342
funded_amnt_inv      0.019640373  -0.1763164532  0.200019767
term                 0.016751570  -0.1099149796  0.035007698
int_rate             0.003248659   0.0716585723 -0.051091023
grade                0.012263922   0.0682546177 -0.053843871
sub_grade            0.012868153   0.0731969805 -0.056658459
emp_title            0.202912262  -0.0289228843  0.040956913
emp_length           1.000000000  -0.0107030746  0.023422116
home_ownership      -0.010703075   1.0000000000 -0.093576012
annual_inc           0.023422116  -0.0935760122  1.000000000
verification_status -0.006910360  -0.0225621661  0.017246187
issue_d             -0.006402532   0.0004425746 -0.000694591
loan_status          0.040605636  -0.0418431286 -0.002209451
purpose             -0.007421217   0.0213473490  0.002117876
title               -0.001581503  -0.0040539954 -0.002830728
addr_state           0.002482745  -0.0585061381 -0.005258259
                    verification_status       issue_d
loan_amnt                    0.174524182 -0.0007216838
funded_amnt                  0.174263972 -0.0007479279
funded_amnt_inv              0.174217441 -0.0004843310
term                         0.088378056  0.0087964183
int_rate                     0.222961468  0.0033066119
grade                        0.246145094 -0.0050756237
sub_grade                    0.258762004 -0.0054946706
emp_title                   -0.003788841 -0.0040143410
emp_length                  -0.006910360 -0.0064025321
home_ownership              -0.022562166  0.0004425746
```

```
annual_inc                  0.017246187 -0.0006945910
verification_status         1.000000000 -0.0082894246
issue_d                    -0.008289425  1.0000000000
loan_status                 0.037248163 -0.0227181196
purpose                     0.029923874 -0.0055780305
title                       0.015508448 -0.0004162859
addr_state                  0.002301992 -0.0008889079
                     loan_status      purpose         title
loan_amnt           -0.101056692 -0.150015119 -0.1529339176
funded_amnt         -0.101243661 -0.150223399 -0.1532377531
funded_amnt_inv     -0.101538836 -0.150668908 -0.1539554307
term                -0.171121268 -0.070374629 -0.0682011853
int_rate            -0.130624341  0.093628952  0.0593621591
grade               -0.074885616  0.097236889  0.0564668019
sub_grade           -0.079180078  0.098499952  0.0568943142
emp_title            0.036596343 -0.006801566 -0.0035251595
emp_length           0.040605636 -0.007421217 -0.0015815027
home_ownership      -0.041843129  0.021347349 -0.0040539954
annual_inc          -0.002209451  0.002117876 -0.0028307281
verification_status  0.037248163  0.029923874  0.0155084475
issue_d             -0.022718120 -0.005578030 -0.0004162859
loan_status          1.000000000  0.003544404  0.0237399136
purpose              0.003544404  1.000000000  0.7537665528
title                0.023739914  0.753766553  1.0000000000
addr_state          -0.001840127 -0.002536174 -0.0001388353
                      addr_state
loan_amnt           0.0033956457
funded_amnt         0.0034295410
funded_amnt_inv     0.0034760230
term                0.0138205699
int_rate            0.0025562970
grade               0.0026850145
sub_grade           0.0032938211
emp_title           0.0096562190
emp_length          0.0024827450
home_ownership     -0.0585061381
annual_inc         -0.0052582589
verification_status 0.0023019921
issue_d            -0.0008889079
loan_status        -0.0018401270
purpose            -0.0025361736
title              -0.0001388353
addr_state          1.0000000000
```

Hide

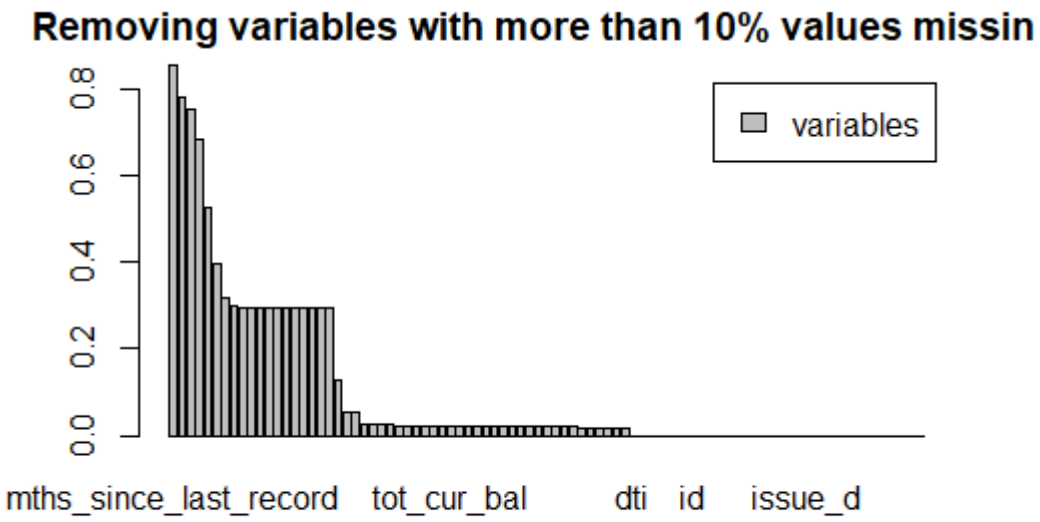
```
corrplot(m[c(1:17),c(1:17)], method="color", tl.col= 'black', is.corr = FALSE)
```



variable clustering / produces a dendogram

Hide

```
missing_values_new = sort(apply(is.na(data_new),2, sum )/nrow(data_new),TRUE)
barplot(missing_values_new,legend.text = 'variables', main='Removing variables wit
h more than 10% values missing')
```



Hide

```
drop_col_missing = names(missing_values_new[c(1:20)])
data_new = data_new[,!names(data_new) %in% drop_col_missing]

#xcat = data_new[,names(data_new) %in% names(cat_var[-1])]
#xcont = data_new[,names(data_new) %in% names(cont_var)]
#tree =kmeansvar(xcont, xcat, init = 4)
#plot(tree)
```