Lending Club Analysis

Max Kutschinski   Shree Karimkolikuzhiyil   Bradley Gravitt

3/9/2022

Missing Values (Part I)

The goal of this section is to identify features that are eligible for feature wise deletion in order to make the data set easier to navigate. Part II will discuss how to handle any remaining missing values.

```r
anyNA(data) # to see if there are any missing values in the dataset
```
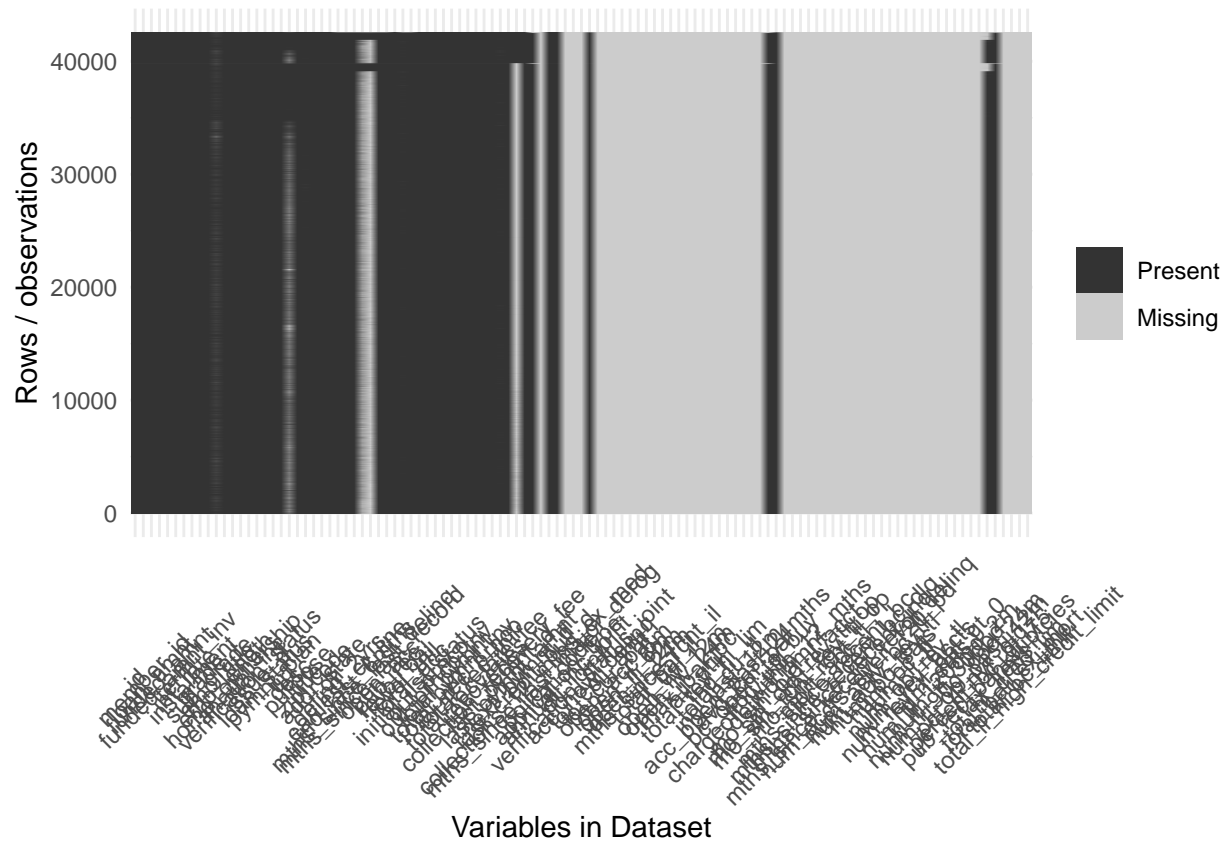
```
## [1] TRUE
```

```r
dim(data) # number of observations and features
```

```
## [1] 42537    111
```

```r
ggplot_missing(data) # visualize missing data
```

```
## Loading required package: reshape2
```

```
## Warning: package 'reshape2' was built under R version 4.0.3
```

The missingness plot indicates that a large amount of features consists of a high percentage of missing values. The following output shows the exact percentages per feature.

```
round(colMeans(is.na(data))*100,3)
```

```
##                              id                     member_id
##                           0.005                         0.005
##                       loan_amnt                   funded_amnt
##                           0.005                         0.005
##                 funded_amnt_inv                          term
##                           0.005                         0.005
##                        int_rate                   installment
##                           0.005                         0.005
##                           grade                     sub_grade
##                           0.005                         0.005
##                       emp_title                    emp_length
##                           6.159                         0.005
##                  home_ownership                    annual_inc
##                           0.005                         0.014
##             verification_status                       issue_d
##                           0.005                         0.005
##                     loan_status                     pymnt_plan
##                           0.005                         0.005
##                             url                          desc
##                           0.005                        31.782
##                         purpose                         title
```

```
##                     0.005                  0.031
##                  zip_code             addr_state
##                     0.005                  0.005
##                       dti            delinq_2yrs
##                     0.005                  0.073
##           earliest_cr_line        inq_last_6mths
##                     0.073                  0.073
##     mths_since_last_delinq mths_since_last_record
##                    63.305                 91.417
##                  open_acc                pub_rec
##                     0.073                  0.073
##                  revol_bal             revol_util
##                     0.005                  0.216
##                 total_acc    initial_list_status
##                     0.073                  0.005
##                 out_prncp           out_prncp_inv
##                     0.005                  0.005
##               total_pymnt         total_pymnt_inv
##                     0.005                  0.005
##            total_rec_prncp          total_rec_int
##                     0.005                  0.005
##           total_rec_late_fee            recoveries
##                     0.005                  0.005
##      collection_recovery_fee           last_pymnt_d
##                     0.005                  0.200
##             last_pymnt_amnt            next_pymnt_d
##                     0.005                 91.581
##           last_credit_pull_d collections_12_mths_ex_med
##                     0.014                  0.346
##  mths_since_last_major_derog            policy_code
##                   100.000                  0.005
##          application_type        annual_inc_joint
##                     0.005                100.000
##                  dti_joint verification_status_joint
##                   100.000                100.000
##             acc_now_delinq            tot_coll_amt
##                     0.073                100.000
##                tot_cur_bal             open_acc_6m
##                   100.000                100.000
##                 open_il_6m             open_il_12m
##                   100.000                100.000
##                open_il_24m         mths_since_rcnt_il
##                   100.000                100.000
##               total_bal_il                 il_util
##                   100.000                100.000
##                open_rv_12m             open_rv_24m
##                   100.000                100.000
##                 max_bal_bc                all_util
##                   100.000                100.000
##           total_rev_hi_lim                  inq_fi
##                   100.000                100.000
##                total_cu_tl            inq_last_12m
##                   100.000                100.000
##         acc_open_past_24mths             avg_cur_bal
```

```
##                       100.000                      100.000
##               bc_open_to_buy                      bc_util
##                       100.000                      100.000
##       chargeoff_within_12_mths                  delinq_amnt
##                         0.346                        0.073
##             mo_sin_old_il_acct            mo_sin_old_rev_tl_op
##                       100.000                      100.000
##           mo_sin_rcnt_rev_tl_op               mo_sin_rcnt_tl
##                       100.000                      100.000
##                       mort_acc           mths_since_recent_bc
##                       100.000                      100.000
##          mths_since_recent_bc_dlq          mths_since_recent_inq
##                       100.000                      100.000
## mths_since_recent_revol_delinq          num_accts_ever_120_pd
##                       100.000                      100.000
##                  num_actv_bc_tl                num_actv_rev_tl
##                       100.000                      100.000
##                     num_bc_sats                      num_bc_tl
##                       100.000                      100.000
##                       num_il_tl                   num_op_rev_tl
##                       100.000                      100.000
##                   num_rev_accts             num_rev_tl_bal_gt_0
##                       100.000                      100.000
##                       num_sats              num_tl_120dpd_2m
##                       100.000                      100.000
##                    num_tl_30dpd             num_tl_90g_dpd_24m
##                       100.000                      100.000
##              num_tl_op_past_12m                  pct_tl_nvr_dlq
##                       100.000                      100.000
##                 percent_bc_gt_75            pub_rec_bankruptcies
##                       100.000                        3.214
##                       tax_liens                tot_hi_cred_lim
##                         0.252                      100.000
##               total_bal_ex_mort                total_bc_limit
##                       100.000                      100.000
##        total_il_high_credit_limit
##                       100.000
```

Thus, it makes sense to delete features with a large percentage of missing values, say 33% or more.

```
data = data %>% select_if(~mean(is.na(.))<=0.33) # drop features with a lot of NAs
dim(data) # dimensions of altered dataset
```

```
## [1] 42537    54
```

Data structures

The following section of code explores the data structures in order to identify any qualitative features that might be coded as quantitative features and vice versa.

```r
table(sapply(data[1,],class)) # number of features per data type
```

```
## 
## character   logical   numeric 
##        22         1        31
```

```r
str(data) # overview of data types
```

```
## tibble [42,537 x 54] (S3: tbl_df/tbl/data.frame)
##  $ id                    : num [1:42537] 1077501 1077430 1077175 1076863 1075358 ...
##  $ member_id             : num [1:42537] 1296599 1314167 1313524 1277178 1311748 ...
##  $ loan_amnt             : num [1:42537] 5000 2500 2400 10000 3000 ...
##  $ funded_amnt           : num [1:42537] 5000 2500 2400 10000 3000 ...
##  $ funded_amnt_inv       : num [1:42537] 4975 2500 2400 10000 3000 ...
##  $ term                  : chr [1:42537] "36 months" "60 months" "36 months" "36 months" ...
##  $ int_rate              : chr [1:42537] "10.65%" "15.27%" "15.96%" "13.49%" ...
##  $ installment           : num [1:42537] 162.9 59.8 84.3 339.3 67.8 ...
##  $ grade                 : chr [1:42537] "B" "C" "C" "C" ...
##  $ sub_grade             : chr [1:42537] "B2" "C4" "C5" "C1" ...
##  $ emp_title             : chr [1:42537] NA "Ryder" NA "AIR RESOURCES BOARD" ...
##  $ emp_length            : chr [1:42537] "10+ years" "< 1 year" "10+ years" "10+ years" ...
##  $ home_ownership        : chr [1:42537] "RENT" "RENT" "RENT" "RENT" ...
##  $ annual_inc            : num [1:42537] 24000 30000 12252 49200 80000 ...
##  $ verification_status   : chr [1:42537] "Verified" "Source Verified" "Not Verified" "Source Veri
##  $ issue_d               : chr [1:42537] "Dec-11" "Dec-11" "Dec-11" "Dec-11" ...
##  $ loan_status           : chr [1:42537] "Fully Paid" "Charged Off" "Fully Paid" "Fully Paid" ..
##  $ pymnt_plan            : chr [1:42537] "n" "n" "n" "n" ...
##  $ url                   : chr [1:42537] "https://lendingclub.com/browse/loanDetail.action?loan_
##  $ desc                  : chr [1:42537] "Borrower added on 12/22/11 > I need to upgrade my busi
##  $ purpose               : chr [1:42537] "credit_card" "car" "small_business" "other" ...
##  $ title                 : chr [1:42537] "Computer" "bike" "real estate business" "personel" ...
##  $ zip_code              : chr [1:42537] "860xx" "309xx" "606xx" "917xx" ...
##  $ addr_state            : chr [1:42537] "AZ" "GA" "IL" "CA" ...
##  $ dti                   : num [1:42537] 27.65 1 8.72 20 17.94 ...
##  $ delinq_2yrs           : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
##  $ earliest_cr_line      : chr [1:42537] "Jan-85" "Apr-99" "Nov-01" "Feb-96" ...
##  $ inq_last_6mths        : num [1:42537] 1 5 2 1 0 3 1 2 2 0 ...
##  $ open_acc              : num [1:42537] 3 3 2 10 15 9 7 4 11 2 ...
##  $ pub_rec               : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
##  $ revol_bal             : num [1:42537] 13648 1687 2956 5598 27783 ...
##  $ revol_util            : chr [1:42537] "83.70%" "9.40%" "98.50%" "21%" ...
##  $ total_acc             : num [1:42537] 9 4 10 37 38 12 11 4 13 3 ...
##  $ initial_list_status   : logi [1:42537] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ out_prncp             : num [1:42537] 0 0 0 0 335 ...
##  $ out_prncp_inv         : num [1:42537] 0 0 0 0 335 ...
##  $ total_pymnt           : num [1:42537] 5863 1009 3006 12232 3717 ...
##  $ total_pymnt_inv       : num [1:42537] 5834 1009 3006 12232 3717 ...
##  $ total_rec_prncp       : num [1:42537] 5000 456 2400 10000 2665 ...
##  $ total_rec_int         : num [1:42537] 863 435 606 2215 1052 ...
##  $ total_rec_late_fee    : num [1:42537] 0 0 0 17 0 ...
##  $ recoveries            : num [1:42537] 0 117 0 0 0 ...
##  $ collection_recovery_fee : num [1:42537] 0 1.11 0 0 0 0 0 0 2.09 2.52 ...
```

```
## $ last_pymnt_d             : chr [1:42537] "Jan-15" "Apr-13" "Jun-14" "Jan-15" ...
## $ last_pymnt_amnt          : num [1:42537] 171.6 119.7 649.9 357.5 67.8 ...
## $ last_credit_pull_d       : chr [1:42537] "Jul-16" "Sep-13" "Jul-16" "Apr-16" ...
## $ collections_12_mths_ex_med: num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
## $ policy_code              : num [1:42537] 1 1 1 1 1 1 1 1 1 1 1 ...
## $ application_type         : chr [1:42537] "INDIVIDUAL" "INDIVIDUAL" "INDIVIDUAL" "INDIVIDUAL" ...
## $ acc_now_delinq           : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
## $ chargeoff_within_12_mths : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
## $ delinq_amnt              : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
## $ pub_rec_bankruptcies     : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
## $ tax_liens                : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
## - attr(*, "problems")= tibble [1 x 5] (S3: tbl_df/tbl/data.frame)
##  ..$ row     : int 39788
##  ..$ col     : chr "id"
##  ..$ expected: chr "a double"
##  ..$ actual  : chr "Loans that do not meet the credit policy"
##  ..$ file    : chr "'Data/LoanStats3a.csv'"
```

```r
sapply(data,function(x){ length(unique(x))})
```

```
##                     id              member_id
##                  42536                  42536
##              loan_amnt            funded_amnt
##                    899                   1052
##         funded_amnt_inv                  term
##                   9242                      3
##               int_rate            installment
##                    395                  16460
##                  grade              sub_grade
##                      8                     36
##              emp_title             emp_length
##                  30449                     13
##          home_ownership             annual_inc
##                      6                   5598
##    verification_status                issue_d
##                      4                     56
##            loan_status             pymnt_plan
##                     10                      3
##                    url                   desc
##                  42536                  28951
##                purpose                  title
##                     15                  20965
##               zip_code             addr_state
##                    838                     51
##                    dti            delinq_2yrs
##                   2895                     13
##       earliest_cr_line         inq_last_6mths
##                    531                     29
##               open_acc                pub_rec
##                     45                      7
##              revol_bal             revol_util
##                  22710                   1120
##              total_acc     initial_list_status
##                     84                      2
```

```
##                  out_prncp                  out_prncp_inv
##                       831                            833
##                total_pymnt                total_pymnt_inv
##                     40610                          40118
##             total_rec_prncp               total_rec_int
##                      8471                          37582
##          total_rec_late_fee                  recoveries
##                      1562                           4518
##      collection_recovery_fee                 last_pymnt_d
##                      2843                            106
##              last_pymnt_amnt            last_credit_pull_d
##                     37080                            111
## collections_12_mths_ex_med                 policy_code
##                         2                              2
##             application_type               acc_now_delinq
##                         2                              3
##      chargeoff_within_12_mths                 delinq_amnt
##                         2                              4
##           pub_rec_bankruptcies                   tax_liens
##                         4                              3
```

There are some qualitative features, some coded as character and some as numeric, that need to be converted to factors. Furthermore, it seems like some features only have one value (besides NA) and should therefore be dropped.

```r
##### Dates converted to factors for now

#issue_d: date
#last_pymnt_d: date
#last_credit_pull_d: date
#earliest_cr_line: date

#####

# features with only one value:
#     collections_12_mths_ex_med,
#     application_type,
#     policy_code,
#     chargeoff_within_12_mths


# get rid of percent signs and convert to numeric
data$revol_util = as.numeric(sub("%","",data$revol_util))
data$int_rate = as.numeric(sub("%","",data$int_rate))

#convert data types
dataQual = data %>% select(c(term,grade,sub_grade,home_ownership,verification_status,loan_status,pymnt_p
                    purpose, initial_list_status, addr_state, zip_code, id, member_id, emp_title,t:
                    last_credit_pull_d,earliest_cr_line, emp_length, url)) %>% mutate_all(factor)
dataQuan = data %>% select(-c(names(dataQual),collections_12_mths_ex_med,application_type,policy_code,cl

#final result
str(dataQual)
```

7

```
## tibble [42,537 x 21] (S3: tbl_df/tbl/data.frame)
##  $ term               : Factor w/ 2 levels "36 months","60 months": 1 2 1 1 2 1 2 1 2 2 ...
##  $ grade              : Factor w/ 7 levels "A","B","C","D",..: 2 3 3 3 2 1 3 5 6 2 ...
##  $ sub_grade          : Factor w/ 35 levels "A1","A2","A3",..: 7 14 15 11 10 4 15 21 27 10 ...
##  $ home_ownership     : Factor w/ 5 levels "MORTGAGE","NONE",..: 5 5 5 5 5 5 5 5 5 4 5 ...
##  $ verification_status: Factor w/ 3 levels "Not Verified",..: 3 2 1 2 2 2 1 2 2 3 ...
##  $ loan_status        : Factor w/ 9 levels "Charged Off",..: 6 1 6 6 2 6 6 6 1 1 ...
##  $ pymnt_plan         : Factor w/ 2 levels "n","y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ purpose            : Factor w/ 14 levels "car","credit_card",..: 2 1 12 10 10 14 3 1 12 10 ...
##  $ initial_list_status: Factor w/ 1 level "FALSE": 1 1 1 1 1 1 1 1 1 1 ...
##  $ addr_state         : Factor w/ 50 levels "AK","AL","AR",..: 4 11 15 5 37 4 28 5 5 43 ...
##  $ zip_code           : Factor w/ 837 levels "007xx","010xx",..: 727 281 513 764 813 721 252 749 802
##  $ id                 : Factor w/ 42535 levels "54734","55521",..: 42535 42534 42533 42532 42531 4253
##  $ member_id          : Factor w/ 42535 levels "70473","70626",..: 42206 42535 42534 40932 42533 4253
##  $ emp_title          : Factor w/ 30448 levels "$260M '06 vintage technology venture capital firm",.
##  $ title              : Factor w/ 20964 levels "'08 & '09 Roth IRA Investments",..: 3614 1817 16941
##  $ issue_d            : Factor w/ 55 levels "Apr-08","Apr-09",..: 14 14 14 14 14 14 14 14 14 14 ...
##  $ last_pymnt_d       : Factor w/ 105 levels "Apr-08","Apr-09",..: 44 6 61 44 18 44 81 44 5 86 ...
##  $ last_credit_pull_d : Factor w/ 110 levels "Apr-09","Apr-10",..: 55 108 55 8 55 45 84 26 14 71 ...
##  $ earliest_cr_line   : Factor w/ 530 levels "Apr-00","Apr-01",..: 202 44 390 171 213 393 222 182 5 4
##  $ emp_length         : Factor w/ 12 levels "< 1 year","1 year",..: 3 1 3 3 2 5 10 11 6 1 ...
##  $ url                : Factor w/ 42535 levels "https://lendingclub.com/browse/loanDetail.action?loan
##  - attr(*, "problems")= tibble [1 x 5] (S3: tbl_df/tbl/data.frame)
##   ..$ row     : int 39788
##   ..$ col     : chr "id"
##   ..$ expected: chr "a double"
##   ..$ actual  : chr "Loans that do not meet the credit policy"
##   ..$ file    : chr "'Data/LoanStats3a.csv'"
```

str(dataQuan)

```
## tibble [42,537 x 29] (S3: tbl_df/tbl/data.frame)
##  $ loan_amnt         : num [1:42537] 5000 2500 2400 10000 3000 ...
##  $ funded_amnt       : num [1:42537] 5000 2500 2400 10000 3000 ...
##  $ funded_amnt_inv   : num [1:42537] 4975 2500 2400 10000 3000 ...
##  $ int_rate          : num [1:42537] 10.6 15.3 16 13.5 12.7 ...
##  $ installment       : num [1:42537] 162.9 59.8 84.3 339.3 67.8 ...
##  $ annual_inc        : num [1:42537] 24000 30000 12252 49200 80000 ...
##  $ desc              : chr [1:42537] "Borrower added on 12/22/11 > I need to upgrade my busines
##  $ dti               : num [1:42537] 27.65 1 8.72 20 17.94 ...
##  $ delinq_2yrs       : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
##  $ inq_last_6mths    : num [1:42537] 1 5 2 1 0 3 1 2 2 0 ...
##  $ open_acc          : num [1:42537] 3 3 2 10 15 9 7 4 11 2 ...
##  $ pub_rec           : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
##  $ revol_bal         : num [1:42537] 13648 1687 2956 5598 27783 ...
##  $ revol_util        : num [1:42537] 83.7 9.4 98.5 21 53.9 28.3 85.6 87.5 32.6 36.5 ...
##  $ total_acc         : num [1:42537] 9 4 10 37 38 12 11 4 13 3 ...
##  $ out_prncp         : num [1:42537] 0 0 0 0 335 ...
##  $ out_prncp_inv     : num [1:42537] 0 0 0 0 335 ...
##  $ total_pymnt       : num [1:42537] 5863 1009 3006 12232 3717 ...
##  $ total_pymnt_inv   : num [1:42537] 5834 1009 3006 12232 3717 ...
##  $ total_rec_prncp   : num [1:42537] 5000 456 2400 10000 2665 ...
##  $ total_rec_int     : num [1:42537] 863 435 606 2215 1052 ...
##  $ total_rec_late_fee: num [1:42537] 0 0 0 17 0 ...
```

```
##  $ recoveries            : num [1:42537] 0 117 0 0 0 ...
##  $ collection_recovery_fee: num [1:42537] 0 1.11 0 0 0 0 0 0 2.09 2.52 ...
##  $ last_pymnt_amnt        : num [1:42537] 171.6 119.7 649.9 357.5 67.8 ...
##  $ acc_now_delinq         : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
##  $ delinq_amnt            : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
##  $ pub_rec_bankruptcies   : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
##  $ tax_liens              : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "problems")= tibble [1 x 5] (S3: tbl_df/tbl/data.frame)
##   ..$ row     : int 39788
##   ..$ col     : chr "id"
##   ..$ expected: chr "a double"
##   ..$ actual  : chr "Loans that do not meet the credit policy"
##   ..$ file    : chr "'Data/LoanStats3a.csv'"
```

Missing Values (Part II)

Before using any kind of imputation method the data is split into a training and test set.

```
#dataImpute = preProcess(data,method = 'medianImpute', k = 5)
```

Removing correlated variables

```
#datacorr = cor(data)
#corrplot(datacorr, order= 'hclust', t1.cex= .35)
#highCorr = findCorrelation(datacorr, .85, verbose=T, names=T)
#dataRemoveCorr= select(all_of(data), -any_of((highCorr)))
#dim(dataRemoveCorr)
```