Analisis Data Eksploratif (EDA) Dataset Lending Club

Muhammad Azizul Hakim

7/13/2019

1. Overview Data

Sebelum memulai analisis data, sebaiknya kita lihat gambaran umum dari dataset yang kita miliki, untuk memberikan gambaran mengenai analisis yang akan kita lakukan selanjutnya.

Load dataset, dan tampilkan 5 observasi teratas:

```
loan <- readRDS("lending_club_loan_data.rds")
```

Tampilkan seluruh nama kolom yang ada beserta struktur datanya:

```
str(loan)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 707838 obs. of  144 variables:
##  $ member_id                                  : logi  NA NA NA NA NA NA ...
##  $ loan_amnt                                  : num  24700 35000 20000 20000 14025 ...
##  $ funded_amnt                                : num  24700 35000 20000 20000 14025 ...
##  $ funded_amnt_inv                            : num  24700 35000 20000 20000 14025 ...
##  $ term                                       : chr  "36 months" "60 months" "60 months" "36 months"
##  $ int_rate                                   : num  12 12.9 14 12.9 18.5 ...
##  $ installment                                : num  820 794 465 673 360 ...
##  $ grade                                      : chr  "C" "C" "C" "C" ...
##  $ sub_grade                                  : chr  "C1" "C2" "C4" "C2" ...
##  $ emp_title                                  : chr  "Engineer" "Lieutenant" "Facilities Coordinator"
##  $ emp_length                                 : chr  "10+ years" "7 years" "10+ years" "10+ years" ..
##  $ home_ownership                             : chr  "MORTGAGE" "MORTGAGE" "OWN" "MORTGAGE" ...
##  $ annual_inc                                 : num  65000 106000 70000 145000 39000 ...
##  $ verification_status                        : chr  "Not Verified" "Source Verified" "Source Verifie
##  $ issue_d                                    : chr  "Dec-2015" "Dec-2015" "Dec-2015" "Dec-2015" ...
##  $ loan_status                                : chr  "Fully Paid" "Charged Off" "Charged Off" "Late (
##  $ pymnt_plan                                 : chr  "n" "n" "n" "n" ...
##  $ url                                        : logi  NA NA NA NA NA NA ...
##  $ desc                                       : chr  NA NA NA NA ...
##  $ purpose                                    : chr  "small_business" "debt_consolidation" "debt_cons
##  $ title                                      : chr  "Business" "Debt consolidation" "Debt consolidat
##  $ zip_code                                   : chr  "577xx" "351xx" "210xx" "029xx" ...
##  $ addr_state                                 : chr  "SD" "AL" "MD" "RI" ...
##  $ dti                                        : num  16.1 17.4 16.9 12.3 18 ...
##  $ delinq_2yrs                                : num  1 0 0 0 0 0 0 0 0 0 ...
##  $ earliest_cr_line                           : chr  "Dec-1999" "Apr-2002" "Jun-2001" "Feb-2004" ...
##  $ inq_last_6mths                             : num  4 0 0 0 1 0 0 0 1 0 ...
##  $ mths_since_last_delinq                     : num  6 NA 33 NA NA 39 27 NA 30 NA ...
##  $ mths_since_last_record                     : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ open_acc                                   : num  22 8 20 12 12 10 7 6 7 13 ...
##  $ pub_rec                                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ revol_bal                                  : num  21470 39055 31200 22551 15646 ...
##  $ revol_util                                 : num  19.2 72.1 42 80.2 74.9 51.2 52.3 56.2 29.7 11.6
##  $ total_acc                                  : num  38 27 35 21 21 24 13 18 13 17 ...
##  $ initial_list_status                        : chr  "w" "w" "w" "w" ...
```

```
##  $ out_prncp                          : num  0 0 0 673 0 ...
##  $ out_prncp_inv                      : num  0 0 0 673 0 ...
##  $ total_pymnt                        : num  25680 22208 15473 23510 16341 ...
##  $ total_pymnt_inv                    : num  25680 22208 15473 23510 16341 ...
##  $ total_rec_prncp                    : num  24700 10324 2965 19327 14025 ...
##  $ total_rec_int                      : num  980 7086 2593 4183 2316 ...
##  $ total_rec_late_fee                 : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ recoveries                         : num  0 4798 9915 0 0 ...
##  $ collection_recovery_fee            : num  0 864 1785 0 0 ...
##  $ last_pymnt_d                       : chr  "Jun-2016" "Nov-2017" "Jan-2017" "Dec-2018" ...
##  $ last_pymnt_amnt                    : num  926 794 465 673 12764 ...
##  $ next_pymnt_d                       : chr  NA NA NA "Mar-2019" ...
##  $ last_credit_pull_d                 : chr  "Feb-2019" "Feb-2019" "Aug-2018" "Feb-2019" ...
##  $ collections_12_mths_ex_med         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ mths_since_last_major_derog        : num  NA NA 69 NA NA 61 NA NA 30 NA ...
##  $ policy_code                        : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ application_type                   : chr  "Individual" "Individual" "Individual" "Individua
##  $ annual_inc_joint                   : num  NA NA NA NA NA NA NA 71000 NA NA ...
##  $ dti_joint                          : num  NA NA NA NA NA ...
##  $ verification_status_joint          : chr  NA NA NA NA ...
##  $ acc_now_delinq                     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ tot_coll_amt                       : num  0 0 0 0 0 0 0 0 722 0 ...
##  $ tot_cur_bal                        : num  204396 299890 34856 429218 60381 ...
##  $ open_acc_6m                        : num  1 0 2 2 1 3 0 0 2 1 ...
##  $ open_act_il                        : num  1 3 1 3 0 3 3 1 2 1 ...
##  $ open_il_12m                        : num  0 0 0 1 0 1 0 0 0 0 ...
##  $ open_il_24m                        : num  1 0 0 2 0 2 1 4 1 1 ...
##  $ mths_since_rcnt_il                 : num  19 35 43 2 91 3 18 19 21 23 ...
##  $ total_bal_il                       : num  18005 16012 3656 9935 0 ...
##  $ il_util                            : num  73 43 37 43 NA 82 73 73 36 70 ...
##  $ open_rv_12m                        : num  2 0 2 2 2 2 1 0 3 1 ...
##  $ open_rv_24m                        : num  3 0 3 3 5 4 2 2 3 1 ...
##  $ max_bal_bc                         : num  6472 9701 4367 4627 4879 ...
##  $ all_util                           : num  29 60 41 62 75 71 64 65 34 45 ...
##  $ total_rev_hi_lim                   : num  111800 54200 74200 28500 20900 ...
##  $ inq_fi                             : num  0 0 0 1 0 1 1 2 3 0 ...
##  $ total_cu_tl                        : num  0 4 1 0 0 1 0 5 1 1 ...
##  $ inq_last_12m                       : num  6 0 0 6 2 0 3 1 4 0 ...
##  $ acc_open_past_24mths               : num  4 0 3 5 5 6 3 6 4 2 ...
##  $ avg_cur_bal                        : num  9733 37486 1835 39020 5032 ...
##  $ bc_open_to_buy                     : num  57830 12875 18527 2822 578 ...
##  $ bc_util                            : num  27.1 67.2 44 81.6 93.4 63 52.3 55.9 37.2 12.1 ..
##  $ chargeoff_within_12_mths           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ delinq_amnt                        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ mo_sin_old_il_acct                 : num  113 135 46 79 138 165 106 125 148 36 ...
##  $ mo_sin_old_rev_tl_op               : num  192 164 174 142 172 159 163 184 128 87 ...
##  $ mo_sin_rcnt_rev_tl_op              : num  2 38 5 4 2 3 11 14 3 2 ...
##  $ mo_sin_rcnt_tl                     : num  2 35 5 2 2 3 11 14 3 2 ...
##  $ mort_acc                           : num  4 4 0 5 3 3 0 5 1 1 ...
##  $ mths_since_recent_bc               : num  2 38 6 27 2 3 11 101 4 2 ...
##  $ mths_since_recent_bc_dlq           : num  NA NA NA NA NA NA 30 NA 69 NA ...
##  $ mths_since_recent_inq              : num  0 NA 13 1 2 17 11 10 4 NA ...
##  $ mths_since_recent_revol_delinq     : num  6 NA NA NA NA 39 27 NA 69 NA ...
##  $ num_accts_ever_120_pd              : num  0 0 1 0 0 1 0 0 2 0 ...
```

```
##  $ num_actv_bc_tl                            : num  5 3 6 4 3 4 2 2 2 4 ...
##  $ num_actv_rev_tl                           : num  5 4 11 6 11 5 2 3 4 5 ...
##  $ num_bc_sats                               : num  13 3 9 5 3 4 4 2 2 8 ...
##  $ num_bc_tl                                 : num  17 6 19 8 3 7 5 4 5 10 ...
##  $ num_il_tl                                 : num  6 13 2 5 4 9 6 6 3 2 ...
##  $ num_op_rev_tl                             : num  20 4 19 7 11 6 4 4 4 10 ...
##  $ num_rev_accts                             : num  27 10 32 11 13 12 7 7 9 13 ...
##  $ num_rev_tl_bal_gt_0                       : num  5 4 11 6 11 6 2 3 4 5 ...
##  $ num_sats                                  : num  22 8 20 12 12 10 7 6 7 13 ...
##  $ num_tl_120dpd_2m                          : num  0 0 0 0 0 0 0 0 0 0 ...
##   [list output truncated]
```

Buat summary statistik data secara umum:

```
summary(loan)
```

```
##  member_id        loan_amnt      funded_amnt    funded_amnt_inv
##  Mode:logical   Min.   :  500   Min.   :  500   Min.   :    0
##  NA's:707838    1st Qu.: 8000   1st Qu.: 8000   1st Qu.: 8000
##                 Median :13000   Median :13000   Median :13000
##                 Mean   :14768   Mean   :14755   Mean   :14723
##                 3rd Qu.:20000   3rd Qu.:20000   3rd Qu.:20000
##                 Max.   :35000   Max.   :35000   Max.   :35000
##
##      term              int_rate      installment         grade
##  Length:707838      Min.   : 5.32   Min.   :   4.93   Length:707838
##  Class :character   1st Qu.: 9.99   1st Qu.: 261.30   Class :character
##  Mode  :character   Median :12.99   Median : 382.55   Mode  :character
##                     Mean   :13.24   Mean   : 437.01
##                     3rd Qu.:16.20   3rd Qu.: 572.61
##                     Max.   :28.99   Max.   :1445.46
##
##   sub_grade          emp_title          emp_length
##  Length:707838      Length:707838      Length:707838
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  home_ownership       annual_inc      verification_status
##  Length:707838      Min.   :      0   Length:707838
##  Class :character   1st Qu.:  45000   Class :character
##  Mode  :character   Median :  65000   Mode  :character
##                     Mean   :  74996
##                     3rd Qu.:  90000
##                     Max.   :9500000
##
##    issue_d          loan_status         pymnt_plan          url
##  Length:707838      Length:707838      Length:707838      Mode:logical
##  Class :character   Class :character   Class :character   NA's:707838
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
##
##      desc             purpose             title
##  Length:707838      Length:707838      Length:707838
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    zip_code          addr_state             dti          delinq_2yrs
##  Length:707838      Length:707838      Min.   :  0.00   Min.   : 0.0000
##  Class :character   Class :character   1st Qu.: 11.92   1st Qu.: 0.0000
##  Mode  :character   Mode  :character   Median : 17.67   Median : 0.0000
##                                        Mean   : 18.15   Mean   : 0.3145
##                                        3rd Qu.: 23.96   3rd Qu.: 0.0000
##                                        Max.   :672.52   Max.   :30.0000
##                                        NA's   :1
##  earliest_cr_line   inq_last_6mths   mths_since_last_delinq
##  Length:707838      Min.   :0.0000   Min.   :  0.0
##  Class :character   1st Qu.:0.0000   1st Qu.: 15.0
##  Mode  :character   Median :0.0000   Median : 31.0
##                     Mean   :0.6845   Mean   : 34.1
##                     3rd Qu.:1.0000   3rd Qu.: 50.0
##                     Max.   :8.0000   Max.   :188.0
##                                      NA's   :362471
##  mths_since_last_record    open_acc        pub_rec
##  Min.   :  0.0          Min.   : 0.00   Min.   : 0.0000
##  1st Qu.: 51.0          1st Qu.: 8.00   1st Qu.: 0.0000
##  Median : 70.0          Median :11.00   Median : 0.0000
##  Mean   : 70.4          Mean   :11.55   Mean   : 0.1957
##  3rd Qu.: 92.0          3rd Qu.:14.00   3rd Qu.: 0.0000
##  Max.   :129.0          Max.   :90.00   Max.   :86.0000
##  NA's   :598712
##    revol_bal         revol_util       total_acc      initial_list_status
##  Min.   :      0   Min.   :  0.00   Min.   :  2.00   Length:707838
##  1st Qu.:   6455   1st Qu.: 37.70   1st Qu.: 17.00   Class :character
##  Median :  11886   Median : 56.00   Median : 24.00   Mode  :character
##  Mean   :  16886   Mean   : 55.09   Mean   : 25.27
##  3rd Qu.:  20815   3rd Qu.: 73.60   3rd Qu.: 32.00
##  Max.   :2904836   Max.   :892.30   Max.   :169.00
##                    NA's   :366
##    out_prncp       out_prncp_inv      total_pymnt    total_pymnt_inv
##  Min.   :    0.0   Min.   :    0.0   Min.   :    0   Min.   :    0
##  1st Qu.:    0.0   1st Qu.:    0.0   1st Qu.: 7969   1st Qu.: 7928
##  Median :    0.0   Median :    0.0   Median :13466   Median :13436
##  Mean   :  433.5   Mean   :  433.3   Mean   :15887   Mean   :15853
##  3rd Qu.:    0.0   3rd Qu.:    0.0   3rd Qu.:21836   3rd Qu.:21805
##  Max.   :23220.4   Max.   :23220.4   Max.   :63297   Max.   :63297
##
##  total_rec_prncp total_rec_int   total_rec_late_fee   recoveries
##  Min.   :    0   Min.   :    0   Min.   :   0.00    Min.   :    0.0
##  1st Qu.: 6000   1st Qu.: 1028   1st Qu.:   0.00    1st Qu.:    0.0
##  Median :10500   Median : 2012   Median :   0.00    Median :    0.0
##  Mean   :12555   Mean   : 3115   Mean   :   1.52    Mean   :  215.5
```

```
##  3rd Qu.:17650   3rd Qu.: 4027   3rd Qu.:   0.00    3rd Qu.:    0.0
##  Max.   :35000   Max.   :28193   Max.   :1098.36    Max.   :35581.9
##
##  collection_recovery_fee last_pymnt_d       last_pymnt_amnt
##  Min.   :   0.00         Length:707838      Min.   :    0.0
##  1st Qu.:   0.00         Class :character   1st Qu.:  348.9
##  Median :   0.00         Mode  :character   Median :  866.0
##  Mean   :  34.78                            Mean   : 4237.2
##  3rd Qu.:   0.00                            3rd Qu.: 6192.8
##  Max.   :6972.59                            Max.   :36475.6
##
##  next_pymnt_d       last_credit_pull_d collections_12_mths_ex_med
##  Length:707838      Length:707838      Min.   : 0.00000
##  Class :character   Class :character   1st Qu.: 0.00000
##  Mode  :character   Mode  :character   Median : 0.00000
##                                        Mean   : 0.01438
##                                        3rd Qu.: 0.00000
##                                        Max.   :20.00000
##                                        NA's   :44
##  mths_since_last_major_derog  policy_code application_type
##  Min.   :  0.0               Min.   :1    Length:707838
##  1st Qu.: 26.0               1st Qu.:1    Class :character
##  Median : 43.0               Median :1    Mode  :character
##  Mean   : 43.1               Mean   :1
##  3rd Qu.: 60.0               3rd Qu.:1
##  Max.   :188.0               Max.   :1
##  NA's   :530440
##  annual_inc_joint   dti_joint      verification_status_joint
##  Min.   : 18480   Min.   : 3.0     Length:707838
##  1st Qu.: 76800   1st Qu.:13.1     Class :character
##  Median :100000   Median :17.6     Mode  :character
##  Mean   :110494   Mean   :18.2
##  3rd Qu.:131250   3rd Qu.:22.6
##  Max.   :500000   Max.   :39.5
##  NA's   :707430   NA's   :707432
##  acc_now_delinq      tot_coll_amt      tot_cur_bal       open_acc_6m
##  Min.   :0.000000   Min.   :      0   Min.   :      0   Min.   : 0.0
##  1st Qu.:0.000000   1st Qu.:      0   1st Qu.:  29839   1st Qu.: 0.0
##  Median :0.000000   Median :      0   Median :  80587   Median : 1.0
##  Mean   :0.004938   Mean   :    229   Mean   : 139309   Mean   : 1.1
##  3rd Qu.:0.000000   3rd Qu.:      0   3rd Qu.: 208053   3rd Qu.: 2.0
##  Max.   :6.000000   Max.   :9152545   Max.   :8000078   Max.   :14.0
##                     NA's   :54105     NA's   :54105     NA's   :690633
##   open_act_il      open_il_12m      open_il_24m     mths_since_rcnt_il
##  Min.   : 0.0     Min.   : 0.0     Min.   : 0.0     Min.   :  0.0
##  1st Qu.: 1.0     1st Qu.: 0.0     1st Qu.: 0.0     1st Qu.:  6.0
##  Median : 2.0     Median : 0.0     Median : 1.0     Median : 12.0
##  Mean   : 2.9     Mean   : 0.8     Mean   : 1.7     Mean   : 20.9
##  3rd Qu.: 4.0     3rd Qu.: 1.0     3rd Qu.: 2.0     3rd Qu.: 23.0
##  Max.   :40.0     Max.   :12.0     Max.   :19.0     Max.   :363.0
##  NA's   :690633   NA's   :690633   NA's   :690633   NA's   :691087
##   total_bal_il       il_util        open_rv_12m      open_rv_24m
##  Min.   :     0   Min.   :  0.0    Min.   : 0.0     Min.   : 0
##  1st Qu.: 10105   1st Qu.: 59.0    1st Qu.: 0.0     1st Qu.: 1
```

```
##  Median : 24600   Median : 75.0    Median : 1.0     Median : 2
##  Mean   : 36642   Mean   : 71.5    Mean   : 1.4     Mean   : 3
##  3rd Qu.: 47839   3rd Qu.: 88.0    3rd Qu.: 2.0     3rd Qu.: 4
##  Max.   :878459   Max.   :223.0    Max.   :22.0     Max.   :43
##  NA's   :690633   NA's   :692830   NA's   :690633   NA's   :690633
##    max_bal_bc        all_util      total_rev_hi_lim      inq_fi
##  Min.   :    0    Min.   :  0.0    Min.   :      0   Min.   : 0.0
##  1st Qu.: 2405    1st Qu.: 48.0    1st Qu.:  14000   1st Qu.: 0.0
##  Median : 4471    Median : 62.0    Median :  23700   Median : 0.0
##  Mean   : 5873    Mean   : 60.8    Mean   :  32054   Mean   : 0.9
##  3rd Qu.: 7771    3rd Qu.: 75.0    3rd Qu.:  39700   3rd Qu.: 1.0
##  Max.   :83047    Max.   :151.0    Max.   :9999999   Max.   :17.0
##  NA's   :690633   NA's   :690633   NA's   :54105     NA's   :690633
##   total_cu_tl      inq_last_12m    acc_open_past_24mths  avg_cur_bal
##  Min.   : 0.0     Min.   : 0.0     Min.   : 0.00        Min.   :     0
##  1st Qu.: 0.0     1st Qu.: 0.0     1st Qu.: 2.00        1st Qu.:  3139
##  Median : 0.0     Median : 2.0     Median : 4.00        Median :  7429
##  Mean   : 1.5     Mean   : 2.2     Mean   : 4.43        Mean   : 13325
##  3rd Qu.: 2.0     3rd Qu.: 3.0     3rd Qu.: 6.00        3rd Qu.: 18531
##  Max.   :32.0     Max.   :32.0     Max.   :64.00        Max.   :958084
##  NA's   :690633   NA's   :690633   NA's   :37879        NA's   :54116
##  bc_open_to_buy      bc_util       chargeoff_within_12_mths
##  Min.   :     0   Min.   :  0.00   Min.   : 0.00000
##  1st Qu.:  1201   1st Qu.: 44.40   1st Qu.: 0.00000
##  Median :  3950   Median : 68.30   Median : 0.00000
##  Mean   :  8995   Mean   : 63.93   Mean   : 0.00893
##  3rd Qu.: 10667   3rd Qu.: 87.40   3rd Qu.: 0.00000
##  Max.   :559912   Max.   :339.60   Max.   :10.00000
##  NA's   :44220    NA's   :44639    NA's   :44
##   delinq_amnt       mo_sin_old_il_acct mo_sin_old_rev_tl_op
##  Min.   :    0.00   Min.   :  0.0      Min.   :  3.0
##  1st Qu.:    0.00   1st Qu.:100.0      1st Qu.:119.0
##  Median :    0.00   Median :130.0      Median :167.0
##  Mean   :   10.95   Mean   :127.3      Mean   :184.7
##  3rd Qu.:    0.00   3rd Qu.:153.0      3rd Qu.:233.0
##  Max.   :94521.00   Max.   :720.0      Max.   :851.0
##                     NA's   :74555      NA's   :54105
##  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 :  8.00        Median :  6.00   Median : 1.00
##  Mean   : 13.47        Mean   :  8.19   Mean   : 1.76
##  3rd Qu.: 16.00        3rd Qu.: 10.00   3rd Qu.: 3.00
##  Max.   :330.00        Max.   :263.00   Max.   :52.00
##  NA's   :54105         NA's   :54105    NA's   :37879
##  mths_since_recent_bc mths_since_recent_bc_dlq mths_since_recent_inq
##  Min.   :  0.00       Min.   :  0              Min.   : 0.00
##  1st Qu.:  6.00       1st Qu.: 21              1st Qu.: 2.00
##  Median : 14.00       Median : 39              Median : 5.00
##  Mean   : 24.95       Mean   : 40              Mean   : 6.85
##  3rd Qu.: 30.00       3rd Qu.: 59              3rd Qu.:10.00
##  Max.   :615.00       Max.   :195              Max.   :25.00
##  NA's   :43760        NA's   :541664           NA's   :107168
##  mths_since_recent_revol_delinq num_accts_ever_120_pd num_actv_bc_tl
```

```
##  Min.   :  0.0                 Min.   : 0.00         Min.   : 0.00
##  1st Qu.: 17.0                 1st Qu.: 0.00         1st Qu.: 2.00
##  Median : 33.0                 Median : 0.00         Median : 3.00
##  Mean   : 35.9                 Mean   : 0.48         Mean   : 3.73
##  3rd Qu.: 53.0                 3rd Qu.: 0.00         3rd Qu.: 5.00
##  Max.   :176.0                 Max.   :35.00         Max.   :32.00
##  NA's   :474946                NA's   :54105         NA's   :54105
##  num_actv_rev_tl  num_bc_sats      num_bc_tl       num_il_tl
##  Min.   : 0.0    Min.   : 0.00   Min.   : 0.00   Min.   :  0.00
##  1st Qu.: 4.0    1st Qu.: 3.00   1st Qu.: 5.00   1st Qu.:  3.00
##  Median : 5.0    Median : 4.00   Median : 8.00   Median :  7.00
##  Mean   : 5.8    Mean   : 4.73   Mean   : 8.42   Mean   :  8.44
##  3rd Qu.: 7.0    3rd Qu.: 6.00   3rd Qu.:11.00   3rd Qu.: 11.00
##  Max.   :47.0    Max.   :63.00   Max.   :70.00   Max.   :150.00
##  NA's   :54105   NA's   :44737   NA's   :54105   NA's   :54105
##  num_op_rev_tl   num_rev_accts    num_rev_tl_bal_gt_0    num_sats
##  Min.   : 0.00   Min.   :  1.00   Min.   : 0.00       Min.   : 0.00
##  1st Qu.: 5.00   1st Qu.:  9.00   1st Qu.: 4.00       1st Qu.: 8.00
##  Median : 7.00   Median : 14.00   Median : 5.00       Median :11.00
##  Mean   : 8.32   Mean   : 15.01   Mean   : 5.77       Mean   :11.64
##  3rd Qu.:10.00   3rd Qu.: 19.00   3rd Qu.: 7.00       3rd Qu.:14.00
##  Max.   :83.00   Max.   :118.00   Max.   :43.00       Max.   :90.00
##  NA's   :54105   NA's   :54106    NA's   :54105       NA's   :44737
##  num_tl_120dpd_2m  num_tl_30dpd   num_tl_90g_dpd_24m num_tl_op_past_12m
##  Min.   :0        Min.   :0       Min.   : 0.00      Min.   : 0.00
##  1st Qu.:0        1st Qu.:0       1st Qu.: 0.00      1st Qu.: 1.00
##  Median :0        Median :0       Median : 0.00      Median : 2.00
##  Mean   :0        Mean   :0       Mean   : 0.09      Mean   : 2.05
##  3rd Qu.:0        3rd Qu.:0       3rd Qu.: 0.00      3rd Qu.: 3.00
##  Max.   :6        Max.   :4       Max.   :30.00      Max.   :30.00
##  NA's   :75894    NA's   :54105   NA's   :54105      NA's   :54105
##  pct_tl_nvr_dlq   percent_bc_gt_75 pub_rec_bankruptcies   tax_liens
##  Min.   :  7.70   Min.   :  0.00   Min.   : 0.0000      Min.   : 0.00000
##  1st Qu.: 91.70   1st Qu.: 20.00   1st Qu.: 0.0000      1st Qu.: 0.00000
##  Median : 98.00   Median : 50.00   Median : 0.0000      Median : 0.00000
##  Mean   : 94.31   Mean   : 49.79   Mean   : 0.1196      Mean   : 0.04862
##  3rd Qu.:100.00   3rd Qu.: 80.00   3rd Qu.: 0.0000      3rd Qu.: 0.00000
##  Max.   :100.00   Max.   :100.00   Max.   :12.0000      Max.   :85.00000
##  NA's   :54226    NA's   :44538    NA's   :576          NA's   :30
##  tot_hi_cred_lim   total_bal_ex_mort total_bc_limit
##  Min.   :      0   Min.   :      0   Min.   :      0
##  1st Qu.:  48800   1st Qu.:  21165   1st Qu.:   7600
##  Median : 110947   Median :  37226   Median :  14700
##  Mean   : 170763   Mean   :  49108   Mean   :  21023
##  3rd Qu.: 247660   3rd Qu.:  61681   3rd Qu.:  27400
##  Max.   :9999999   Max.   :2921551   Max.   :1090700
##  NA's   :54105     NA's   :37879     NA's   :37879
##  total_il_high_credit_limit revol_bal_joint  sec_app_earliest_cr_line
##  Min.   :      0            Min.   : NA      Length:707838
##  1st Qu.:  13965            1st Qu.: NA      Class :character
##  Median :  30554            Median : NA      Mode  :character
##  Mean   :  40635            Mean   :NaN
##  3rd Qu.:  54727            3rd Qu.: NA
##  Max.   :2101913            Max.   : NA
```

```
##  NA's   :54105              NA's   :707838
##  sec_app_inq_last_6mths sec_app_mort_acc sec_app_open_acc
##  Min.   : NA            Min.   : NA      Min.   : NA
##  1st Qu.: NA            1st Qu.: NA      1st Qu.: NA
##  Median : NA            Median : NA      Median : NA
##  Mean   :NaN            Mean   :NaN      Mean   :NaN
##  3rd Qu.: NA            3rd Qu.: NA      3rd Qu.: NA
##  Max.   : NA            Max.   : NA      Max.   : NA
##  NA's   :707838         NA's   :707838   NA's   :707838
##  sec_app_revol_util sec_app_open_act_il sec_app_num_rev_accts
##  Min.   : NA        Min.   : NA         Min.   : NA
##  1st Qu.: NA        1st Qu.: NA         1st Qu.: NA
##  Median : NA        Median : NA         Median : NA
##  Mean   :NaN        Mean   :NaN         Mean   :NaN
##  3rd Qu.: NA        3rd Qu.: NA         3rd Qu.: NA
##  Max.   : NA        Max.   : NA         Max.   : NA
##  NA's   :707838     NA's   :707838      NA's   :707838
##  sec_app_chargeoff_within_12_mths sec_app_collections_12_mths_ex_med
##  Min.   : NA                      Min.   : NA
##  1st Qu.: NA                      1st Qu.: NA
##  Median : NA                      Median : NA
##  Mean   :NaN                      Mean   :NaN
##  3rd Qu.: NA                      3rd Qu.: NA
##  Max.   : NA                      Max.   : NA
##  NA's   :707838                   NA's   :707838
##  sec_app_mths_since_last_major_derog hardship_flag      hardship_type
##  Min.   : NA                         Length:707838      Length:707838
##  1st Qu.: NA                         Class :character   Class :character
##  Median : NA                         Mode  :character   Mode  :character
##  Mean   :NaN
##  3rd Qu.: NA
##  Max.   : NA
##  NA's   :707838
##  hardship_reason    hardship_status    deferral_term    hardship_amount
##  Length:707838      Length:707838      Min.   :3        Min.   :  0.6
##  Class :character   Class :character   1st Qu.:3        1st Qu.: 44.6
##  Mode  :character   Mode  :character   Median :3        Median : 95.0
##                                        Mean   :3        Mean   :117.4
##                                        3rd Qu.:3        3rd Qu.:163.9
##                                        Max.   :3        Max.   :629.7
##                                        NA's   :705370   NA's   :705370
##  hardship_start_date hardship_end_date  payment_plan_start_date
##  Length:707838       Length:707838      Length:707838
##  Class :character    Class :character   Class :character
##  Mode  :character    Mode  :character   Mode  :character
##
##
##
##
##  hardship_length   hardship_dpd    hardship_loan_status
##  Min.   :3        Min.   : 0.0     Length:707838
##  1st Qu.:3        1st Qu.: 7.0     Class :character
##  Median :3        Median :15.0     Mode  :character
##  Mean   :3        Mean   :14.1
```

```
##  3rd Qu.:3        3rd Qu.:23.0
##  Max.   :3        Max.   :32.0
##  NA's   :705370   NA's   :705370
##  orig_projected_additional_accrued_interest hardship_payoff_balance_amount
##  Min.   :   1.9                             Min.   :   55.7
##  1st Qu.: 128.9                             1st Qu.: 4246.7
##  Median : 281.9                             Median : 7956.3
##  Mean   : 345.8                             Mean   : 9020.8
##  3rd Qu.: 484.5                             3rd Qu.:12801.6
##  Max.   :1889.1                             Max.   :29401.0
##  NA's   :705844                             NA's   :705370
##  hardship_last_payment_amount disbursement_method debt_settlement_flag
##  Min.   :  0.0                Length:707838       Length:707838
##  1st Qu.: 44.4                Class :character    Class :character
##  Median :136.2                Mode  :character    Mode  :character
##  Mean   :182.0
##  3rd Qu.:270.4
##  Max.   :991.8
##  NA's   :705370
##  debt_settlement_flag_date settlement_status  settlement_date
##  Length:707838             Length:707838      Length:707838
##  Class :character          Class :character   Class :character
##  Mode  :character          Mode  :character   Mode  :character
##
##
##
##
##  settlement_amount settlement_percentage settlement_term
##  Min.   :   44.2   Min.   :  0.2         Min.   :  0
##  1st Qu.: 2065.8   1st Qu.: 45.0         1st Qu.:  1
##  Median : 3957.8   Median : 45.0         Median : 12
##  Mean   : 4674.1   Mean   : 47.0         Mean   : 11
##  3rd Qu.: 6408.2   3rd Qu.: 50.0         3rd Qu.: 18
##  Max.   :33601.0   Max.   :521.4         Max.   :112
##  NA's   :693575    NA's   :693575        NA's   :693575
```

###1.1. Identifikasi Missing Values

Sebelum memulai analisis, terlebih dahulu kita identifikasi dahulu data yang tidak lengkap, agar tidak merusak performa analisis maupun model machine learning yang akan kita gunakan untuk Credit Scoring Model. Identifikasi missing values:

```
colSums(is.na(loan))
```

```
##                                member_id
##                                   707838
##                                loan_amnt
##                                        0
##                              funded_amnt
##                                        0
##                            funded_amnt_inv
##                                        0
##                                     term
##                                        0
##                                 int_rate
##                                        0
```

```
##                          installment
##                                    0
##                                grade
##                                    0
##                            sub_grade
##                                    0
##                            emp_title
##                                40974
##                           emp_length
##                                    0
##                          home_ownership
##                                    0
##                           annual_inc
##                                    0
##                        verification_status
##                                    0
##                              issue_d
##                                    0
##                          loan_status
##                                    0
##                           pymnt_plan
##                                    0
##                                  url
##                               707838
##                                 desc
##                               609114
##                              purpose
##                                    0
##                                title
##                                  134
##                             zip_code
##                                    0
##                           addr_state
##                                    0
##                                  dti
##                                    1
##                          delinq_2yrs
##                                    0
##                        earliest_cr_line
##                                    0
##                               inq_last_6mths
##                                    0
##                     mths_since_last_delinq
##                                362471
##                     mths_since_last_record
##                                598712
##                                open_acc
##                                    0
##                                 pub_rec
##                                    0
##                               revol_bal
##                                    0
##                              revol_util
##                                   366
```

```
##                              total_acc
##                                      0
##                      initial_list_status
##                                      0
##                                out_prncp
##                                      0
##                            out_prncp_inv
##                                      0
##                              total_pymnt
##                                      0
##                          total_pymnt_inv
##                                      0
##                          total_rec_prncp
##                                      0
##                            total_rec_int
##                                      0
##                       total_rec_late_fee
##                                      0
##                               recoveries
##                                      0
##                   collection_recovery_fee
##                                      0
##                             last_pymnt_d
##                                    522
##                          last_pymnt_amnt
##                                      0
##                             next_pymnt_d
##                                   658374
##                         last_credit_pull_d
##                                     35
##                  collections_12_mths_ex_med
##                                     44
##                mths_since_last_major_derog
##                                   530440
##                              policy_code
##                                      0
##                         application_type
##                                      0
##                         annual_inc_joint
##                                   707430
##                                dti_joint
##                                   707432
##                  verification_status_joint
##                                   707430
##                           acc_now_delinq
##                                      0
##                             tot_coll_amt
##                                    54105
##                              tot_cur_bal
##                                    54105
##                             open_acc_6m
##                                   690633
##                             open_act_il
##                                   690633
```

```
##                           open_il_12m
##                                690633
##                           open_il_24m
##                                690633
##                      mths_since_rcnt_il
##                                691087
##                            total_bal_il
##                                690633
##                               il_util
##                                692830
##                           open_rv_12m
##                                690633
##                           open_rv_24m
##                                690633
##                            max_bal_bc
##                                690633
##                              all_util
##                                690633
##                         total_rev_hi_lim
##                                 54105
##                                inq_fi
##                                690633
##                           total_cu_tl
##                                690633
##                          inq_last_12m
##                                690633
##                      acc_open_past_24mths
##                                 37879
##                           avg_cur_bal
##                                 54116
##                         bc_open_to_buy
##                                 44220
##                                bc_util
##                                 44639
##                    chargeoff_within_12_mths
##                                    44
##                             delinq_amnt
##                                     0
##                          mo_sin_old_il_acct
##                                 74555
##                      mo_sin_old_rev_tl_op
##                                 54105
##                     mo_sin_rcnt_rev_tl_op
##                                 54105
##                           mo_sin_rcnt_tl
##                                 54105
##                               mort_acc
##                                 37879
##                       mths_since_recent_bc
##                                 43760
##                    mths_since_recent_bc_dlq
##                                541664
##                     mths_since_recent_inq
##                                107168
```

```
##             mths_since_recent_revol_delinq
##                                     474946
##                       num_accts_ever_120_pd
##                                      54105
##                              num_actv_bc_tl
##                                      54105
##                             num_actv_rev_tl
##                                      54105
##                                num_bc_sats
##                                      44737
##                                 num_bc_tl
##                                      54105
##                                 num_il_tl
##                                      54105
##                               num_op_rev_tl
##                                      54105
##                               num_rev_accts
##                                      54106
##                          num_rev_tl_bal_gt_0
##                                      54105
##                                   num_sats
##                                      44737
##                             num_tl_120dpd_2m
##                                      75894
##                                num_tl_30dpd
##                                      54105
##                           num_tl_90g_dpd_24m
##                                      54105
##                           num_tl_op_past_12m
##                                      54105
##                               pct_tl_nvr_dlq
##                                      54226
##                             percent_bc_gt_75
##                                      44538
##                          pub_rec_bankruptcies
##                                        576
##                                   tax_liens
##                                         30
##                               tot_hi_cred_lim
##                                      54105
##                          total_bal_ex_mort
##                                      37879
##                               total_bc_limit
##                                      37879
##                    total_il_high_credit_limit
##                                      54105
##                                revol_bal_joint
##                                     707838
##                      sec_app_earliest_cr_line
##                                     707838
##                         sec_app_inq_last_6mths
##                                     707838
##                               sec_app_mort_acc
##                                     707838
```

```
##                           sec_app_open_acc
##                                     707838
##                          sec_app_revol_util
##                                     707838
##                        sec_app_open_act_il
##                                     707838
##                       sec_app_num_rev_accts
##                                     707838
##             sec_app_chargeoff_within_12_mths
##                                     707838
##           sec_app_collections_12_mths_ex_med
##                                     707838
##          sec_app_mths_since_last_major_derog
##                                     707838
##                                hardship_flag
##                                           0
##                                hardship_type
##                                     705370
##                              hardship_reason
##                                     705370
##                              hardship_status
##                                     705370
##                                deferral_term
##                                     705370
##                              hardship_amount
##                                     705370
##                           hardship_start_date
##                                     705370
##                             hardship_end_date
##                                     705370
##                       payment_plan_start_date
##                                     705370
##                                hardship_length
##                                     705370
##                                  hardship_dpd
##                                     705370
##                          hardship_loan_status
##                                     705370
## orig_projected_additional_accrued_interest
##                                     705844
##                hardship_payoff_balance_amount
##                                     705370
##                 hardship_last_payment_amount
##                                     705370
##                           disbursement_method
##                                           0
##                          debt_settlement_flag
##                                           0
##                    debt_settlement_flag_date
##                                     693575
##                            settlement_status
##                                     693575
##                              settlement_date
##                                     693575
```

```
##                          settlement_amount
##                                     693575
##                       settlement_percentage
##                                     693575
##                                settlement_term
##                                     693575
```

Kita lihat bahwa dari total 707838 observasi dari dataset Lending Club Loan, terdapat 13 variabel yang seluruh datanya kosong, terdapat variabel-variabel yang memiliki NA lebih dari 80% keseluruhan observasi, serta terdapat 2 variabel waktu yang memiliki missing value. Dikarenakan variabel waktu terpenting adalah issue_d dan earliest_cr_line, sehingga untuk mengurangi kompleksitas interpolasi data tipe "date", maka kedua variabel waktu tersebut (last_pymnt_d dan last_credit_pull_d) juga dihilangkan, karena kita juga tidak bisa mengambil risiko untuk menghapus baris observasi, karena mengandung "loan_status" yang akan diprediksi nanti.

Berikut ini proses penghapusan variabel-variabel tersebut, dan simpan menjadi variabel data "loan2":

```
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
```

```
# Drop the columns of the dataframe
loan2 <- select (loan,-c(member_id,
                         desc,
                         url,
                         last_pymnt_d,
                         last_credit_pull_d,
                         mths_since_last_record,
                         revol_bal_joint,
                         sec_app_earliest_cr_line,
                         sec_app_inq_last_6mths,
                         sec_app_mort_acc,
                         sec_app_open_acc,
                         sec_app_revol_util,
                         sec_app_open_act_il,
                         sec_app_num_rev_accts,
                         sec_app_chargeoff_within_12_mths,
                         sec_app_collections_12_mths_ex_med,
                         sec_app_mths_since_last_major_derog,
                         hardship_type,
                         hardship_reason,
                         hardship_status,
                         deferral_term,
                         hardship_amount,
                         hardship_start_date,
                         hardship_end_date,
                         payment_plan_start_date,
                         hardship_length,
```

```
                        hardship_dpd,
                        hardship_loan_status,
                        orig_projected_additional_accrued_interest,
                        hardship_payoff_balance_amount,
                        hardship_last_payment_amount,
                        debt_settlement_flag_date,
                        settlement_status,
                        settlement_date,
                        settlement_amount,
                        settlement_percentage,
                        settlement_term,
                        inq_last_12m,
                        inq_fi,
                        total_cu_tl,
                        all_util,
                        open_acc_6m,
                        open_act_il,
                        open_il_12m,
                        open_il_24m,
                        mths_since_rcnt_il,
                        total_bal_il,
                        il_util,
                        open_rv_12m,
                        open_rv_24m,
                        max_bal_bc,
                        verification_status_joint,
                        annual_inc_joint,
                        dti_joint,
                        next_pymnt_d))
```

Cek kembali struktur data:

```
str(loan2)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 707838 obs. of  89 variables:
##  $ loan_amnt                   : num  24700 35000 20000 20000 14025 ...
##  $ funded_amnt                 : num  24700 35000 20000 20000 14025 ...
##  $ funded_amnt_inv             : num  24700 35000 20000 20000 14025 ...
##  $ term                        : chr  "36 months" "60 months" "60 months" "36 months" ...
##  $ int_rate                    : num  12 12.9 14 12.9 18.5 ...
##  $ installment                 : num  820 794 465 673 360 ...
##  $ grade                       : chr  "C" "C" "C" "C" ...
##  $ sub_grade                   : chr  "C1" "C2" "C4" "C2" ...
##  $ emp_title                   : chr  "Engineer" "Lieutenant" "Facilities Coordinator" "President"
##  $ emp_length                  : chr  "10+ years" "7 years" "10+ years" "10+ years" ...
##  $ home_ownership              : chr  "MORTGAGE" "MORTGAGE" "OWN" "MORTGAGE" ...
##  $ annual_inc                  : num  65000 106000 70000 145000 39000 ...
##  $ verification_status         : chr  "Not Verified" "Source Verified" "Source Verified" "Not Veri
##  $ issue_d                     : chr  "Dec-2015" "Dec-2015" "Dec-2015" "Dec-2015" ...
##  $ loan_status                 : chr  "Fully Paid" "Charged Off" "Charged Off" "Late (31-120 days)
##  $ pymnt_plan                  : chr  "n" "n" "n" "n" ...
##  $ purpose                     : chr  "small_business" "debt_consolidation" "debt_consolidation" "
##  $ title                       : chr  "Business" "Debt consolidation" "Debt consolidation" "Debt c
##  $ zip_code                    : chr  "577xx" "351xx" "210xx" "029xx" ...
##  $ addr_state                  : chr  "SD" "AL" "MD" "RI" ...
```

```
##  $ dti                            : num  16.1 17.4 16.9 12.3 18 ...
##  $ delinq_2yrs                    : num  1 0 0 0 0 0 0 0 0 0 ...
##  $ earliest_cr_line               : chr  "Dec-1999" "Apr-2002" "Jun-2001" "Feb-2004" ...
##  $ inq_last_6mths                 : num  4 0 0 0 1 0 0 0 1 0 ...
##  $ mths_since_last_delinq         : num  6 NA 33 NA NA 39 27 NA 30 NA ...
##  $ open_acc                       : num  22 8 20 12 12 10 7 6 7 13 ...
##  $ pub_rec                        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ revol_bal                      : num  21470 39055 31200 22551 15646 ...
##  $ revol_util                     : num  19.2 72.1 42 80.2 74.9 51.2 52.3 56.2 29.7 11.6 ...
##  $ total_acc                      : num  38 27 35 21 21 24 13 18 13 17 ...
##  $ initial_list_status            : chr  "w" "w" "w" "w" ...
##  $ out_prncp                      : num  0 0 0 673 0 ...
##  $ out_prncp_inv                  : num  0 0 0 673 0 ...
##  $ total_pymnt                    : num  25680 22208 15473 23510 16341 ...
##  $ total_pymnt_inv                : num  25680 22208 15473 23510 16341 ...
##  $ total_rec_prncp                : num  24700 10324 2965 19327 14025 ...
##  $ total_rec_int                  : num  980 7086 2593 4183 2316 ...
##  $ total_rec_late_fee             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ recoveries                     : num  0 4798 9915 0 0 ...
##  $ collection_recovery_fee        : num  0 864 1785 0 0 ...
##  $ last_pymnt_amnt                : num  926 794 465 673 12764 ...
##  $ collections_12_mths_ex_med     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ mths_since_last_major_derog    : num  NA NA 69 NA NA 61 NA NA 30 NA ...
##  $ policy_code                    : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ application_type               : chr  "Individual" "Individual" "Individual" "Individual" ...
##  $ acc_now_delinq                 : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ tot_coll_amt                   : num  0 0 0 0 0 0 0 0 722 0 ...
##  $ tot_cur_bal                    : num  204396 299890 34856 429218 60381 ...
##  $ total_rev_hi_lim               : num  111800 54200 74200 28500 20900 ...
##  $ acc_open_past_24mths           : num  4 0 3 5 5 6 3 6 4 2 ...
##  $ avg_cur_bal                    : num  9733 37486 1835 39020 5032 ...
##  $ bc_open_to_buy                 : num  57830 12875 18527 2822 578 ...
##  $ bc_util                        : num  27.1 67.2 44 81.6 93.4 63 52.3 55.9 37.2 12.1 ...
##  $ chargeoff_within_12_mths       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ delinq_amnt                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ mo_sin_old_il_acct             : num  113 135 46 79 138 165 106 125 148 36 ...
##  $ mo_sin_old_rev_tl_op           : num  192 164 174 142 172 159 163 184 128 87 ...
##  $ mo_sin_rcnt_rev_tl_op          : num  2 38 5 4 2 3 11 14 3 2 ...
##  $ mo_sin_rcnt_tl                 : num  2 35 5 2 2 3 11 14 3 2 ...
##  $ mort_acc                       : num  4 4 0 5 3 3 0 5 1 1 ...
##  $ mths_since_recent_bc           : num  2 38 6 27 2 3 11 101 4 2 ...
##  $ mths_since_recent_bc_dlq       : num  NA NA NA NA NA NA 30 NA 69 NA ...
##  $ mths_since_recent_inq          : num  0 NA 13 1 2 17 11 10 4 NA ...
##  $ mths_since_recent_revol_delinq: num  6 NA NA NA NA 39 27 NA 69 NA ...
##  $ num_accts_ever_120_pd          : num  0 0 1 0 0 1 0 0 2 0 ...
##  $ num_actv_bc_tl                 : num  5 3 6 4 3 4 2 2 2 4 ...
##  $ num_actv_rev_tl                : num  5 4 11 6 11 5 2 3 4 5 ...
##  $ num_bc_sats                    : num  13 3 9 5 3 4 4 2 2 8 ...
##  $ num_bc_tl                      : num  17 6 19 8 3 7 5 4 5 10 ...
##  $ num_il_tl                      : num  6 13 2 5 4 9 6 6 3 2 ...
##  $ num_op_rev_tl                  : num  20 4 19 7 11 6 4 4 4 10 ...
##  $ num_rev_accts                  : num  27 10 32 11 13 12 7 7 9 13 ...
##  $ num_rev_tl_bal_gt_0            : num  5 4 11 6 11 6 2 3 4 5 ...
##  $ num_sats                       : num  22 8 20 12 12 10 7 6 7 13 ...
```

```
##  $ num_tl_120dpd_2m              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ num_tl_30dpd                  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ num_tl_90g_dpd_24m            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ num_tl_op_past_12m            : num  2 0 2 3 2 3 1 0 3 1 ...
##  $ pct_tl_nvr_dlq                : num  97.4 100 94.3 100 100 87.5 84.6 100 76.9 100 ...
##  $ percent_bc_gt_75              : num  7.7 0 44.4 75 66.7 50 0 50 0 0 ...
##  $ pub_rec_bankruptcies          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ tax_liens                     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ tot_hi_cred_lim               : num  314017 366974 84200 472759 72286 ...
##  $ total_bal_ex_mort             : num  39475 55067 34856 32486 15646 ...
##  $ total_bc_limit                : num  79300 39200 33100 13500 8800 23500 21000 6200 2400 62500 ...
##  $ total_il_high_credit_limit    : num  24667 36981 10000 23259 0 ...
##  $ hardship_flag                 : chr  "N" "N" "N" "N" ...
##  $ disbursement_method           : chr  "Cash" "Cash" "Cash" "Cash" ...
##  $ debt_settlement_flag          : chr  "N" "N" "Y" "N" ...
```

Selanjutnya, untuk data dengan tipe "character" dan masih memiliki missing values atau NA, kita ganti NA tersebut dengan character "None":

```
library(forcats)
```

```
## Warning: package 'forcats' was built under R version 3.6.1
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.1
```

```
## -- Attaching packages ---------------------------------------------------------------------------- t
```

```
## v ggplot2 3.2.0     v readr   1.3.1
## v tibble  2.1.3     v purrr   0.3.2
## v tidyr   0.8.3     v stringr 1.4.0
```

```
## Warning: package 'readr' was built under R version 3.6.1
```

```
## -- Conflicts ------------------------------------------------------------------------------- tidyver
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
loan2 %<>% mutate(title = fct_explicit_na(title, na_level = "None"))
loan2 %<>% mutate(emp_title = fct_explicit_na(emp_title, na_level = "None"))
```

Mengatasi missing values pada data numerik dengan mengganti NA nilai rata-rata atau mean:

```
loan2 = transform(loan2, dti = ifelse(is.na(dti), mean(dti, na.rm=TRUE), dti))
loan2 = transform(loan2, revol_util = ifelse(is.na(revol_util), mean(revol_util, na.rm=TRUE), revol_uti
loan2 = transform(loan2, collections_12_mths_ex_med = ifelse(is.na(collections_12_mths_ex_med), mean(co
loan2 = transform(loan2, total_il_high_credit_limit = ifelse(is.na(total_il_high_credit_limit), mean(to
```

```
loan2 = transform(loan2, total_bc_limit = ifelse(is.na(total_bc_limit), mean(total_bc_limit, na.rm=TRUE)
loan2 = transform(loan2, total_bal_ex_mort = ifelse(is.na(total_bal_ex_mort), mean(total_bal_ex_mort, na
loan2 = transform(loan2, tot_hi_cred_lim = ifelse(is.na(tot_hi_cred_lim), mean(tot_hi_cred_lim, na.rm=TR
loan2 = transform(loan2, tax_liens = ifelse(is.na(tax_liens), mean(tax_liens, na.rm=TRUE), tax_liens))
loan2 = transform(loan2, pub_rec_bankruptcies = ifelse(is.na(pub_rec_bankruptcies), mean(pub_rec_bankrup
loan2 = transform(loan2, percent_bc_gt_75 = ifelse(is.na(percent_bc_gt_75), mean(percent_bc_gt_75, na.rm
loan2 = transform(loan2, pct_tl_nvr_dlq = ifelse(is.na(pct_tl_nvr_dlq), mean(pct_tl_nvr_dlq, na.rm=TRUE)
loan2 = transform(loan2, tot_coll_amt = ifelse(is.na(tot_coll_amt), mean(tot_coll_amt, na.rm=TRUE), tot_
loan2 = transform(loan2, tot_cur_bal = ifelse(is.na(tot_cur_bal), mean(tot_cur_bal, na.rm=TRUE), tot_cur
loan2 = transform(loan2, total_rev_hi_lim = ifelse(is.na(total_rev_hi_lim), mean(total_rev_hi_lim, na.rm
loan2 = transform(loan2, mths_since_recent_revol_delinq = ifelse(is.na(mths_since_recent_revol_delinq),
loan2 = transform(loan2, avg_cur_bal = ifelse(is.na(avg_cur_bal), mean(avg_cur_bal, na.rm=TRUE), avg_cur
loan2 = transform(loan2, bc_open_to_buy = ifelse(is.na(bc_open_to_buy), mean(bc_open_to_buy, na.rm=TRUE)
loan2 = transform(loan2, bc_util = ifelse(is.na(bc_util), mean(bc_util, na.rm=TRUE), bc_util))
loan2 = transform(loan2, num_bc_sats = ifelse(is.na(num_bc_sats), mean(num_bc_sats, na.rm=TRUE), num_bc_
loan2 = transform(loan2, mort_acc = ifelse(is.na(mort_acc), mean(mort_acc, na.rm=TRUE), mort_acc))
```

Mengatasi missing values pada data numerik dengan mengganti "NA" dengan "0" (khusus data numerik yang memiliki banyak (>= 50%) data kosong, dan berkaitan dengan tenggat waktu atau status pinjaman tertentu):

```
loan2 = transform(loan2, mths_since_recent_bc_dlq = ifelse(is.na(mths_since_recent_bc_dlq), 0, mths_sinc
loan2 = transform(loan2, mths_since_last_delinq = ifelse(is.na(mths_since_last_delinq), 0, mths_since_la
loan2 = transform(loan2, mo_sin_old_il_acct = ifelse(is.na(mo_sin_old_il_acct), 0, mo_sin_old_il_acct))
loan2 = transform(loan2, mo_sin_old_rev_tl_op = ifelse(is.na(mo_sin_old_rev_tl_op), 0, mo_sin_old_rev_tl
loan2 = transform(loan2, mo_sin_rcnt_rev_tl_op = ifelse(is.na(mo_sin_rcnt_rev_tl_op), 0, mo_sin_rcnt_rev
loan2 = transform(loan2, mo_sin_rcnt_tl = ifelse(is.na(mo_sin_rcnt_tl), 0, mo_sin_rcnt_tl))
loan2 = transform(loan2, mths_since_recent_bc = ifelse(is.na(mths_since_recent_bc), 0, mths_since_recent
loan2 = transform(loan2, num_tl_op_past_12m = ifelse(is.na(num_tl_op_past_12m), 0, num_tl_op_past_12m))
loan2 = transform(loan2, num_tl_90g_dpd_24m = ifelse(is.na(num_tl_90g_dpd_24m), 0, num_tl_90g_dpd_24m))
loan2 = transform(loan2, chargeoff_within_12_mths = ifelse(is.na(chargeoff_within_12_mths), 0, chargeoff
loan2 = transform(loan2, mths_since_last_major_derog = ifelse(is.na(mths_since_last_major_derog), 0, mth
loan2 = transform(loan2, num_tl_30dpd = ifelse(is.na(num_tl_30dpd), 0, num_tl_30dpd))
loan2 = transform(loan2, num_tl_120dpd_2m = ifelse(is.na(num_tl_120dpd_2m), 0, num_tl_120dpd_2m))
loan2 = transform(loan2, num_accts_ever_120_pd = ifelse(is.na(num_accts_ever_120_pd), 0, num_accts_ever_
loan2 = transform(loan2, mths_since_recent_inq = ifelse(is.na(mths_since_recent_inq), 0, mths_since_rece
loan2 = transform(loan2, mths_since_recent_revol_delinq = ifelse(is.na(mths_since_recent_revol_delinq),
loan2 = transform(loan2, acc_open_past_24mths = ifelse(is.na(acc_open_past_24mths), 0, acc_open_past_24m
loan2 = transform(loan2, num_actv_bc_tl = ifelse(is.na(num_actv_bc_tl), 0, num_actv_bc_tl))
loan2 = transform(loan2, num_actv_rev_tl = ifelse(is.na(num_actv_rev_tl), 0, num_actv_rev_tl))
loan2 = transform(loan2, num_bc_tl = ifelse(is.na(num_bc_tl), 0, num_bc_tl))
loan2 = transform(loan2, num_il_tl = ifelse(is.na(num_il_tl), 0, num_il_tl))
loan2 = transform(loan2, num_op_rev_tl = ifelse(is.na(num_op_rev_tl), 0, num_op_rev_tl))
loan2 = transform(loan2, num_rev_accts = ifelse(is.na(num_rev_accts), 0, num_rev_accts))
loan2 = transform(loan2, num_rev_tl_bal_gt_0 = ifelse(is.na(num_rev_tl_bal_gt_0), 0, num_rev_tl_bal_gt_
```

1.2. Export Data yang Telah Bersih ke File Baru

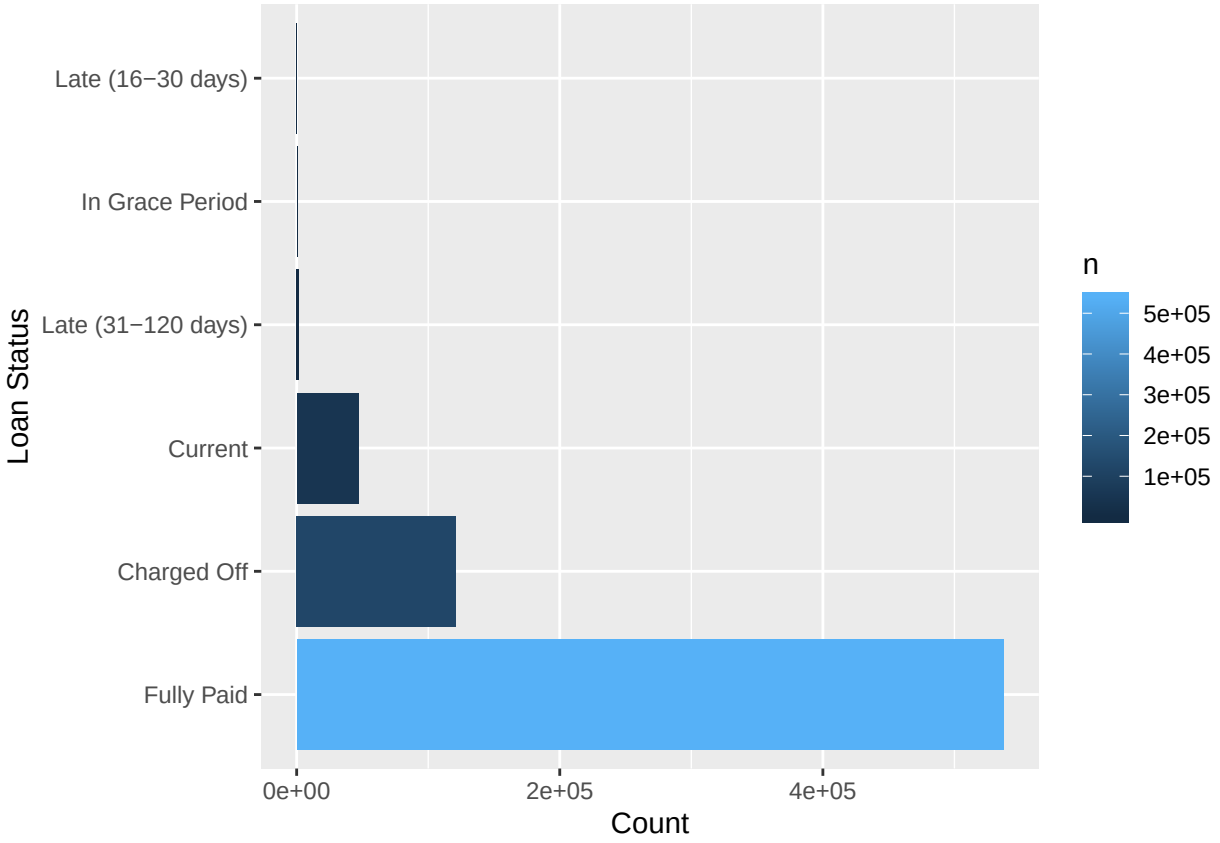
Simpan data yang telah dibersihkan menjadi file loan2.rds, untuk digunakan ketika membuat Credit Scoring Model:

```
# Save a single object to a file
saveRDS(loan2, "loan2.rds")
```

2. Analisis Data Eksploratif (EDA)

Status Pinjaman (loan_status)

```
library(dplyr)
library(ggplot2)

loan2 %>%
  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")
```



Untuk lebih memperjelas hasil analisis distribusi, diplotkan pula frekuensi loan_status dalam satuan %:

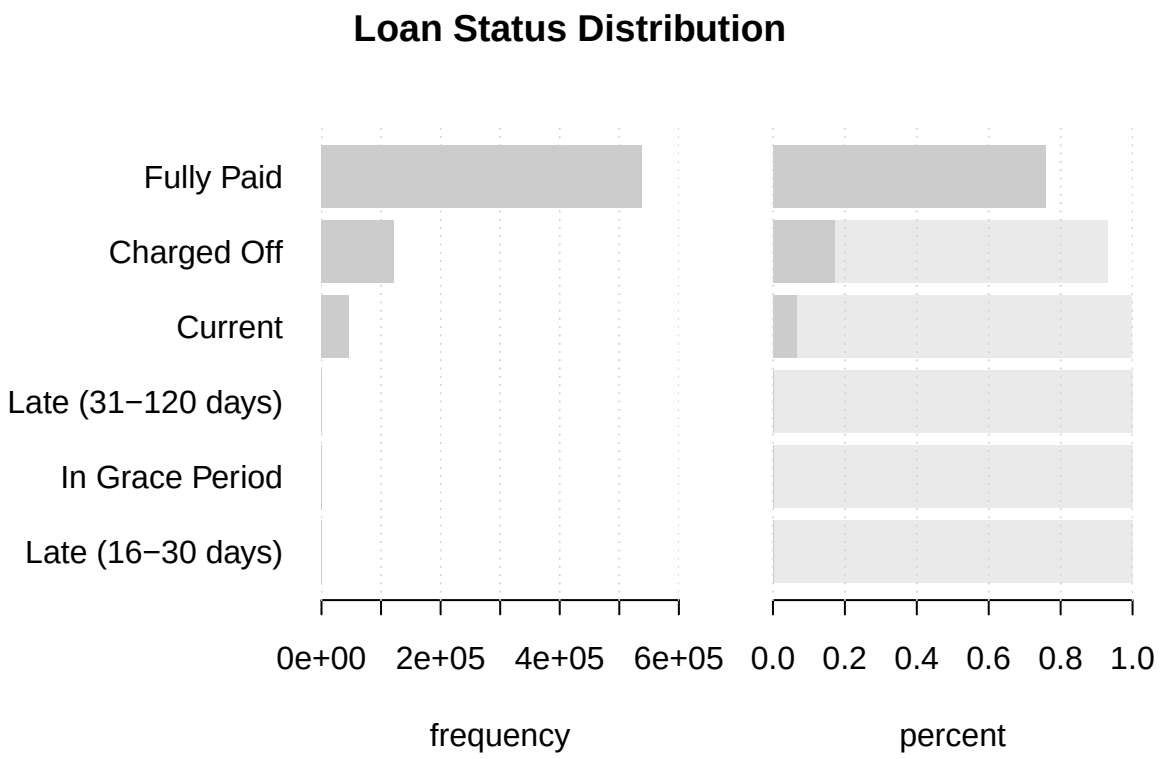
```
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 3.6.1
```

```
Desc(loan2$loan_status, main="Loan Status Distribution", plotit = TRUE)
```

```
## ------------------------------------------------------------------------
## Loan Status Distribution
##
##   length      n    NAs unique levels  dupes
##    7e+05  7e+05      0  6e+00  6e+00      y
##          100.0%   0.0%
##
```

```
##                level   freq   perc  cumfreq  cumperc
## 1          Fully Paid  5e+05  75.9%    5e+05    75.9%
## 2         Charged Off  1e+05  17.1%    7e+05    93.0%
## 3             Current  5e+04   6.6%    7e+05    99.6%
## 4  Late (31-120 days)  2e+03   0.2%    7e+05    99.9%
## 5     In Grace Period  7e+02   0.1%    7e+05   100.0%
## 6   Late (16-30 days)  3e+02   0.0%    7e+05   100.0%
```



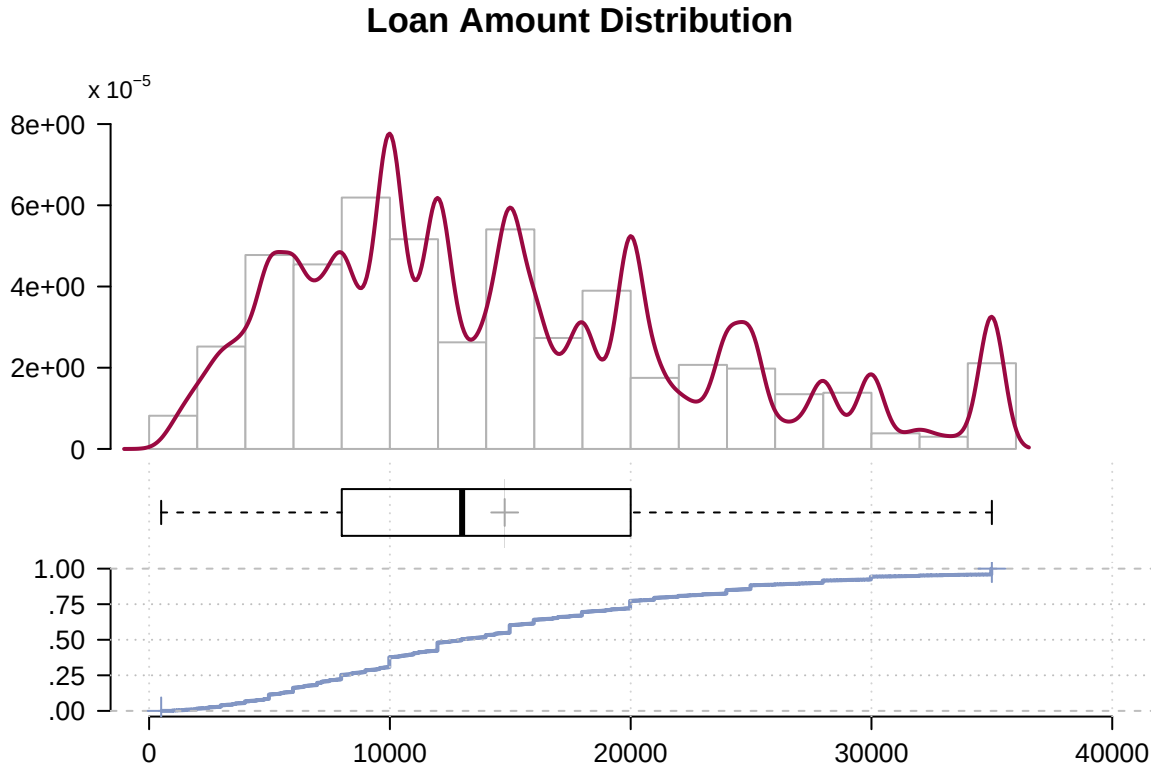
Jumlah Pinjaman (loan_amnt)

Berikut ini adalah visualisasi loan_amnt menggunakan density plot, box plot, dan empirical distribution function plot:

```
Desc(loan2$loan_amnt, main="Loan Amount Distribution", plotit=TRUE)
```

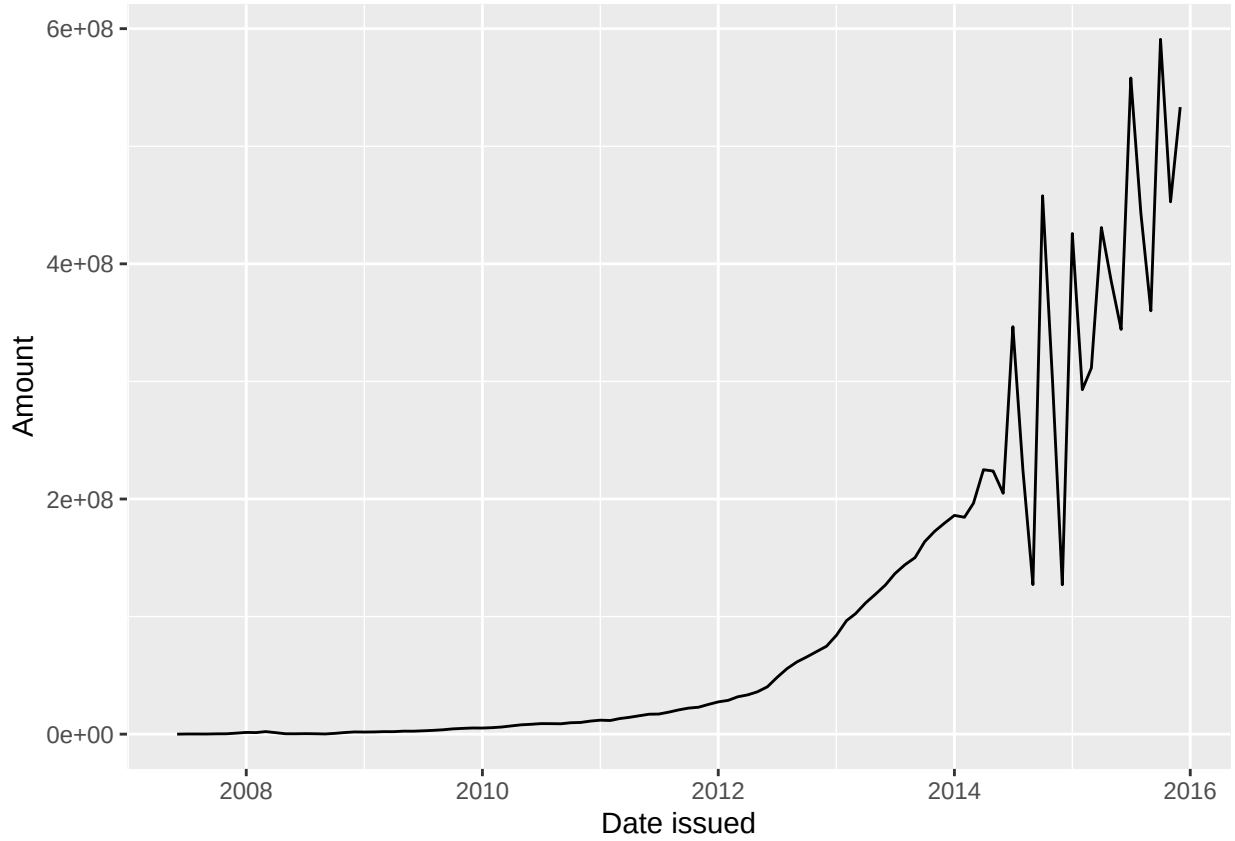
```
## ------------------------------------------------------------------------
## Loan Amount Distribution
##
##     length         n       NAs    unique        0s      mean     meanCI
##      7e+05     7e+05         0     1e+03         0  1.48e+04   1.47e+04
##             100.0%      0.0%              0.0%            1.48e+04
##
##        .05       .10       .25    median       .75       .90        .95
##   3.60e+03  5.00e+03  8.00e+03  1.30e+04  2.00e+04  2.80e+04   3.20e+04
##
##      range        sd     vcoef       mad       IQR      skew       kurt
##   3.45e+04  8.43e+03  5.71e-01  8.60e+03  1.20e+04  6.82e-01  -2.56e-01
```

```
##
## lowest : 5.00e+02 (5e+00), 7.00e+02, 7.50e+02, 8.00e+02, 9.00e+02
## highest: 3.49e+04 (1e+01), 3.49e+04 (5e+00), 3.50e+04 (1e+01), 3.50e+04 (2e+01), 3.50e+04 (3e+04)
```



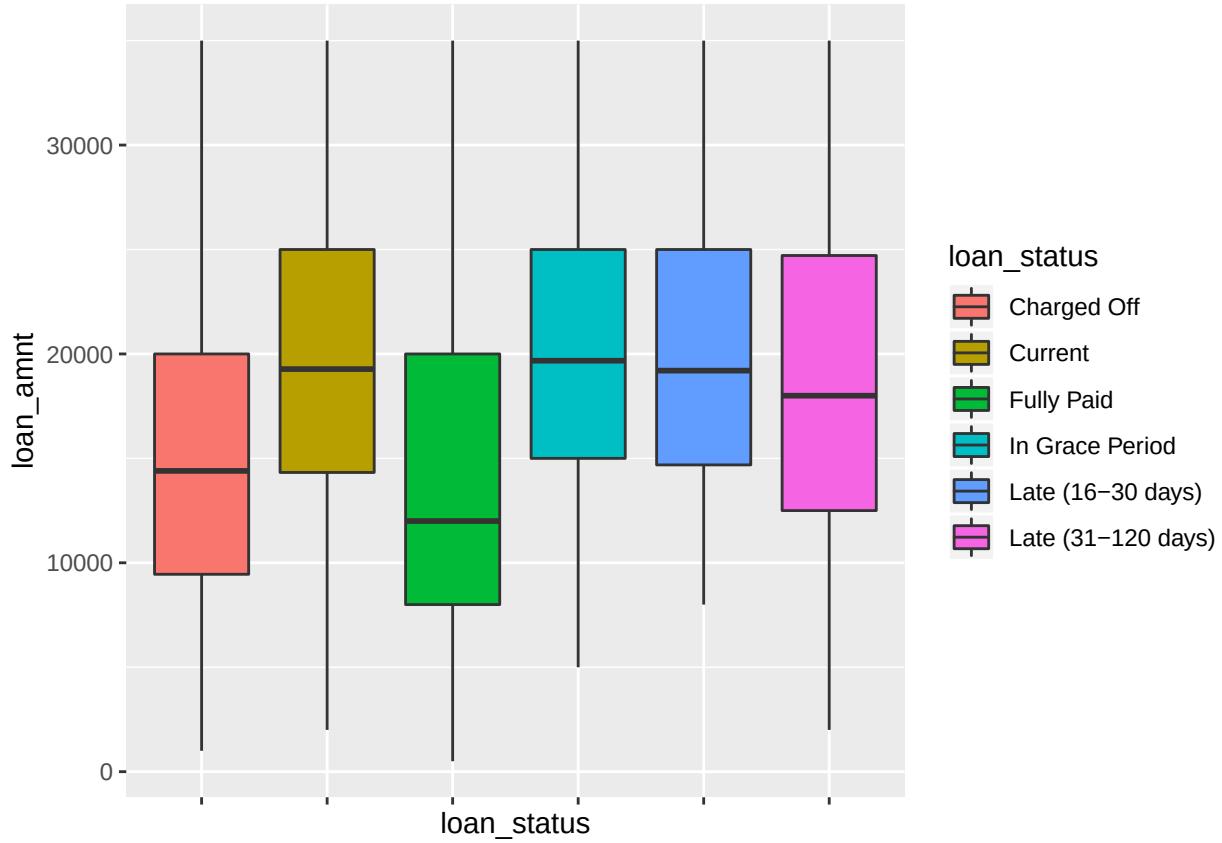
Dengan pertumbuhan jumlah pinjaman terhadap waktu sebagai berikut:

```
loan2$issue_d <- as.Date(gsub("^", "01-", loan2$issue_d), format="%d-%b-%Y")

loan_amnt_df <- loan2 %>%
  select(issue_d, loan_amnt) %>%
  group_by(issue_d) %>%
  summarise(Amount = sum(loan_amnt))

loan_amnt_ts <- ggplot(loan_amnt_df,
                  aes(x = issue_d, y = Amount))
loan_amnt_ts + geom_line() + xlab("Date issued")
```



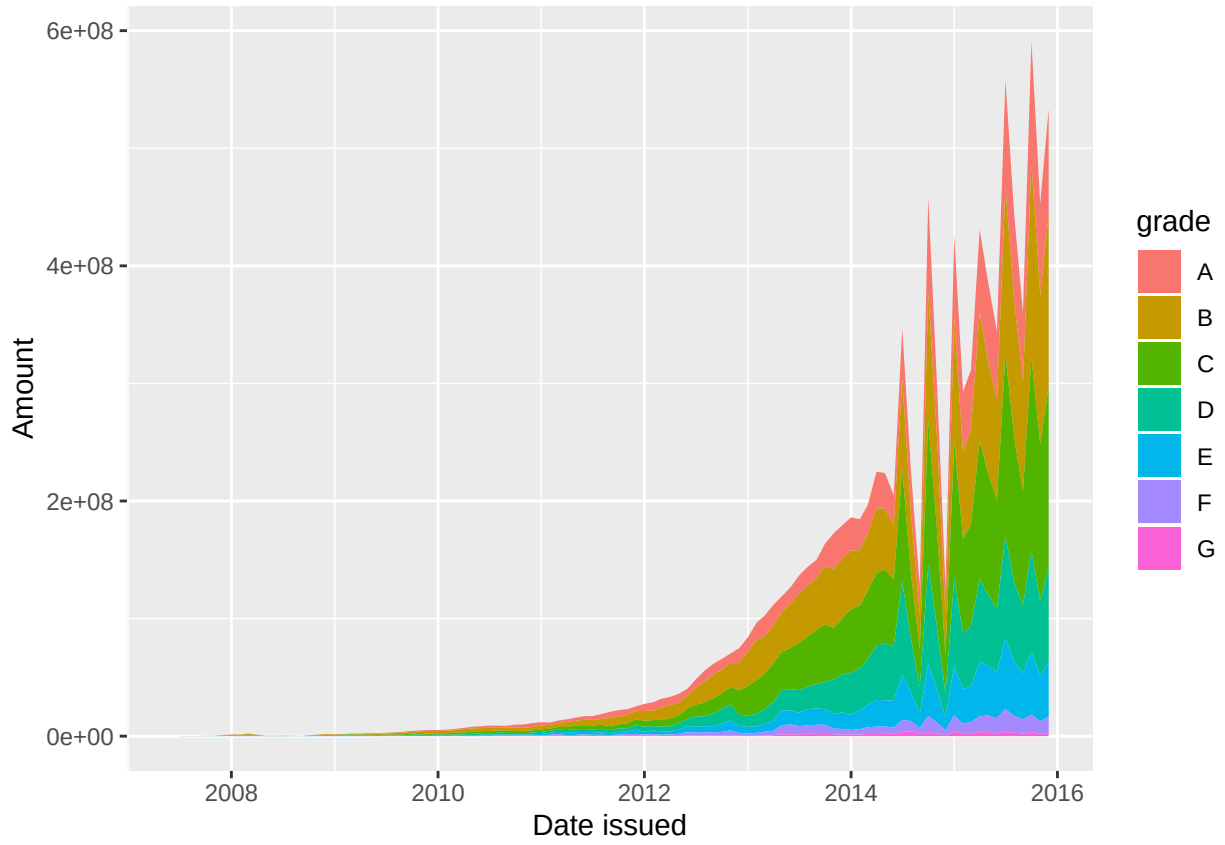
Distribusi jumlah pinjaman terhadap status pinjaman:

```
box_status <- ggplot(loan2, 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"))
```



Dengan pertumbuhan jumlah pinjaman terhadap waktu, apabila dikelompokkan berdasarkan grade, sebagai berikut:

```
loan_amnt_df_grade <- loan2 %>%
  select(issue_d, loan_amnt, grade) %>%
  group_by(issue_d, grade) %>%
  summarise(Amount = sum(loan_amnt))

loan_amnt_ts_grade <- ggplot(loan_amnt_df_grade,
                  aes(x = issue_d, y = Amount))
loan_amnt_ts_grade + geom_area(aes(fill=grade)) + xlab("Date issued")
```


6e+08
4e+08
2e+08
0e+00
Amount
2008
2010
2012
2014
2016
Date issued
grade
A
B
C
D
E
F
G