LendingClub

Jake Bowmer

23 February 2016

This is the R Markdown version of a brief analysis done on the publicly available lending club loan dataset.

```
#library imports and data import:
library(stringr)
library(dplyr)
library(lubridate)
library(randomForest)
library(reshape2)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.3
```

```
data = read.csv("/Users/Jake/Projects/LendingClub/LoanStats3a.csv", header=TRUE, stringsAsFactors=FALSE

head(data)
```

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

```
## 2 https://www.lendingclub.com/browse/loanDetail.action?loan_id=1077430
## 3 https://www.lendingclub.com/browse/loanDetail.action?loan_id=1077175
## 4 https://www.lendingclub.com/browse/loanDetail.action?loan_id=1076863
## 5 https://www.lendingclub.com/browse/loanDetail.action?loan_id=1075358
## 6 https://www.lendingclub.com/browse/loanDetail.action?loan_id=1075269
##
## 1
## 2   Borrower added on 12/22/11 > I plan to use this money to finance the motorcycle i am looking at.
## 3
## 4
## 5
## 6
##          purpose                                    title zip_code addr_state
## 1    credit_card                                Computer    860xx         AZ
## 2            car                                    bike    309xx         GA
## 3 small_business                    real estate business    606xx         IL
## 4          other                                personel    917xx         CA
## 5          other                                Personal    972xx         OR
## 6        wedding My wedding loan I promise to pay back    852xx         AZ
##     dti delinq_2yrs earliest_cr_line inq_last_6mths mths_since_last_delinq
## 1 27.65           0         Jan-1985              1                      NA
## 2  1.00           0         Apr-1999              5                      NA
## 3  8.72           0         Nov-2001              2                      NA
## 4 20.00           0         Feb-1996              1                      35
## 5 17.94           0         Jan-1996              0                      38
## 6 11.20           0         Nov-2004              3                      NA
##   mths_since_last_record open_acc pub_rec revol_bal revol_util total_acc
## 1                     NA        3       0     13648      83.7%         9
## 2                     NA        3       0      1687       9.4%         4
## 3                     NA        2       0      2956      98.5%        10
## 4                     NA       10       0      5598        21%        37
## 5                     NA       15       0     27783      53.9%        38
## 6                     NA        9       0      7963      28.3%        12
##   initial_list_status out_prncp out_prncp_inv total_pymnt total_pymnt_inv
## 1                   f       0.0           0.0    5861.071         5831.78
## 2                   f       0.0           0.0    1008.710         1008.71
## 3                   f       0.0           0.0    3003.654         3003.65
## 4                   f       0.0           0.0   12226.302        12226.30
## 5                   f     766.9         766.9    3242.170         3242.17
## 6                   f       0.0           0.0    5631.378         5631.38
##   total_rec_prncp total_rec_int total_rec_late_fee recoveries
## 1         5000.00        861.07               0.00       0.00
## 2          456.46        435.17               0.00     117.08
## 3         2400.00        603.65               0.00       0.00
## 4        10000.00       2209.33              16.97       0.00
## 5         2233.10       1009.07               0.00       0.00
## 6         5000.00        631.38               0.00       0.00
##   collection_recovery_fee last_pymnt_d last_pymnt_amnt next_pymnt_d
## 1                    0.00     Jan-2015          171.62
## 2                    1.11     Apr-2013          119.66
## 3                    0.00     Jun-2014          649.91
## 4                    0.00     Jan-2015          357.48
## 5                    0.00     Jan-2016           67.79     Feb-2016
## 6                    0.00     Jan-2015          161.03
```

```
##   last_credit_pull_d collections_12_mths_ex_med
## 1           Jan-2016                         0
## 2           Sep-2013                         0
## 3           Jan-2016                         0
## 4           Jan-2015                         0
## 5           Jan-2016                         0
## 6           Sep-2015                         0
##   mths_since_last_major_derog policy_code application_type
## 1                          NA           1       INDIVIDUAL
## 2                          NA           1       INDIVIDUAL
## 3                          NA           1       INDIVIDUAL
## 4                          NA           1       INDIVIDUAL
## 5                          NA           1       INDIVIDUAL
## 6                          NA           1       INDIVIDUAL
##   annual_inc_joint dti_joint verification_status_joint acc_now_delinq
## 1               NA        NA                        NA              0
## 2               NA        NA                        NA              0
## 3               NA        NA                        NA              0
## 4               NA        NA                        NA              0
## 5               NA        NA                        NA              0
## 6               NA        NA                        NA              0
##   tot_coll_amt tot_cur_bal open_acc_6m open_il_6m open_il_12m open_il_24m
## 1           NA          NA          NA         NA          NA          NA
## 2           NA          NA          NA         NA          NA          NA
## 3           NA          NA          NA         NA          NA          NA
## 4           NA          NA          NA         NA          NA          NA
## 5           NA          NA          NA         NA          NA          NA
## 6           NA          NA          NA         NA          NA          NA
##   mths_since_rcnt_il total_bal_il il_util open_rv_12m open_rv_24m
## 1                 NA           NA      NA          NA          NA
## 2                 NA           NA      NA          NA          NA
## 3                 NA           NA      NA          NA          NA
## 4                 NA           NA      NA          NA          NA
## 5                 NA           NA      NA          NA          NA
## 6                 NA           NA      NA          NA          NA
##   max_bal_bc all_util total_rev_hi_lim inq_fi total_cu_tl inq_last_12m
## 1         NA       NA               NA     NA          NA           NA
## 2         NA       NA               NA     NA          NA           NA
## 3         NA       NA               NA     NA          NA           NA
## 4         NA       NA               NA     NA          NA           NA
## 5         NA       NA               NA     NA          NA           NA
## 6         NA       NA               NA     NA          NA           NA
```

```
dim(data)
```

```
## [1] 42538    74
```

```
names(data)
```

```
##  [1] "id"                         "member_id"
##  [3] "loan_amnt"                  "funded_amnt"
##  [5] "funded_amnt_inv"            "term"
##  [7] "int_rate"                   "installment"
```

```
##  [9] "grade"                       "sub_grade"
## [11] "emp_title"                   "emp_length"
## [13] "home_ownership"              "annual_inc"
## [15] "verification_status"         "issue_d"
## [17] "loan_status"                 "pymnt_plan"
## [19] "url"                         "desc"
## [21] "purpose"                     "title"
## [23] "zip_code"                    "addr_state"
## [25] "dti"                         "delinq_2yrs"
## [27] "earliest_cr_line"            "inq_last_6mths"
## [29] "mths_since_last_delinq"      "mths_since_last_record"
## [31] "open_acc"                    "pub_rec"
## [33] "revol_bal"                   "revol_util"
## [35] "total_acc"                   "initial_list_status"
## [37] "out_prncp"                   "out_prncp_inv"
## [39] "total_pymnt"                 "total_pymnt_inv"
## [41] "total_rec_prncp"             "total_rec_int"
## [43] "total_rec_late_fee"          "recoveries"
## [45] "collection_recovery_fee"     "last_pymnt_d"
## [47] "last_pymnt_amnt"             "next_pymnt_d"
## [49] "last_credit_pull_d"          "collections_12_mths_ex_med"
## [51] "mths_since_last_major_derog" "policy_code"
## [53] "application_type"            "annual_inc_joint"
## [55] "dti_joint"                   "verification_status_joint"
## [57] "acc_now_delinq"              "tot_coll_amt"
## [59] "tot_cur_bal"                 "open_acc_6m"
## [61] "open_il_6m"                  "open_il_12m"
## [63] "open_il_24m"                 "mths_since_rcnt_il"
## [65] "total_bal_il"                "il_util"
## [67] "open_rv_12m"                 "open_rv_24m"
## [69] "max_bal_bc"                  "all_util"
## [71] "total_rev_hi_lim"            "inq_fi"
## [73] "total_cu_tl"                 "inq_last_12m"
```

```
summary(data)
```

```
##       id              member_id         loan_amnt      funded_amnt
##  Length:42538       Min.   :  70473   Min.   :  500   Min.   :  500
##  Class :character   1st Qu.: 638480   1st Qu.: 5200   1st Qu.: 5000
##  Mode  :character   Median : 824178   Median : 9700   Median : 9600
##                     Mean   : 825703   Mean   :11090   Mean   :10822
##                     3rd Qu.:1033946   3rd Qu.:15000   3rd Qu.:15000
##                     Max.   :1314167   Max.   :35000   Max.   :35000
##                     NA's   :3         NA's   :3       NA's   :3
##  funded_amnt_inv     term             int_rate          installment
##  Min.   :    0   Length:42538       Length:42538       Min.   :  15.67
##  1st Qu.: 4950   Class :character   Class :character   1st Qu.: 165.52
##  Median : 8500   Mode  :character   Mode  :character   Median : 277.69
##  Mean   :10140                                         Mean   : 322.62
##  3rd Qu.:14000                                         3rd Qu.: 428.18
##  Max.   :35000                                         Max.   :1305.19
##  NA's   :3                                             NA's   :3
##     grade            sub_grade          emp_title
##  Length:42538       Length:42538       Length:42538
```

```
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   emp_length        home_ownership       annual_inc
##  Length:42538       Length:42538       Min.   :   1896
##  Class :character   Class :character   1st Qu.:  40000
##  Mode  :character   Mode  :character   Median :  59000
##                                        Mean   :  69137
##                                        3rd Qu.:  82500
##                                        Max.   :6000000
##                                        NA's   :7
##  verification_status   issue_d          loan_status
##  Length:42538        Length:42538       Length:42538
##  Class :character    Class :character   Class :character
##  Mode  :character    Mode  :character   Mode  :character
##
##
##
##
##   pymnt_plan            url               desc
##  Length:42538       Length:42538       Length:42538
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    purpose             title             zip_code
##  Length:42538       Length:42538       Length:42538
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   addr_state             dti         delinq_2yrs      earliest_cr_line
##  Length:42538       Min.   : 0.00   Min.   : 0.0000   Length:42538
##  Class :character   1st Qu.: 8.20   1st Qu.: 0.0000   Class :character
##  Mode  :character   Median :13.47   Median : 0.0000   Mode  :character
##                     Mean   :13.37   Mean   : 0.1525
##                     3rd Qu.:18.68   3rd Qu.: 0.0000
##                     Max.   :29.99   Max.   :13.0000
##                     NA's   :3       NA's   :32
##  inq_last_6mths   mths_since_last_delinq mths_since_last_record
##  Min.   : 0.000   Min.   :  0.00         Min.   :  0.00
##  1st Qu.: 0.000   1st Qu.: 17.00         1st Qu.:  0.00
##  Median : 1.000   Median : 33.00         Median : 85.00
##  Mean   : 1.081   Mean   : 35.02         Mean   : 59.18
##  3rd Qu.: 2.000   3rd Qu.: 51.00         3rd Qu.:101.00
##  Max.   :33.000   Max.   :120.00         Max.   :129.00
##  NA's   :32       NA's   :26929          NA's   :38887
```

```
##     open_acc         pub_rec          revol_bal        revol_util
##  Min.   : 1.000   Min.   :0.00000   Min.   :      0   Length:42538
##  1st Qu.: 6.000   1st Qu.:0.00000   1st Qu.:   3635   Class :character
##  Median : 9.000   Median :0.00000   Median :   8821   Mode  :character
##  Mean   : 9.344   Mean   :0.05816   Mean   :  14298
##  3rd Qu.:12.000   3rd Qu.:0.00000   3rd Qu.:  17251
##  Max.   :47.000   Max.   :5.00000   Max.   :1207359
##  NA's   :32       NA's   :32        NA's   :3
##    total_acc     initial_list_status   out_prncp       out_prncp_inv
##  Min.   : 1.00   Length:42538        Min.   :    0.0   Min.   :    0.0
##  1st Qu.:13.00   Class :character    1st Qu.:    0.0   1st Qu.:    0.0
##  Median :20.00   Mode  :character    Median :    0.0   Median :    0.0
##  Mean   :22.12                       Mean   :  128.3   Mean   :  127.5
##  3rd Qu.:29.00                       3rd Qu.:    0.0   3rd Qu.:    0.0
##  Max.   :90.00                       Max.   :10418.1   Max.   :10410.6
##  NA's   :32                          NA's   :3         NA's   :3
##   total_pymnt    total_pymnt_inv total_rec_prncp total_rec_int
##  Min.   :    0   Min.   :    0   Min.   :    0   Min.   :    0
##  1st Qu.: 5456   1st Qu.: 4781   1st Qu.: 4382   1st Qu.:  656
##  Median : 9655   Median : 8930   Median : 8000   Median : 1337
##  Mean   :11881   Mean   :11175   Mean   : 9554   Mean   : 2229
##  3rd Qu.:16255   3rd Qu.:15368   3rd Qu.:13000   3rd Qu.: 2795
##  Max.   :56809   Max.   :56475   Max.   :35000   Max.   :23062
##  NA's   :3       NA's   :3       NA's   :3       NA's   :3
##  total_rec_late_fee   recoveries       collection_recovery_fee
##  Min.   :  0.000    Min.   :    0.00   Min.   :   0.00
##  1st Qu.:  0.000    1st Qu.:    0.00   1st Qu.:   0.00
##  Median :  0.000    Median :    0.00   Median :   0.00
##  Mean   :  1.485    Mean   :   97.28   Mean   :  13.41
##  3rd Qu.:  0.000    3rd Qu.:    0.00   3rd Qu.:   0.00
##  Max.   :208.820    Max.   :29623.35   Max.   :7002.19
##  NA's   :3          NA's   :3          NA's   :3
##  last_pymnt_d       last_pymnt_amnt   next_pymnt_d
##  Length:42538       Min.   :    0.0   Length:42538
##  Class :character   1st Qu.:  211.5   Class :character
##  Mode  :character   Median :  518.6   Mode  :character
##                     Mean   : 2596.0
##                     3rd Qu.: 3114.0
##                     Max.   :36115.2
##                     NA's   :3
##  last_credit_pull_d collections_12_mths_ex_med mths_since_last_major_derog
##  Length:42538       Min.   :0                  Mode:logical
##  Class :character   1st Qu.:0                  NA's:42538
##  Mode  :character   Median :0
##                     Mean   :0
##                     3rd Qu.:0
##                     Max.   :0
##                     NA's   :148
##   policy_code application_type   annual_inc_joint dti_joint
##  Min.   :1    Length:42538       Mode:logical     Mode:logical
##  1st Qu.:1    Class :character   NA's:42538       NA's:42538
##  Median :1    Mode  :character
##  Mean   :1
##  3rd Qu.:1
```

```
##  Max.   :1
##  NA's   :3
##  verification_status_joint acc_now_delinq  tot_coll_amt   tot_cur_bal
##  Mode:logical              Min.   :0e+00   Mode:logical   Mode:logical
##  NA's:42538                1st Qu.:0e+00   NA's:42538     NA's:42538
##                            Median :0e+00
##                            Mean   :9e-05
##                            3rd Qu.:0e+00
##                            Max.   :1e+00
##                            NA's   :32
##  open_acc_6m    open_il_6m     open_il_12m    open_il_24m
##  Mode:logical   Mode:logical   Mode:logical   Mode:logical
##  NA's:42538     NA's:42538     NA's:42538     NA's:42538
##
##
##
##
##
##  mths_since_rcnt_il total_bal_il   il_util        open_rv_12m
##  Mode:logical       Mode:logical   Mode:logical   Mode:logical
##  NA's:42538         NA's:42538     NA's:42538     NA's:42538
##
##
##
##
##
##  open_rv_24m    max_bal_bc     all_util       total_rev_hi_lim
##  Mode:logical   Mode:logical   Mode:logical   Mode:logical
##  NA's:42538     NA's:42538     NA's:42538     NA's:42538
##
##
##
##
##
##   inq_fi        total_cu_tl    inq_last_12m
##  Mode:logical   Mode:logical   Mode:logical
##  NA's:42538     NA's:42538     NA's:42538
##
##
##
##
##
```

From the above, we can see that there are a number of columns consisting almost entirely of NA values. In addition, the description column, while potentially useful from a text mining point of view, is outside the scope of this example. We remove these columns.

```
data = subset(data, select = -desc)
data = select(data, id:last_credit_pull_d)
```

A simple model predicting whether a loan turns "bad" or not is a binary classification problem. We need to adjust our dataset to highlight the loans that have gone bad.

```r
#create a binary indicator of whether a loan is considered bad or good.
bad = c("Late (31-120 days)", "Default", "Charged Off")
data$bad_loans = ifelse(data$loan_status %in% bad, 1, ifelse(data$loan_status=="", NA, 0))
data$bad_loans = factor(data$bad_loans)
```

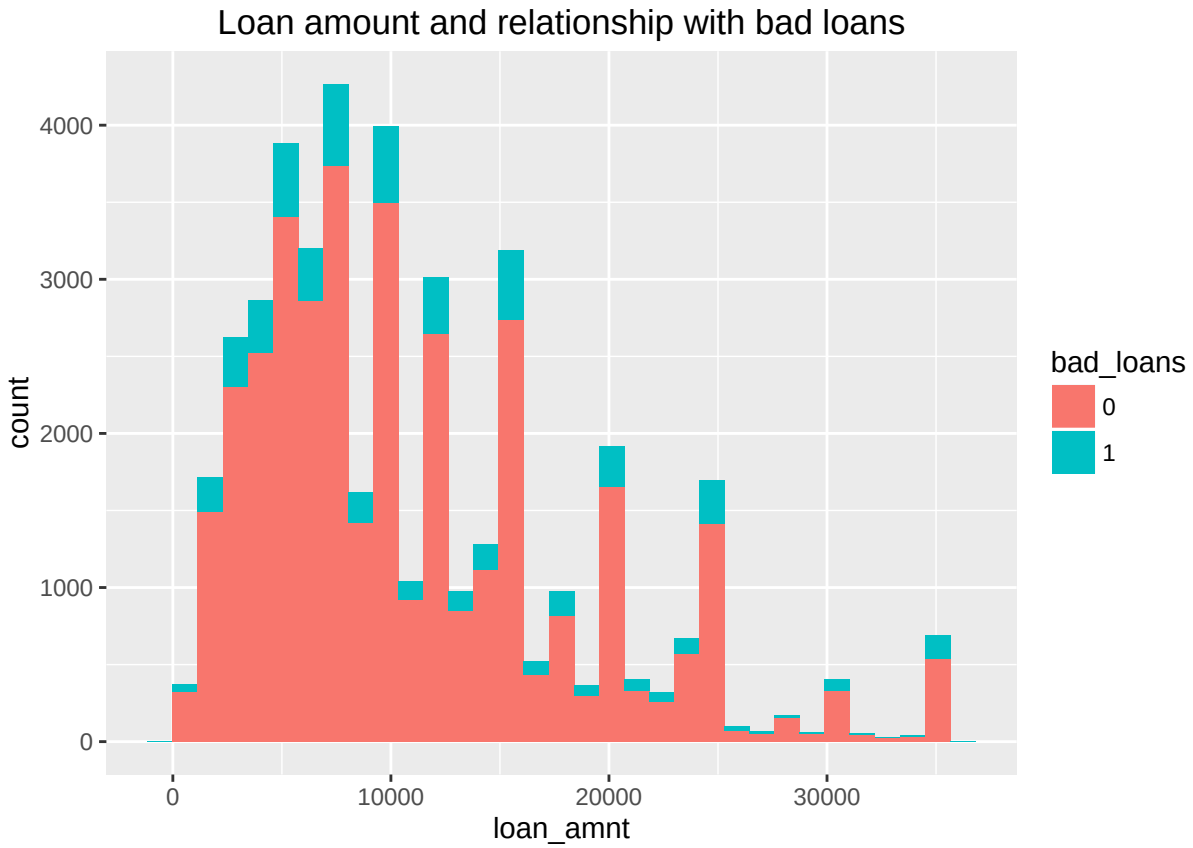
Analysis

This results in 5634 bad loans from a total of 42538. We can use this column to investigates potential explanatory variables. We briefly look at, for examples sake, loan amount, term, interest rate and grade and effects on loan outcomes. We first look at loan amount and how bad loans vary with the loan amount.

```r
ggplot(data = data, aes(x = loan_amnt, fill = bad_loans)) + geom_histogram() +
  ggtitle ("Loan amount and relationship with bad loans")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

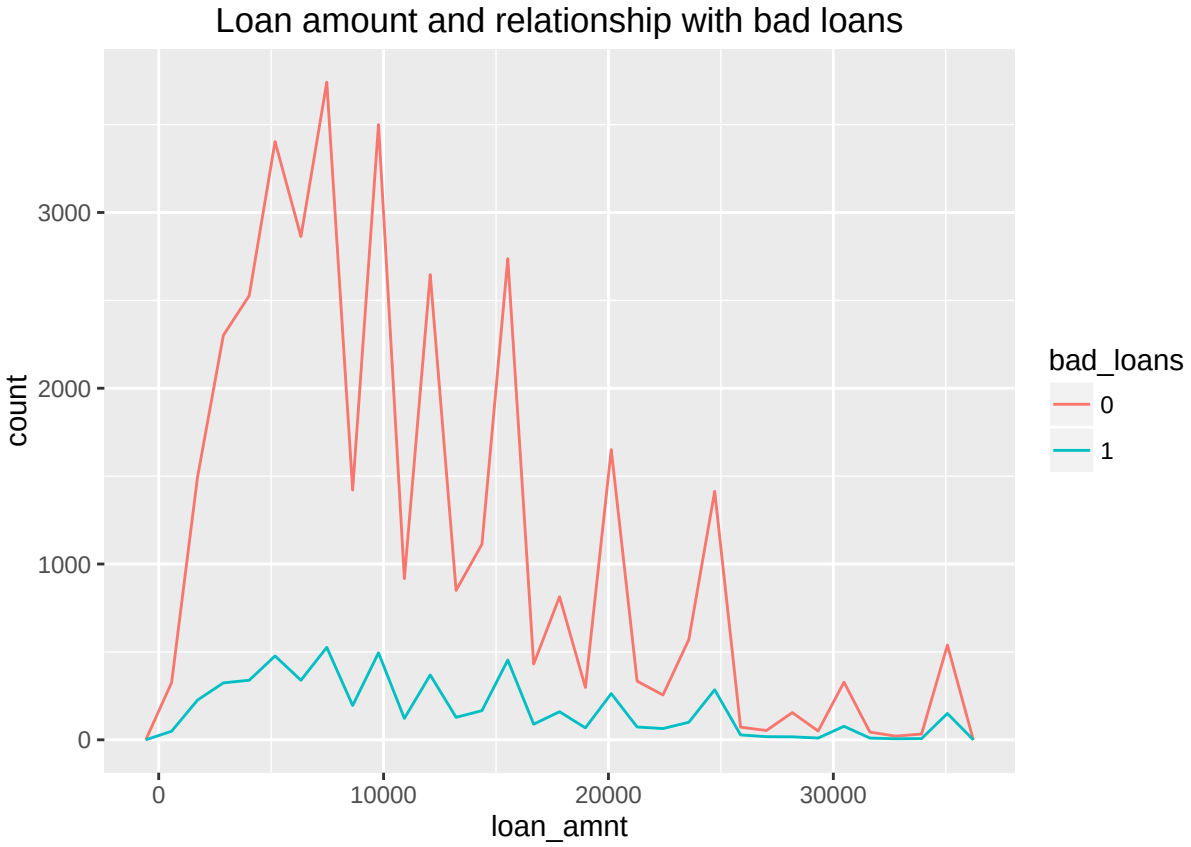
```
## Warning: Removed 3 rows containing non-finite values (stat_bin).
```




```r
ggplot(data = data, aes(x = loan_amnt, color = bad_loans)) + geom_freqpoly() +
  ggtitle("Loan amount and relationship with bad loans")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

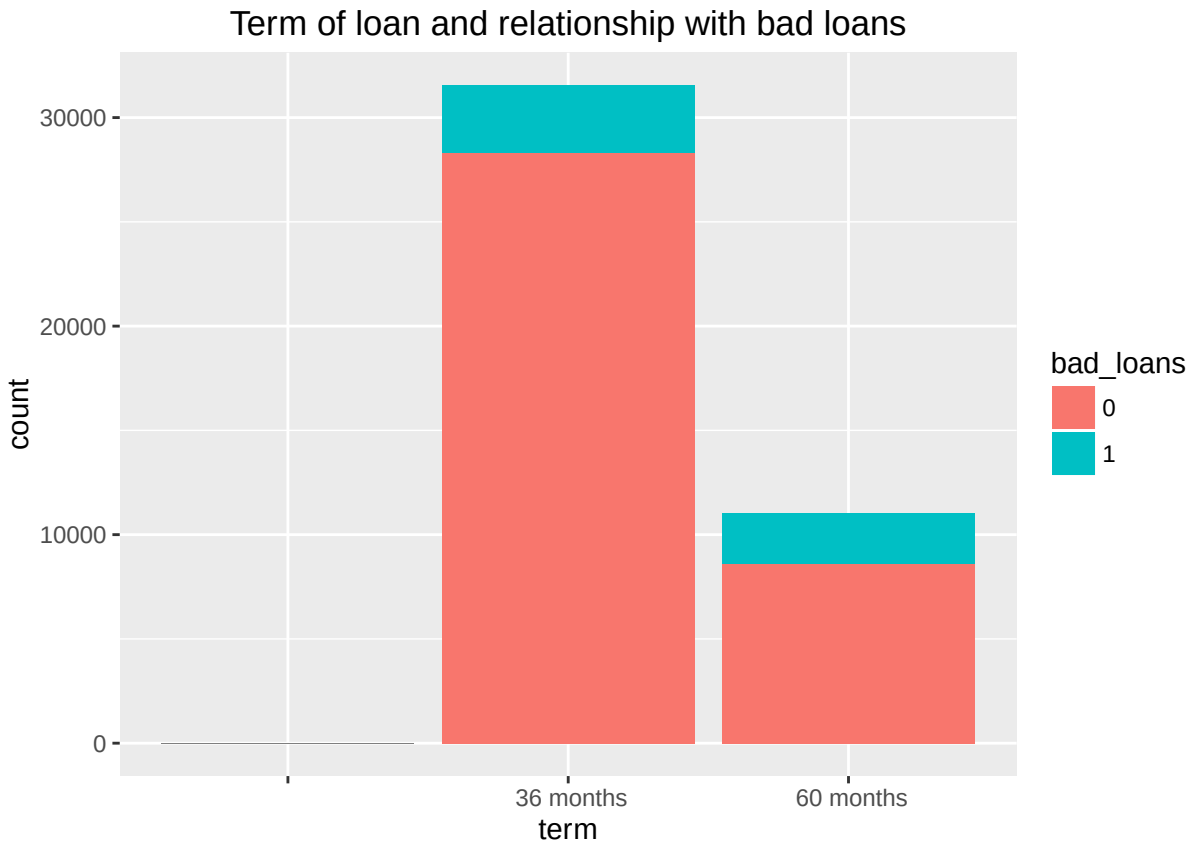
```
## Warning: Removed 3 rows containing non-finite values (stat_bin).
```



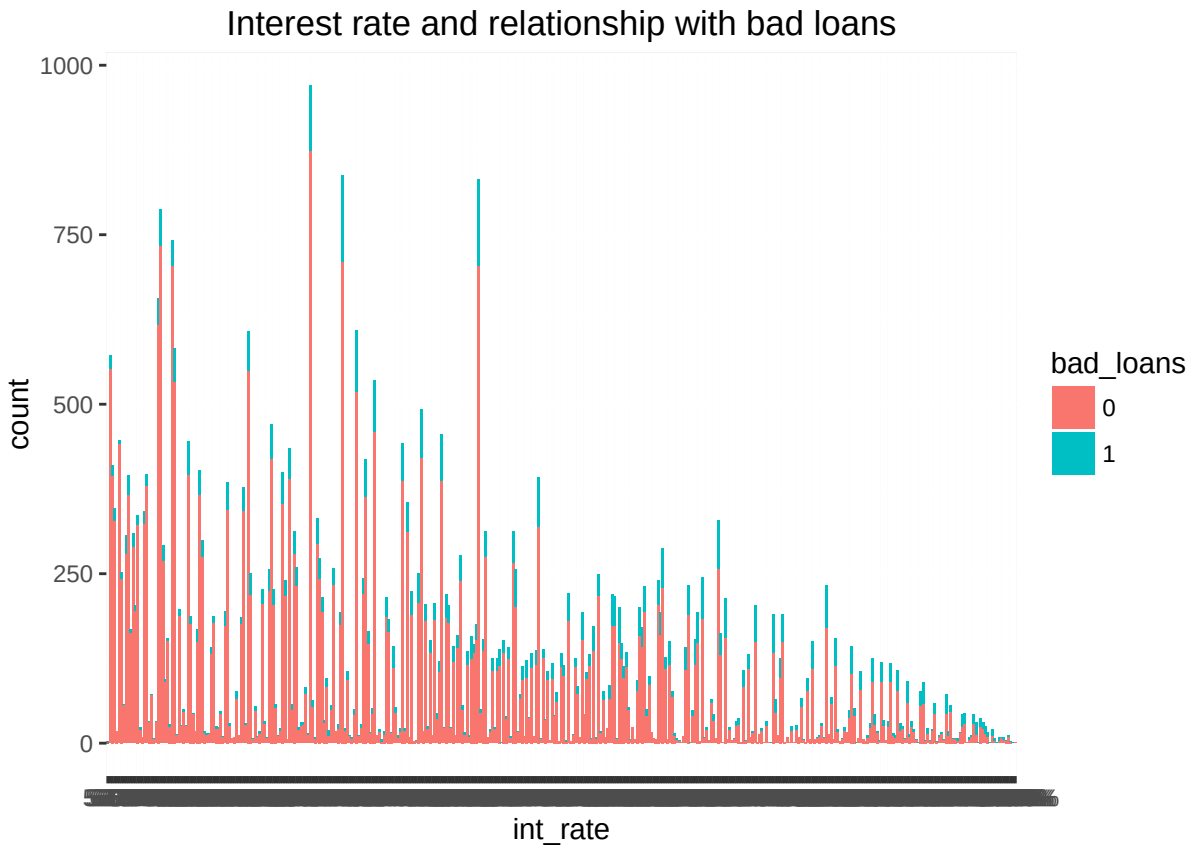
By far the most interesting part of the chart above is the spike in

We also examine some other variables and the relationship with bad loans. Of particular interest is the chart showing the change in bad loans as grade changes - we would expect this to increase as grade falls, but this seems not to be the case at first glance. This would most likely be an interesting avenue for further investigation.

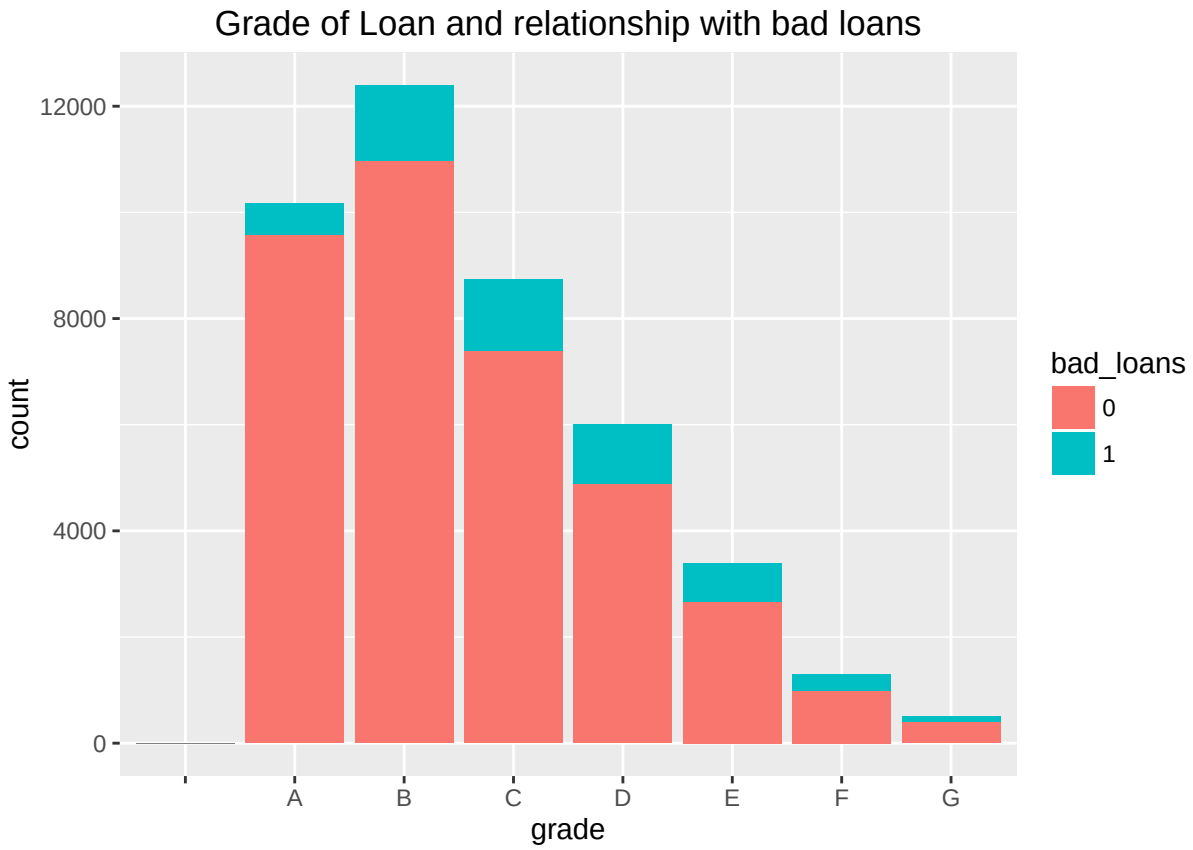
```
ggplot(data = data, aes(x = term,  fill =  bad_loans)) + geom_bar() +
  ggtitle("Term of loan and relationship with bad loans")
```




```
ggplot(data = data, aes(x = int_rate, fill = bad_loans)) + geom_bar() +
  ggtitle("Interest rate and relationship with bad loans") #This chart would be improved with binning i
```




```
ggplot(data = data, aes(x = grade, fill = bad_loans)) + geom_bar() +
  ggtitle("Grade of Loan and relationship with bad loans") #Potentially the most interesting chart of t
```



Finally, we create a simple model and look at its performance on a test set.

```
data$renter = ifelse(data$home_ownership == 'RENT', 1, 0)
data_rf = select(data, loan_amnt, term, grade, renter, annual_inc, bad_loans)
summary(data_rf)
```

```
##    loan_amnt         term               grade              renter
##  Min.   :  500   Length:42538       Length:42538       Min.   :0.0000
##  1st Qu.: 5200   Class :character   Class :character   1st Qu.:0.0000
##  Median : 9700   Mode  :character   Mode  :character   Median :0.0000
##  Mean   :11090                                         Mean   :0.4744
##  3rd Qu.:15000                                         3rd Qu.:1.0000
##  Max.   :35000                                         Max.   :1.0000
##  NA's   :3
##    annual_inc      bad_loans
##  Min.   :   1896   0   :36901
##  1st Qu.:  40000   1   : 5634
##  Median :  59000   NA's:    3
##  Mean   :  69137
##  3rd Qu.:  82500
##  Max.   :6000000
##  NA's   :7
```

```
data_rf$grade = factor(data_rf$grade)
data_rf$term = factor(data_rf$term)
data_rf = na.omit(data_rf)
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
split = createDataPartition(data_rf$bad_loans, p=0.75, list = FALSE)
training = data_rf[split,]
test = data_rf[-split,]

#train model on training set
rf = randomForest(as.factor(bad_loans) ~ loan_amnt + term + grade + renter + annual_inc,
                  type="classification", data=training, importance=TRUE, na.action=na.omit)
#Predict on test set
data_rf_predict = factor(predict(rf, newdata = test))

#We can see that our model is not yet very useful:
table(data_rf_predict, test$bad_loans)
```

```
##
## data_rf_predict    0    1
##               0 9221 1406
##               1    3    2
```

We see that the simple model is not particularly useful. The addition of the fico score makes the problem solvable, but lending club has scrubbed fico data from its public datasets as the addition of fico data causes privacy issues.