Project5

Group 5AA_Black-Thenmozhi Boopathy

2025-12-26

OOS PERFORMANCE COMPETITION - Predict New Observations as Default = Yes or No

In this project, we are using the given scoring data that contains 79243 observations with 36 variables. We are doing the pre processing methods for the scoreing data.For the initial start, just looking at the structure to make a proper execution in predictor variables.

```
df = read.csv("score.csv", stringsAsFactors = TRUE)
train_df = read.csv("train.csv", stringsAsFactors = TRUE)
train_median <- median(train_df$revol_util, na.rm = TRUE)
str(df)
```

```
## 'data.frame':    79243 obs. of  36 variables:
##  $ ID                    : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ loan_amnt             : int  35000 2600 7000 22250 16000 14000 26650 10000 20000 15000 ...
##  $ term                  : Factor w/ 2 levels " 36 months"," 60 months": 1 1 1 1 2 2 2 1 1 2 ...
##  $ int_rate              : num  13.7 13.7 12.1 21 12.5 ...
##  $ installment           : num  1190.6 88.5 232.9 838.3 359.9 ...
##  $ grade                 : Factor w/ 7 levels "A","B","C","D",..: 3 3 2 5 2 3 6 3 2 4 ...
##  $ sub_grade             : Factor w/ 35 levels "A1","A2","A3",..: 14 11 8 21 10 11 28 12 9 17
...
##  $ emp_title             : Factor w/ 43586 levels "","  Credit rev specialist",..: 28879 10774 3
7115 32204 38448 18071 24913 2823 38743 37148 ...
##  $ emp_length            : Factor w/ 12 levels "","< 1 year",..: 4 5 4 4 4 4 11 2 5 8 ...
##  $ home_ownership        : Factor w/ 6 levels "ANY","MORTGAGE",..: 6 2 6 2 6 6 2 6 2 6 ...
##  $ annual_inc            : num  89000 103000 80000 94000 52000 37000 58000 52000 170000 40000
...
##  $ verification_status   : Factor w/ 3 levels "Not Verified",..: 3 2 1 2 2 2 3 2 3 2 ...
##  $ issue_d               : Factor w/ 114 levels "Apr-2008","Apr-2009",..: 93 54 25 91 64 92 8 8
82 42 ...
##  $ purpose               : Factor w/ 14 levels "car","credit_card",..: 3 3 3 3 2 2 2 3 3 10 ...
##  $ title                 : Factor w/ 12237 levels "","'08 & '09 Roth IRA Investments",..: 3870 3
870 8288 3870 3139 3139 3139 3870 3870 1893 ...
##  $ dti                   : num  22.4 26.8 18.9 12.4 19.3 ...
##  $ earliest_cr_line      : Factor w/ 618 levels "Apr-1961","Apr-1962",..: 552 245 241 293 185 55
2 450 549 541 547 ...
##  $ open_acc              : int  8 16 7 10 11 12 22 20 14 8 ...
##  $ pub_rec               : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ revol_bal             : int  14210 31492 2402 14334 8434 13459 18644 4954 9652 0 ...
##  $ revol_util            : num  60 44.4 63.2 69.6 39.6 59.3 39.6 84 36 0 ...
##  $ total_acc             : int  21 36 38 29 15 29 37 44 34 43 ...
##  $ initial_list_status   : Factor w/ 2 levels "f","w": 2 1 1 1 2 2 1 2 1 1 ...
##  $ application_type      : Factor w/ 3 levels "DIRECT_PAY","INDIVIDUAL",..: 2 2 2 2 2 2 2 2 2 2
...
##  $ mort_acc              : int  0 3 1 4 1 0 2 0 6 NA ...
##  $ pub_rec_bankruptcies  : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ address               : Factor w/ 79108 levels "000 Adrian Cliffs\nRandyton, LA 22690",..: 41
639 65134 66277 27304 57318 77233 25031 16448 28914 66627 ...
##  $ delinq_2yrs           : int  1 0 0 1 0 0 0 0 0 0 ...
##  $ fico_range_low        : int  695 710 680 665 740 665 700 685 685 765 ...
##  $ fico_range_high       : int  699 714 684 669 744 669 704 689 689 769 ...
##  $ inq_last_6mths        : int  0 3 0 1 0 1 3 0 1 1 ...
##  $ mths_since_last_delinq: int  17 NA NA 12 NA NA NA 52 48 NA ...
##  $ last_credit_pull_d    : Factor w/ 131 levels "","Apr-2009",..: 42 64 20 21 116 118 56 120 118
89 ...
##  $ acc_now_delinq        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hardship_flag         : Factor w/ 1 level "N": 1 1 1 1 1 1 1 1 1 1 ...
##  $ debt_settlement_flag  : Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 2 1 1 ...
```

Summary Statistics for numeric Variables

We make sure which variables names are comes under the category of bank_numeric and bank_factor using head(bank_numeric) head(bank_factor).

```
#Split data into categorical and numeric datasets
bank_numeric <- select_if(df, is.numeric)
bank_factor <- select_if(df, is.factor)
#Ensure dataframes correctly split
head(bank_numeric)
```

```
##   ID loan_amnt int_rate installment annual_inc   dti open_acc pub_rec revol_bal
## 1  0     35000    13.67     1190.62      89000 22.35        8       0     14210
## 2  1      2600    13.68       88.46     103000 26.77       16       0     31492
## 3  2      7000    12.12      232.91      80000 18.86        7       0      2402
## 4  3     22250    21.00      838.28      94000 12.39       10       0     14334
## 5  4     16000    12.49      359.89      52000 19.32       11       0      8434
## 6  5     14000    12.39      314.19      37000 30.49       12       1     13459
##   revol_util total_acc mort_acc pub_rec_bankruptcies delinq_2yrs fico_range_low
## 1       60.0        21        0                    0           1            695
## 2       44.4        36        3                    0           0            710
## 3       63.2        38        1                    0           0            680
## 4       69.6        29        4                    0           1            665
## 5       39.6        15        1                    0           0            740
## 6       59.3        29        0                    1           0            665
##   fico_range_high inq_last_6mths mths_since_last_delinq acc_now_delinq
## 1             699              0                     17              0
## 2             714              3                     NA              0
## 3             684              0                     NA              0
## 4             669              1                     12              0
## 5             744              0                     NA              0
## 6             669              1                     NA              0
```

```
head(bank_factor)
```

```
##         term grade sub_grade              emp_title emp_length home_ownership
## 1  36 months     C        C4        Platoon sergeant  10+ years           RENT
## 2  36 months     C        C1   Department of Defense    2 years       MORTGAGE
## 3  36 months     B        B3 State of California-CHP  10+ years           RENT
## 4  36 months     E        E1          Rn staff nurse  10+ years       MORTGAGE
## 5  60 months     B        B5   Tariff Representative  10+ years           RENT
## 6  60 months     C        C1           House Keeping  10+ years           RENT
##   verification_status  issue_d            purpose                   title
## 1            Verified Nov-2015 debt_consolidation      Debt consolidation
## 2     Source Verified Jul-2013 debt_consolidation      Debt consolidation
## 3        Not Verified Dec-2012 debt_consolidation                   my loan
## 4     Source Verified Nov-2013 debt_consolidation      Debt consolidation
## 5     Source Verified Jun-2014        credit_card Credit card refinancing
## 6     Source Verified Nov-2014        credit_card Credit card refinancing
##   earliest_cr_line initial_list_status application_type
## 1         Oct-1998                   w       INDIVIDUAL
## 2         Jan-1998                   f       INDIVIDUAL
## 3         Jan-1994                   f       INDIVIDUAL
## 4         Jul-1992                   f       INDIVIDUAL
## 5         Feb-1994                   w       INDIVIDUAL
## 6         Oct-1998                   w       INDIVIDUAL
##                                             address last_credit_pull_d
## 1           5884 Valdez Ridges\nPort Davidfort, DC 86630           Feb-2017
## 2 92228 Cooper Flats Apt. 703\nNorth Alexander, UT 29597           Jul-2016
## 3     93894 Bianca Rest Apt. 713\nPort Jessica, NH 22690           Aug-2015
## 4     38728 Sanders Knolls Apt. 028\nJasonstad, AZ 22690           Aug-2016
## 5       811 Leon Glens Apt. 014\nEast Danielle, LA 70466           Oct-2014
## 6                             USNS Carter\nFPO AE 11650           Oct-2016
##   hardship_flag debt_settlement_flag
## 1             N                    N
## 2             N                    N
## 3             N                    N
## 4             N                    N
## 5             N                    N
## 6             N                    N
```

Summary Statistics for numeric Variables

Next, we are trying to understand the patterns and detect anomalies and simplify the dataset for targeting better prediction. Each variable gives important values such as the mean and median with the min and max values helping identify outliers. We can also find early indicators of numeric variables that could potentially be manipulated into factor variables or factor lumped to help our prediction model.

```
summary(bank_numeric)
```

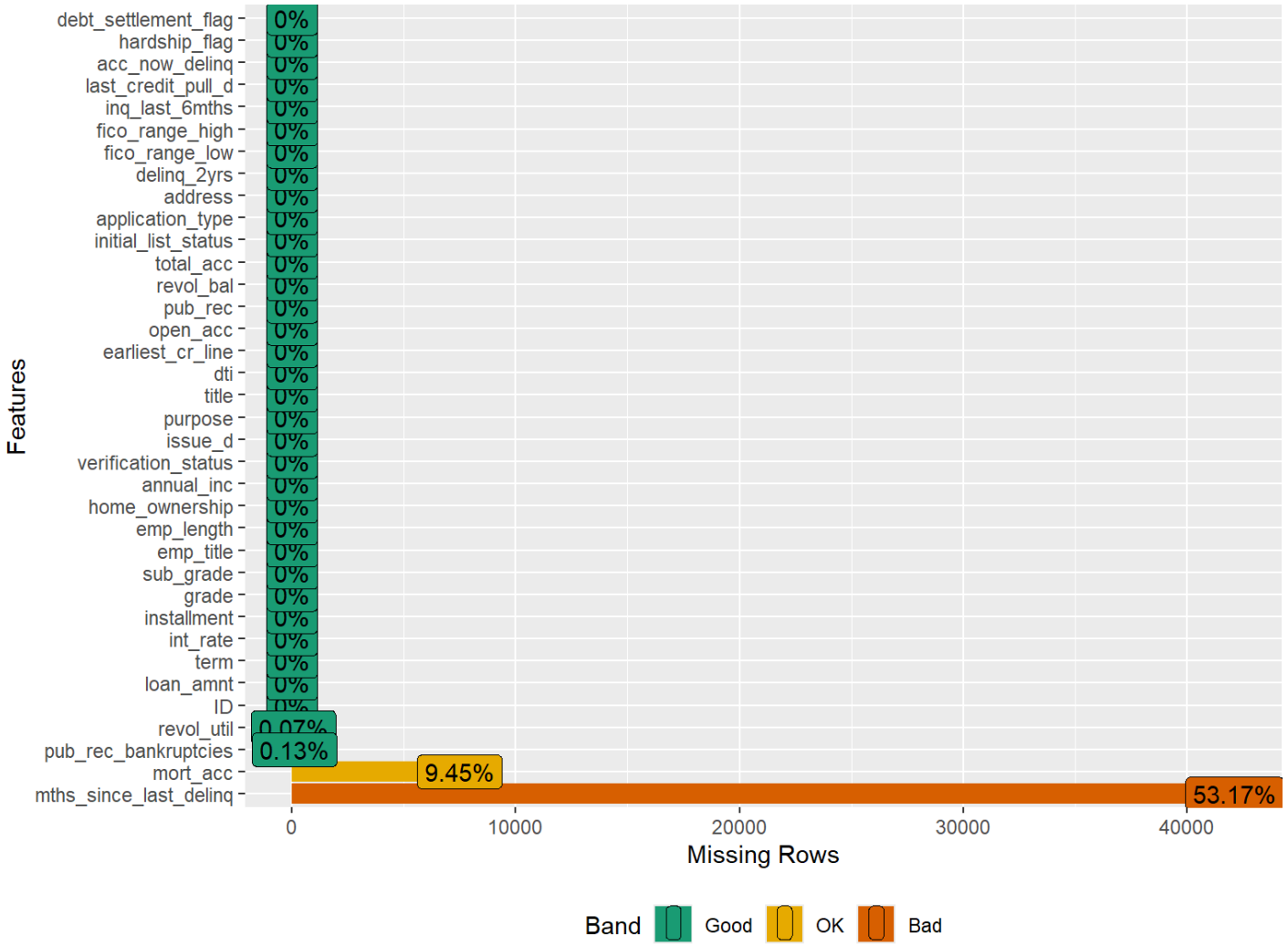
```
##        ID          loan_amnt        int_rate      installment
##  Min.   :    0   Min.   :  950   Min.   : 5.32   Min.   :  22.24
##  1st Qu.:19811   1st Qu.: 8000   1st Qu.:10.49   1st Qu.: 250.33
##  Median :39621   Median :12000   Median :13.33   Median : 375.38
##  Mean   :39621   Mean   :14124   Mean   :13.68   Mean   : 432.05
##  3rd Qu.:59432   3rd Qu.:20000   3rd Qu.:16.55   3rd Qu.: 567.30
##  Max.   :79242   Max.   :40000   Max.   :30.99   Max.   :1503.85
##
##    annual_inc           dti             open_acc        pub_rec
##  Min.   :    600   Min.   :   0.00   Min.   : 1.00   Min.   : 0.0000
##  1st Qu.:  45000   1st Qu.:  11.30   1st Qu.: 8.00   1st Qu.: 0.0000
##  Median :  64000   Median :  16.93   Median :10.00   Median : 0.0000
##  Mean   :  74075   Mean   :  17.37   Mean   :11.29   Mean   : 0.1778
##  3rd Qu.:  90000   3rd Qu.:  22.96   3rd Qu.:14.00   3rd Qu.: 0.0000
##  Max.   :7446395   Max.   :1622.00   Max.   :55.00   Max.   :24.0000
##
##    revol_bal         revol_util       total_acc         mort_acc
##  Min.   :      0   Min.   :  0.00   Min.   :  3.00   Min.   : 0.000
##  1st Qu.:   6018   1st Qu.: 36.00   1st Qu.: 17.00   1st Qu.: 0.000
##  Median :  11189   Median : 55.00   Median : 24.00   Median : 1.000
##  Mean   :  15942   Mean   : 53.93   Mean   : 25.41   Mean   : 1.819
##  3rd Qu.:  19676   3rd Qu.: 73.00   3rd Qu.: 32.00   3rd Qu.: 3.000
##  Max.   :1743266   Max.   :152.50   Max.   :151.00   Max.   :34.000
##                    NA's   :56                        NA's   :7485
##  pub_rec_bankruptcies  delinq_2yrs      fico_range_low  fico_range_high
##  Min.   :0.0000       Min.   : 0.0000   Min.   :660.0   Min.   :664.0
##  1st Qu.:0.0000       1st Qu.: 0.0000   1st Qu.:670.0   1st Qu.:674.0
##  Median :0.0000       Median : 0.0000   Median :690.0   Median :694.0
##  Mean   :0.1222       Mean   : 0.2825   Mean   :696.5   Mean   :700.5
##  3rd Qu.:0.0000       3rd Qu.: 0.0000   3rd Qu.:710.0   3rd Qu.:714.0
##  Max.   :7.0000       Max.   :18.0000   Max.   :845.0   Max.   :850.0
##  NA's   :104
##  inq_last_6mths  mths_since_last_delinq acc_now_delinq
##  Min.   :0.000   Min.   :  0.00         Min.   :0.000000
##  1st Qu.:0.000   1st Qu.: 16.00         1st Qu.:0.000000
##  Median :0.000   Median : 32.00         Median :0.000000
##  Mean   :0.781   Mean   : 34.73         Mean   :0.004253
##  3rd Qu.:1.000   3rd Qu.: 51.00         3rd Qu.:0.000000
##  Max.   :8.000   Max.   :152.00         Max.   :3.000000
##                  NA's   :42134
```

Plot for Missing Values

In this stage, we are focusing on the missing values to decide which predictors need imputation strategy or dropping the variables if its less than 10%.

mths_since_last_delinq has 53.17%. revol_util = 56 missing values mort_acc = 7485 missing values pub__rec_bankruptcies = 104 missing values mths_since_last_delinq = 42134 missing values

```
plot_missing(df)
```



Look at structure and ensure these are all factor variables

```
str(bank_factor)
```

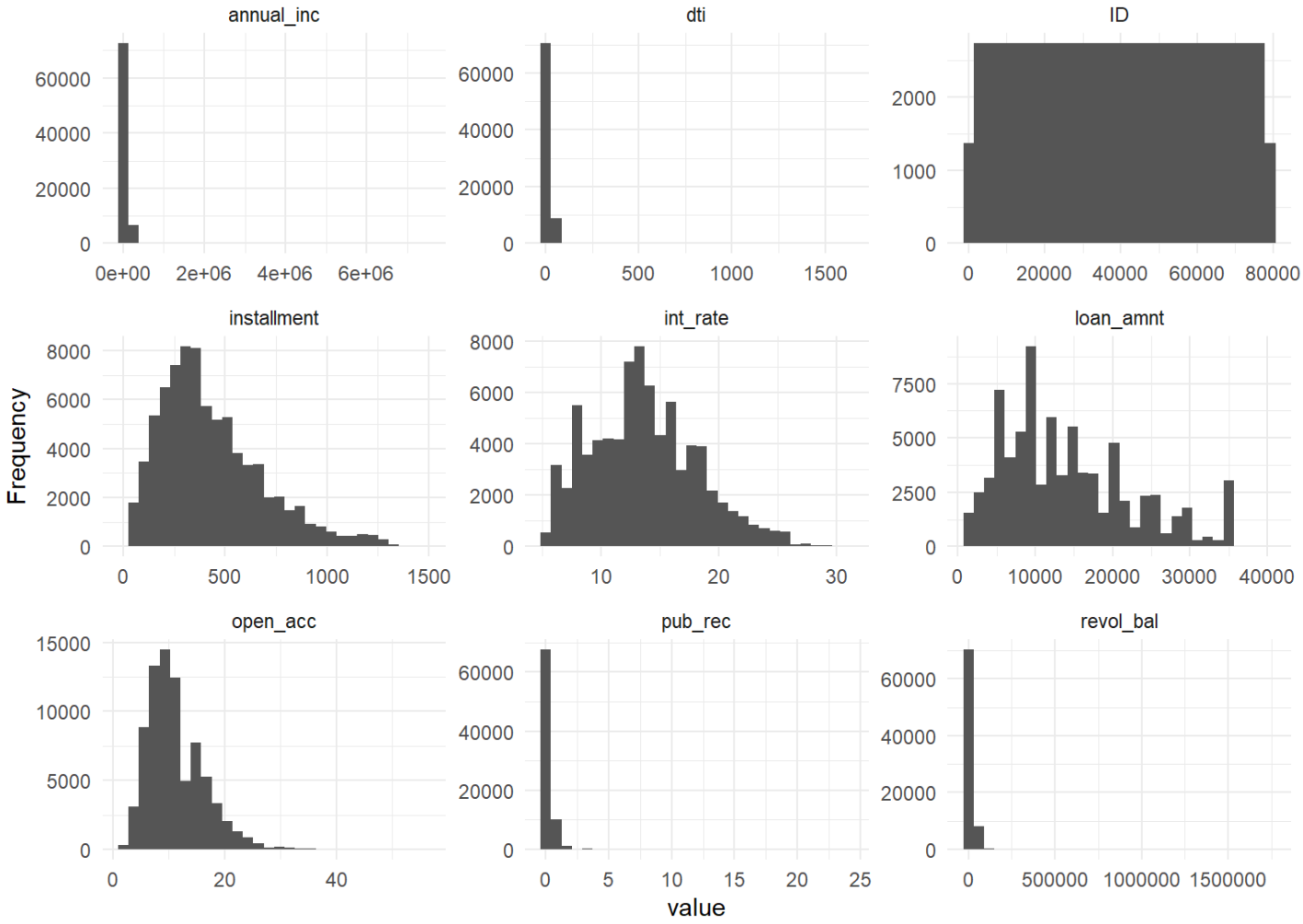
```
## 'data.frame':    79243 obs. of  17 variables:
##  $ term                : Factor w/ 2 levels " 36 months"," 60 months": 1 1 1 1 2 2 2 1 1 2 ...
##  $ grade               : Factor w/ 7 levels "A","B","C","D",..: 3 3 2 5 2 3 6 3 2 4 ...
##  $ sub_grade           : Factor w/ 35 levels "A1","A2","A3",..: 14 11 8 21 10 11 28 12 9 17 ...
##  $ emp_title           : Factor w/ 43586 levels "","  Credit rev specialist",..: 28879 10774 371
15 32204 38448 18071 24913 2823 38743 37148 ...
##  $ emp_length          : Factor w/ 12 levels "","< 1 year",..: 4 5 4 4 4 4 11 2 5 8 ...
##  $ home_ownership      : Factor w/ 6 levels "ANY","MORTGAGE",..: 6 2 6 2 6 6 2 6 2 6 ...
##  $ verification_status : Factor w/ 3 levels "Not Verified",..: 3 2 1 2 2 2 3 2 3 2 ...
##  $ issue_d             : Factor w/ 114 levels "Apr-2008","Apr-2009",..: 93 54 25 91 64 92 8 8 82
42 ...
##  $ purpose             : Factor w/ 14 levels "car","credit_card",..: 3 3 3 3 2 2 2 3 3 10 ...
##  $ title               : Factor w/ 12237 levels "","'08 & '09 Roth IRA Investments",..: 3870 387
0 8288 3870 3139 3139 3139 3870 3870 1893 ...
##  $ earliest_cr_line    : Factor w/ 618 levels "Apr-1961","Apr-1962",..: 552 245 241 293 185 552
450 549 541 547 ...
##  $ initial_list_status : Factor w/ 2 levels "f","w": 2 1 1 1 2 2 1 2 1 1 ...
##  $ application_type    : Factor w/ 3 levels "DIRECT_PAY","INDIVIDUAL",..: 2 2 2 2 2 2 2 2 2 2
...
##  $ address             : Factor w/ 79108 levels "000 Adrian Cliffs\nRandyton, LA 22690",..: 4163
9 65134 66277 27304 57318 77233 25031 16448 28914 66627 ...
##  $ last_credit_pull_d  : Factor w/ 131 levels "","Apr-2009",..: 42 64 20 21 116 118 56 120 118 8
9 ...
##  $ hardship_flag       : Factor w/ 1 level "N": 1 1 1 1 1 1 1 1 1 1 ...
##  $ debt_settlement_flag: Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 1 2 1 1 ...
```

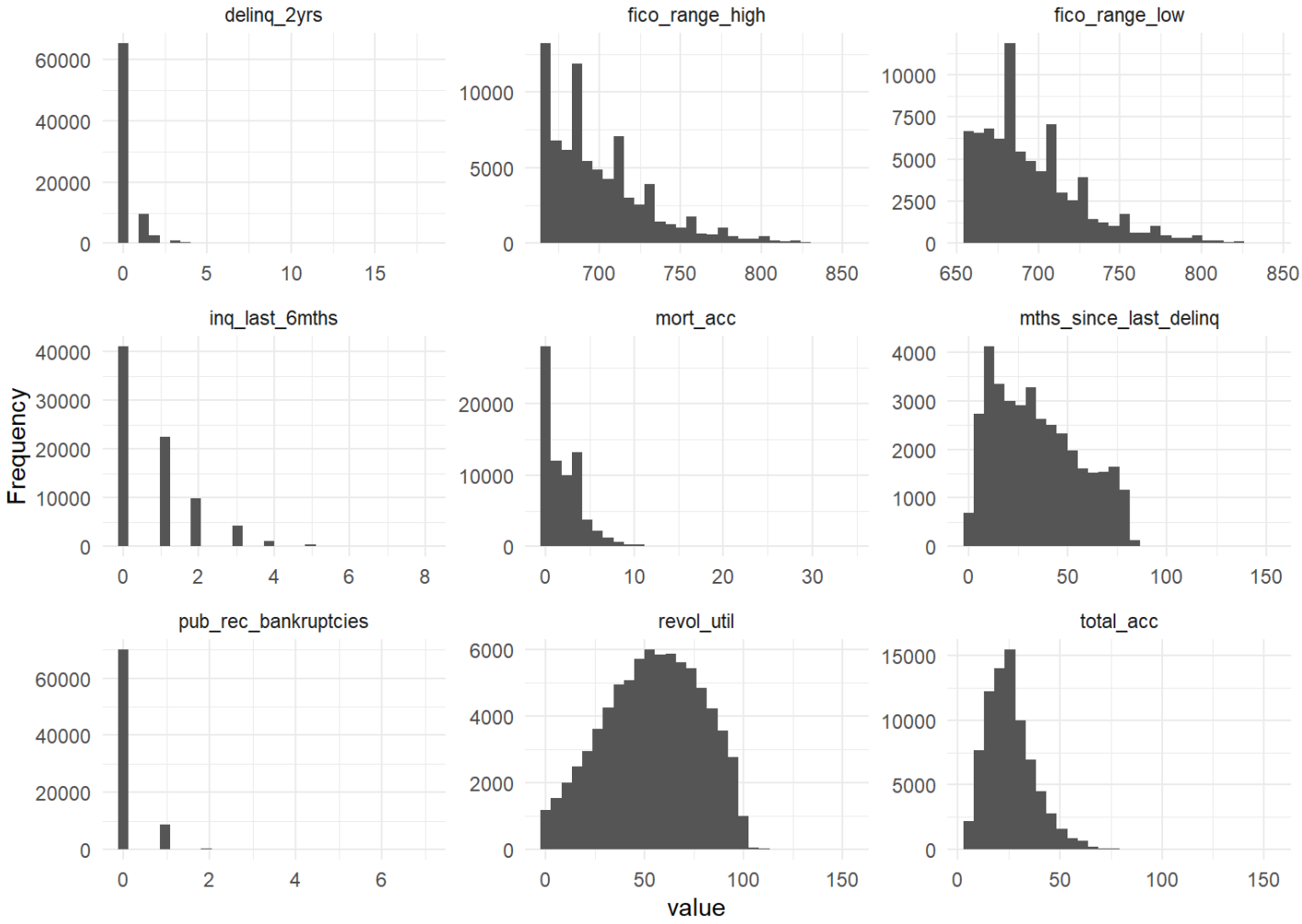
Data visualization

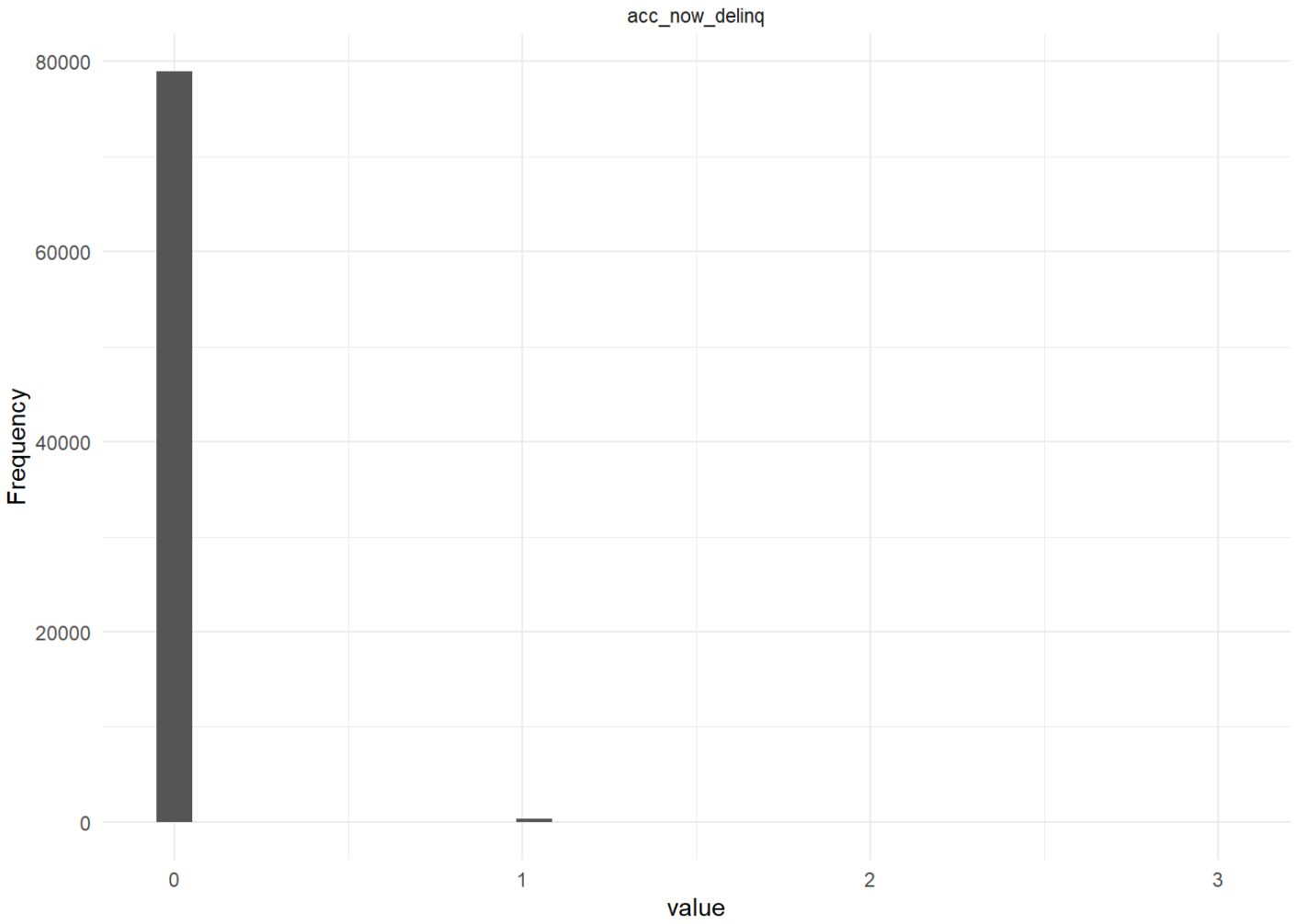
Histograms for numeric variables

These variables can easily be seen for their skew, symmetry, and range. If the range is large from the models, like pub_rec, revol_bal, dti, and annual_inc, this indicates outliers in the data.

```
DataExplorer::plot_histogram(bank_numeric,
                             ggtheme=theme_minimal(),
                             ncol=3,
                             nrow=3)
```


annual_inc
dti
ID
installment
int_rate
loan_amnt
open_acc
pub_rec
revol_bal
Frequency
value

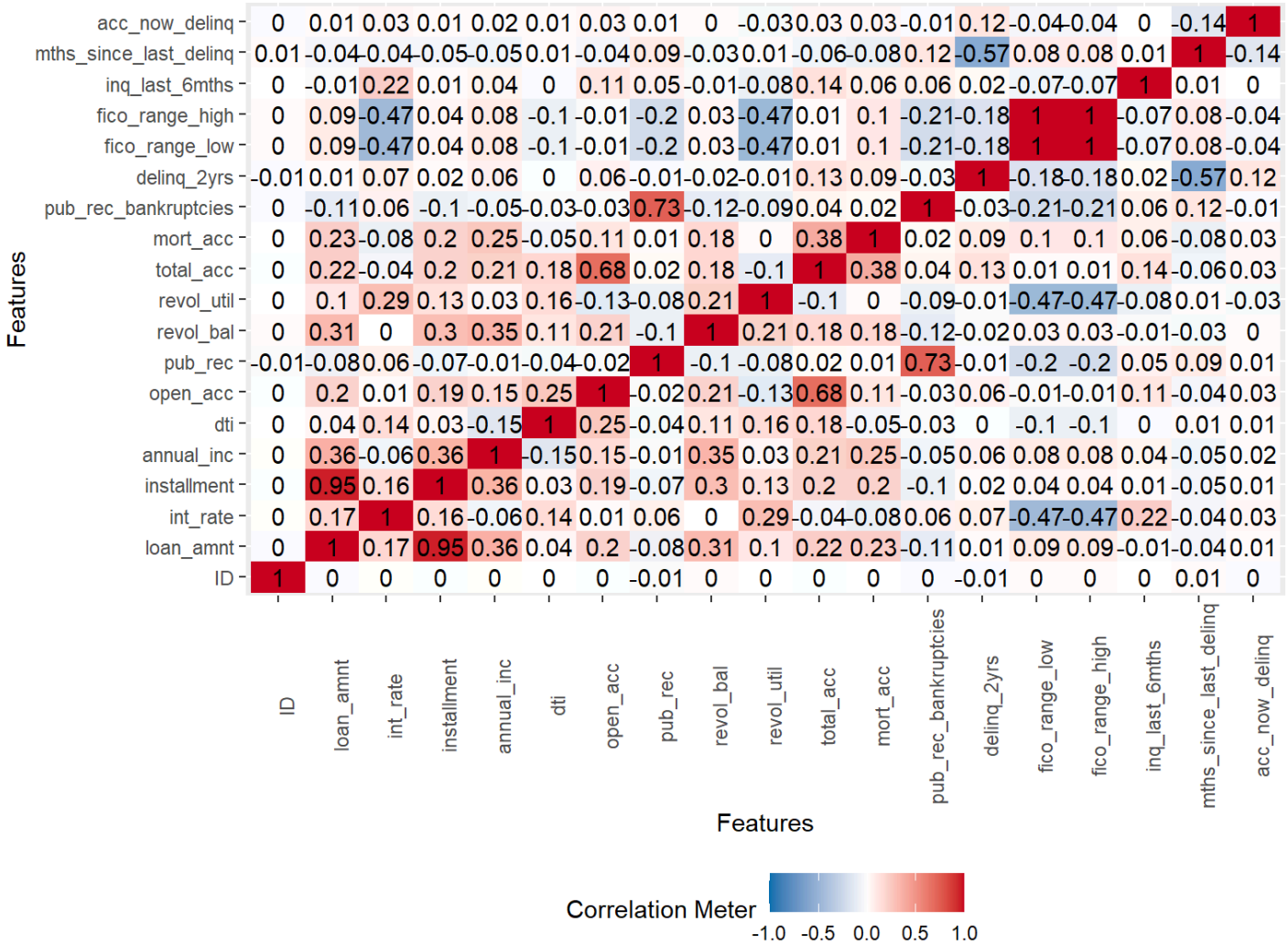

delinq_2yrs
fico_range_high
fico_range_low
inq_last_6mths
mort_acc
mths_since_last_delinq
pub_rec_bankruptcies
revol_util
total_acc
Frequency
value



Create a heat map to check correlation between variables

Based on these results, we can confirm outliers in the data and deal with them accordingly. Also, we see that these numeric variables can have correlation to one defaulting on a loan. Lets look at a heat map to finalize some of the cleaning techniques we will perform

```
DataExplorer::plot_correlation(bank_numeric,
                               type="continuous",
                               cor_args = list("use" = "pairwise.complete.obs")
                               )
```



Total_acc and open_acc has correlated metrics with an R^2 of .68, while pub_rec and pub_rec_bankrupcies have correlated features with an R^2 of 0.68. Installment has an extremely high correlation to loan amount, showing we can remove this variable.

Variables we are Removing

acc_now_delinq: We are removing this variable because most of the values are zero and it wouldn't give us any special indication of loan default. mths_since_last_delinq: We will remove this variable because there are too many null values(53.22%).Most people probably don't have any delinquency, so this variable can be disregarded. delinq_2yrs: Most of the values are zero. but we can also just get rid of it. installment: We are removing this variable because of its high correlation (0.95) to loan amount, meaning they provide the same information.

```
# Remove these variables
df<-df%>%
  select(-installment, -acc_now_delinq, -mths_since_last_delinq, -delinq_2yrs)
dim(df)
```

```
## [1] 79243    32
```

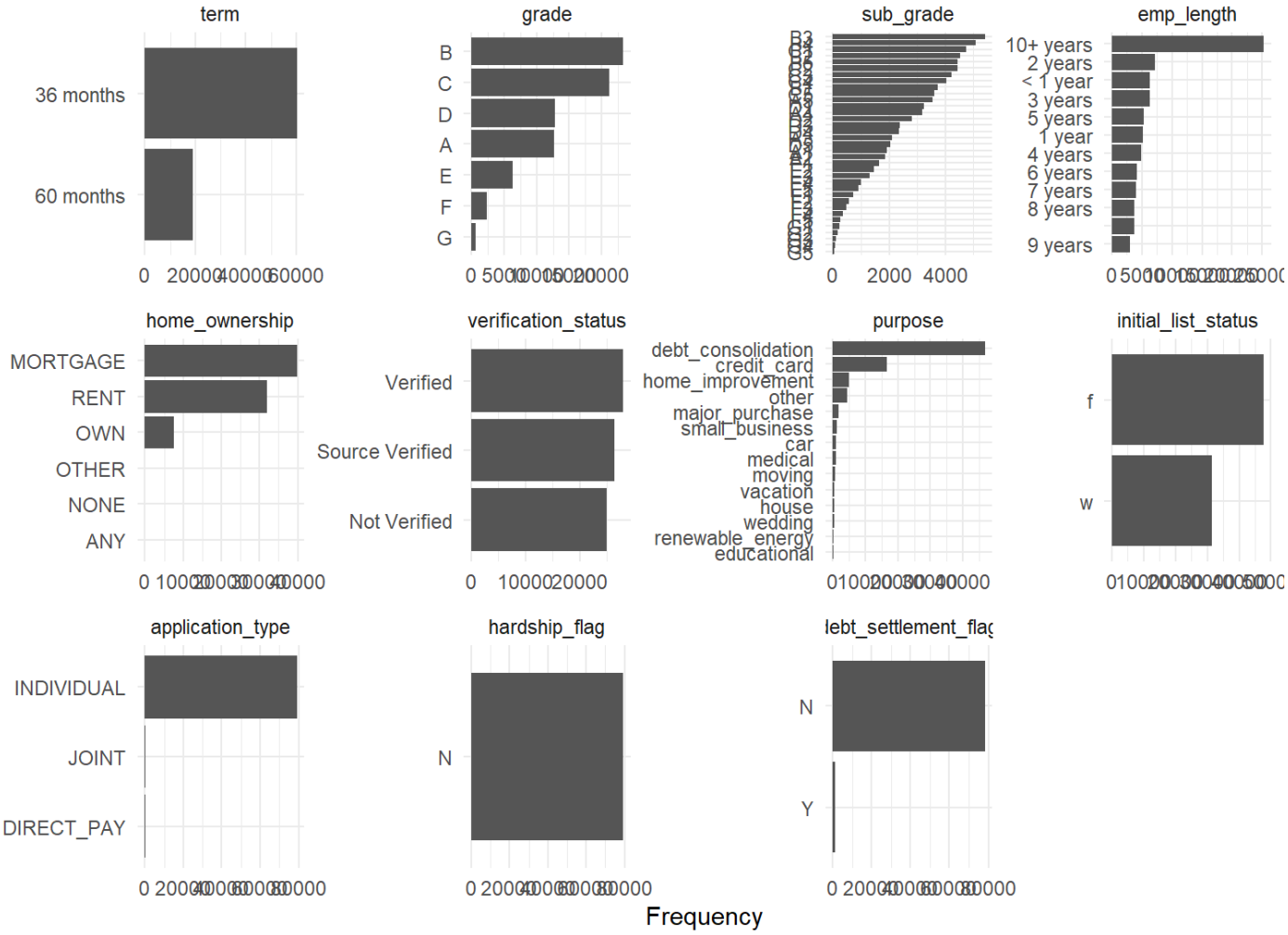
We now have only 32 variables in our original data frame. (36-4=32). None of these variables show major distribution between the 'yes' and 'no decision of a loan default based on the boxplot and amount of null values.

Cateogrical Variables:

Now, we will look at the factor variables and decide where to make changes based on the prediction power and levels. We will also recode and lump variables together if there are too many levels to the data or remove when there's too many levels in itself.

Look at structure and levels for categorical variables

```
DataExplorer::plot_bar(bank_factor,
                       ggtheme = theme_minimal(),
                       ncol=4,
                       nrow=5)
```



There are 12 variables shown here. Notice how some variables are not included in the visual, such as emp_title, issue_d, earliest_cr_line, address, and last_credit_pull_d. These variables all have a significant amount of levels, meaning they could either be manipulated to a numeric variable or be removed because there would way too many dummy variables. Even some of the bar plots have too many levels that will need to be analyzed further later in our analysis.

Now, we will identify what we are doing with each variable. For the categorical variables we are keeping, we will specify changing the leels or leaving the data as it is.

Keep the Same

Term: 2 levels, we will keep the same Grade: 7 levels, keep the same Verification_status: 2 levels, Keep the same Initial_list_status: 2 levels, Keep the same Debt_settlement_flag: a “yes” decision has a high correlation to “yes” to loan_default, keep the same

Variables we are Removing

Emp_title: Remove (too many levels) Issue_d: Remove (No significance to loan_default_ Title: Remove (Too many levels) Application_type: Remove (Vast majority of values in “individual” value) Hardship_flag: Remove (only 1 level) Earliest_cr_line: Dates do not seem significant in predicting loan default Last_credit_pull_d: Dates do not seem significant in predicting loan default Address (zip code): Discriminatory variable, not significant in explaining loan default sub_grade: Has the same properties as grade, not needed in our model

```
# Remove variables that are not seen as needed
df<-df%>%
  select(-emp_title, -issue_d, -title, -application_type, -hardship_flag, -earliest_cr_line, -last_
credit_pull_d, -address, -sub_grade)
dim(df)
```

```
## [1] 79243    23
```

Handling Missing Data

Earlier, we identified some of the missing values in different variables. Although we removed the mnths_since_last_delinq variable, we still have three variables that are missing data:

```
#Checking missing Values
colSums(is.na(df))
```

```
##                   ID            loan_amnt                 term
##                    0                    0                    0
##             int_rate                grade           emp_length
##                    0                    0                    0
##       home_ownership           annual_inc  verification_status
##                    0                    0                    0
##              purpose                  dti             open_acc
##                    0                    0                    0
##              pub_rec            revol_bal           revol_util
##                    0                    0                   56
##            total_acc  initial_list_status             mort_acc
##                    0                    0                 7485
## pub_rec_bankruptcies       fico_range_low      fico_range_high
##                  104                    0                    0
##       inq_last_6mths debt_settlement_flag
##                    0                    0
```

Outlier/Anomaly Detection

Throughout our data cleaning process, we have identified numerous variables with outliers. Below, we will decide what to do with these variables. In deciding what to do, we must ask ourselves: Are these outliers realistic in context of the real world?

```
#DEC
df$fico_score <- (df$fico_range_low + df$fico_range_high) / 2
df <- df %>% select(-fico_range_low, -fico_range_high)

#cap <- quantile(df$annual_inc, 0.99)
#df$annual_inc[df$annual_inc > cap] <- cap


# Median imputation on the mort_acc variable

df$mort_acc[is.na(df$mort_acc)] <- 1
#df$mort_acc[is.na(df$mort_acc)] <- median(df$mort_acc, na.rm = TRUE)
sum(is.na(df$mort_acc))
```

```
## [1] 0
```

```
#Median imputation to the pub_rec_bankrupcies variable
df$pub_rec_bankruptcies[is.na(df$pub_rec_bankruptcies)] <- 0
sum(is.na(df$pub_rec_bankruptcies))
```

```
## [1] 0
```

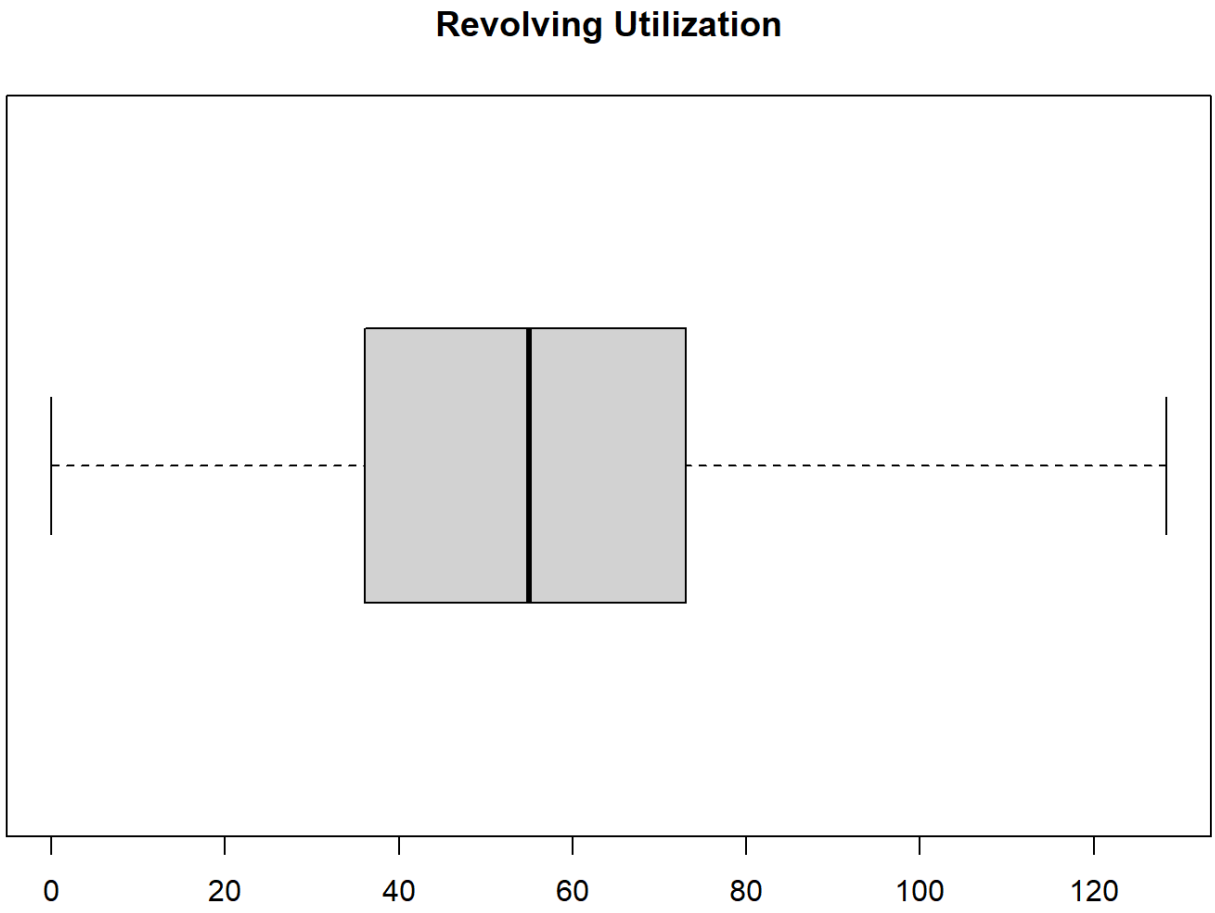
Here, we are making the strategy of using median imputation in revol_util for making the na's are effective in prediction. Because we need all the observations in predictions of loan_default.

```
# Removing rows where revol_util is NA
#df <- df[!is.na(df$revol_util), ]
df$revol_util[is.na(df$revol_util)] <- train_median
#nrow(df)
```

```
# Capping the revol_util variable
Q1 <- quantile(df$revol_util, 0.25, na.rm = TRUE)
Q3 <- quantile(df$revol_util, 0.75, na.rm = TRUE)
IQR <- Q3 - Q1
upper_bound <- Q3 + 1.5 * IQR
df$revol_util[df$revol_util > upper_bound] <- upper_bound
summary(df$revol_util)
```

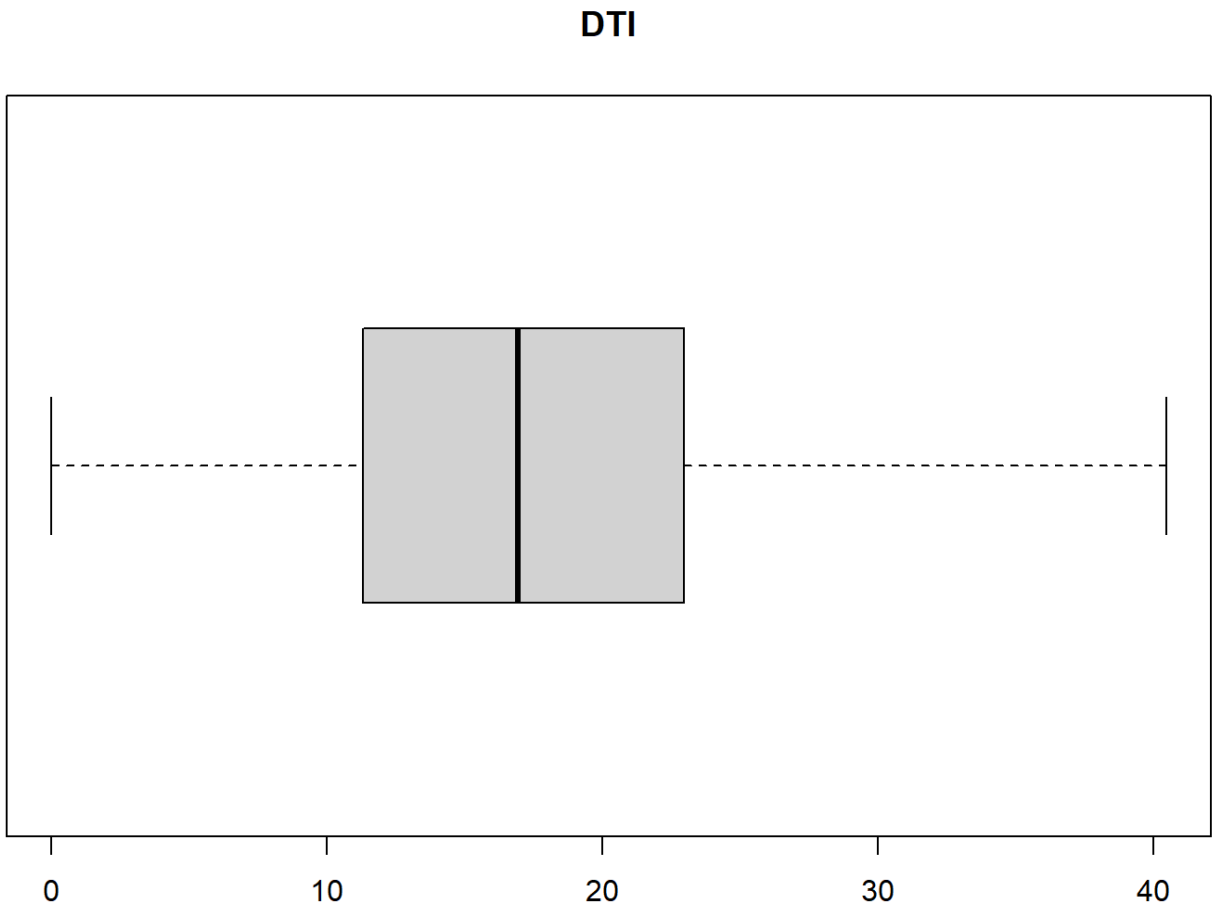
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   36.10   55.00   53.93   73.00  128.35
```

```
boxplot(df$revol_util, main = "Revolving Utilization", horizontal = TRUE)
```



```
# Capping the dti outlier
Q1 <- quantile(df$dti, 0.25, na.rm = TRUE)
Q3 <- quantile(df$dti, 0.75, na.rm = TRUE)
IQR <- Q3 - Q1

upper_bound <- Q3 + 1.5 * IQR
df$dti[df$dti > upper_bound] <- upper_bound
boxplot(df$dti, main = "DTI", horizontal = TRUE)
```




```
#Removing the "other" and "none" values
#df <- df[!df$home_ownership %in% c("OTHER", "NONE"), ]
df$home_ownership[df$home_ownership %in% c("OTHER","NONE")] <- "RENT"
df$home_ownership <- droplevels(df$home_ownership)
table(df$home_ownership)
```

```
##
##      ANY MORTGAGE      OWN     RENT
##         1    39751     7619    31872
```

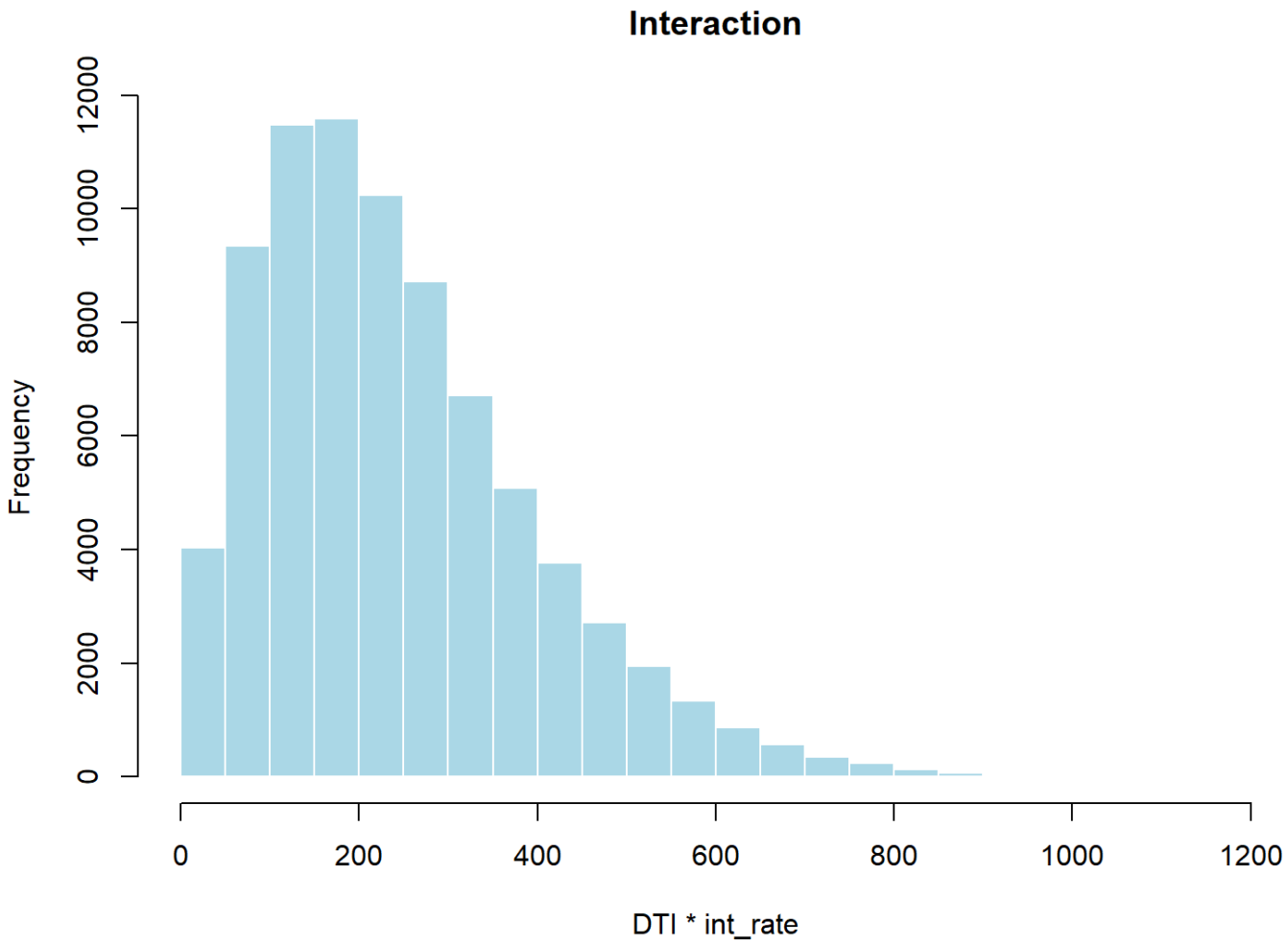
```
#Reduce dimensions in the purpose variable
df$purpose <- fct_other(df$purpose, keep = c("debt_consolidation", "credit_card", "home_improvemen
t"), other_level = "other")
table(df$purpose)
```

```
##
##        credit_card debt_consolidation   home_improvement              other
##              16618              46765               4848              11012
```

```
# Reducing dimensions on emp_length
df <- df %>%
  mutate(emp_length = case_when(
    emp_length %in% c("< 1 year", "1 year", "2 years", "3 years") ~ "0-3 years",
    emp_length %in% c("4 years", "5 years", "6 years") ~ "4-6 years",
    emp_length %in% c("7 years", "8 years", "9 years") ~ "7-9 years",
    emp_length == "10+ years" ~ "10+ years",
    TRUE ~ "Unknown"
  ))
df$emp_length <- factor(df$emp_length,
                                levels = c("0-3 years", "4-6 years", "7-9 years", "10+ years","Unkn
own"),
                                ordered = TRUE)
#df <- df %>%
#  filter(!is.na(emp_length) & emp_length != "Unknown")
```

```
# Dummy encoding all remaining categorical variables
df <- fastDummies::dummy_cols(df,
                              select_columns = c(
                                              "term",
                                              "home_ownership",
                                              "verification_status",
                                              "grade",
                                              "purpose",
                                              "initial_list_status",
                                              "debt_settlement_flag",
                                              "emp_length"),
                              remove_first_dummy = TRUE,      # avoids dummy variable trap
                              remove_selected_columns = TRUE) # removes original factor columns
```

```
# dti * int_rate interaction term
df$interaction_dti_interest <- df$dti * df$int_rate
# Show correlation
hist(df$interaction_dti_interest,
     main = "Interaction",
     xlab = "DTI * int_rate",
     col = "lightblue",
     border = "white")
```



Conclusion: Summary of our Cleaned Data

After performing all of these steps to create the best data frame to train a model, we have resulted in this dataset:

```
summary(df)
```

```
##        ID          loan_amnt        int_rate       annual_inc
##  Min.   :    0   Min.   :  950   Min.   : 5.32   Min.   :    600
##  1st Qu.:19811   1st Qu.: 8000   1st Qu.:10.49   1st Qu.:  45000
##  Median :39621   Median :12000   Median :13.33   Median :  64000
##  Mean   :39621   Mean   :14124   Mean   :13.68   Mean   :  74075
##  3rd Qu.:59432   3rd Qu.:20000   3rd Qu.:16.55   3rd Qu.:  90000
##  Max.   :79242   Max.   :40000   Max.   :30.99   Max.   :7446395
##       dti           open_acc        pub_rec          revol_bal
##  Min.   : 0.00   Min.   : 1.00   Min.   : 0.0000   Min.   :      0
##  1st Qu.:11.30   1st Qu.: 8.00   1st Qu.: 0.0000   1st Qu.:   6018
##  Median :16.93   Median :10.00   Median : 0.0000   Median :  11189
##  Mean   :17.34   Mean   :11.29   Mean   : 0.1778   Mean   :  15942
##  3rd Qu.:22.96   3rd Qu.:14.00   3rd Qu.: 0.0000   3rd Qu.:  19676
##  Max.   :40.45   Max.   :55.00   Max.   :24.0000   Max.   :1743266
##    revol_util       total_acc         mort_acc      pub_rec_bankruptcies
##  Min.   :  0.00   Min.   :  3.00   Min.   : 0.000   Min.   :0.000
##  1st Qu.: 36.10   1st Qu.: 17.00   1st Qu.: 0.000   1st Qu.:0.000
##  Median : 55.00   Median : 24.00   Median : 1.000   Median :0.000
##  Mean   : 53.93   Mean   : 25.41   Mean   : 1.741   Mean   :0.122
##  3rd Qu.: 73.00   3rd Qu.: 32.00   3rd Qu.: 3.000   3rd Qu.:0.000
##  Max.   :128.35   Max.   :151.00   Max.   :34.000   Max.   :7.000
##  inq_last_6mths    fico_score    term_ 60 months  home_ownership_MORTGAGE
##  Min.   :0.000   Min.   :662.0   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:672.0   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :692.0   Median :0.0000   Median :1.0000
##  Mean   :0.781   Mean   :698.5   Mean   :0.2396   Mean   :0.5016
##  3rd Qu.:1.000   3rd Qu.:712.0   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :8.000   Max.   :847.5   Max.   :1.0000   Max.   :1.0000
##  home_ownership_OWN home_ownership_RENT verification_status_Source Verified
##  Min.   :0.00000    Min.   :0.0000      Min.   :0.0000
##  1st Qu.:0.00000    1st Qu.:0.0000      1st Qu.:0.0000
##  Median :0.00000    Median :0.0000      Median :0.0000
##  Mean   :0.09615    Mean   :0.4022      Mean   :0.3324
##  3rd Qu.:0.00000    3rd Qu.:1.0000      3rd Qu.:1.0000
##  Max.   :1.00000    Max.   :1.0000      Max.   :1.0000
##  verification_status_Verified    grade_B          grade_C
##  Min.   :0.0000               Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000               1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000               Median :0.0000   Median :0.0000
##  Mean   :0.3526               Mean   :0.2935   Mean   :0.2662
##  3rd Qu.:1.0000               3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000               Max.   :1.0000   Max.   :1.0000
##     grade_D          grade_E           grade_F           grade_G
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.000000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000000
##  Median :0.0000   Median :0.00000   Median :0.00000   Median :0.000000
##  Mean   :0.1622   Mean   :0.07994   Mean   :0.03035   Mean   :0.008228
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.000000
##  purpose_debt_consolidation purpose_home_improvement purpose_other
##  Min.   :0.0000             Min.   :0.00000          Min.   :0.000
##  1st Qu.:0.0000             1st Qu.:0.00000          1st Qu.:0.000
##  Median :1.0000             Median :0.00000          Median :0.000
```

```
##  Mean   :0.5901            Mean   :0.06118          Mean   :0.139
##  3rd Qu.:1.0000            3rd Qu.:0.00000          3rd Qu.:0.000
##  Max.   :1.0000            Max.   :1.00000          Max.   :1.000
##  initial_list_status_w debt_settlement_flag_Y emp_length_4-6 years
##  Min.   :0.0000        Min.   :0.00000        Min.   :0.0000
##  1st Qu.:0.0000        1st Qu.:0.00000        1st Qu.:0.0000
##  Median :0.0000        Median :0.00000        Median :0.0000
##  Mean   :0.3976        Mean   :0.01262        Mean   :0.1806
##  3rd Qu.:1.0000        3rd Qu.:0.00000        3rd Qu.:0.0000
##  Max.   :1.0000        Max.   :1.00000        Max.   :1.0000
##  emp_length_7-9 years emp_length_10+ years emp_length_Unknown
##  Min.   :0.000        Min.   :0.0000       Min.   :0.00000
##  1st Qu.:0.000        1st Qu.:0.0000       1st Qu.:0.00000
##  Median :0.000        Median :0.0000       Median :0.00000
##  Mean   :0.137        Mean   :0.3189       Mean   :0.04659
##  3rd Qu.:0.000        3rd Qu.:1.0000       3rd Qu.:0.00000
##  Max.   :1.000        Max.   :1.0000       Max.   :1.00000
##  interaction_dti_interest
##  Min.   :   0.0
##  1st Qu.: 128.3
##  Median : 214.5
##  Mean   : 243.7
##  3rd Qu.: 328.4
##  Max.   :1172.6
```

Removing Extra Spaces in Predictor Variables and Save the File for Final Execution

```
names(df) <- gsub(" ", "_", names(df))
names(df) <- gsub("-", "_", names(df))
names(df) <- gsub("\\+", "plus", names(df))
saveRDS(df, "score_df.rds")
```

Evaluate the performance using Logit Model

Here, We are using the logit model for evaluating the performance on loan_default. We already performed the construction on an model in our previous project. so we are going to call the model and doing the predictions with our cleaned scoring data and setting the prob thershold as greater than 0.5. our levels are 0 and 1. Also, we made the new dataframe contains only id and loan default for production environment in the form of csv file.

```
logit_model = readRDS("dec2_xgb_model.rds")
# Predict using the trained model
#y_pred <- predict(logit_Model, newdata=df)
#y_pred <- ifelse(y_pred == "Yes", 1, 0)

pred.prob <- predict(logit_model,
                           newdata = df,
                           type = "prob")[,"Yes"]

y_pred <- factor(ifelse(pred.prob > 0.2246494,
                                      "1","0"),
                               levels=c("1","0"))



# Combine predictions with IDs
predictions <- data.frame(ID = df$ID, loan_status = y_pred)

# Write to CSV
write.csv(predictions, "group5AA_Black-Boopathy_submission_file.csv", row.names = FALSE)
```