G01- Exploratory Data Analysis

Vinayak B. Menon, Xinkai Zhou, Kamaldeep Kaur

7/28/2020

Problem Statement:
CredX is a leading credit card provider that gets thousands of credit card applications every year. But in the past few years, it has experienced an increase in credit loss. The CEO believes that the best strategy to mitigate credit risk is to "acquire the right customers".

The aim of this report is to explore the past data of the bank applicants, i.e their credit history and demographic data, and determine primary factors that has an influence on their default status.

Understanding the data:

We will be working on two datasets i.e the Demographic data and the Credit Bureau data.

```
#Accessing the two datasets
credit <- read.csv("Credit_Bureau.csv",stringsAsFactors = TRUE)
dem <- read.csv("demogs.csv",stringsAsFactors = TRUE)
```

We present below the summary statistics of the demographic data:

```
##  Application.ID           Age        Gender
##  Min.   :1.004e+05   Min.   :-3.00    :    2
##  1st Qu.:2.484e+08   1st Qu.:37.00   F:16837
##  Median :4.976e+08   Median :45.00   M:54456
##  Mean   :4.990e+08   Mean   :44.94
##  3rd Qu.:7.496e+08   3rd Qu.:53.00
##  Max.   :1.000e+09   Max.   :65.00
##
##  Marital.Status..at.the.time.of.application. No.of.dependents
##         :    6                                Min.   :1.000
##  Married:60730                                1st Qu.:2.000
##  Single :10559                                Median :3.000
##                                               Mean   :2.865
##                                               3rd Qu.:4.000
##                                               Max.   :5.000
##                                               NA's   :3
##      Income            Education       Profession
##  Min.   :-0.5               :  119          :   14
##  1st Qu.:14.0   Bachelor    :17697   SAL    :40439
##  Median :27.0   Masters     :23970   SE     :14307
##  Mean   :27.2   Others      :  121   SE_PROF:16535
##  3rd Qu.:40.0   Phd         : 4549
##  Max.   :60.0   Professional:24839
##
##            Type.of.residence No.of.months.in.current.residence
##                     :    8   Min.   :  6.00
##  Company provided   : 1630   1st Qu.:  6.00
```

```
##  Living with Parents: 1818   Median : 11.00
##  Others             :  199   Mean   : 34.56
##  Owned              :14243   3rd Qu.: 60.00
##  Rented             :53397   Max.   :126.00
##
##  No.of.months.in.current.company Performance.Tag
##  Min.   :  3.00                  Min.   :0.0000
##  1st Qu.: 16.00                  1st Qu.:0.0000
##  Median : 34.00                  Median :0.0000
##  Mean   : 33.96                  Mean   :0.0422
##  3rd Qu.: 51.00                  3rd Qu.:0.0000
##  Max.   :133.00                  Max.   :1.0000
##                                  NA's   :1425
```

Now we present the summary statistics of the credit bureau data:

```
##  Application.ID      No.of.times.90.DPD.or.worse.in.last.6.months
##  Min.   :1.004e+05   Min.   :0.0000
##  1st Qu.:2.484e+08   1st Qu.:0.0000
##  Median :4.976e+08   Median :0.0000
##  Mean   :4.990e+08   Mean   :0.2703
##  3rd Qu.:7.496e+08   3rd Qu.:0.0000
##  Max.   :1.000e+09   Max.   :3.0000
##
##  No.of.times.60.DPD.or.worse.in.last.6.months
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.4305
##  3rd Qu.:1.0000
##  Max.   :5.0000
##
##  No.of.times.30.DPD.or.worse.in.last.6.months
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.5772
##  3rd Qu.:1.0000
##  Max.   :7.0000
##
##  No.of.times.90.DPD.or.worse.in.last.12.months
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.4503
##  3rd Qu.:1.0000
##  Max.   :5.0000
##
##  No.of.times.60.DPD.or.worse.in.last.12.months
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.6555
##  3rd Qu.:1.0000
```

```
##  Max.   :7.0000
##
##  No.of.times.30.DPD.or.worse.in.last.12.months
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.8009
##  3rd Qu.:1.0000
##  Max.   :9.0000
##
##  Avgas.CC.Utilization.in.last.12.months
##  Min.   :  0.0
##  1st Qu.:  8.0
##  Median : 15.0
##  Mean   : 29.7
##  3rd Qu.: 46.0
##  Max.   :113.0
##  NA's   :1058
##  No.of.trades.opened.in.last.6.months
##  Min.   : 0.000
##  1st Qu.: 1.000
##  Median : 2.000
##  Mean   : 2.298
##  3rd Qu.: 3.000
##  Max.   :12.000
##  NA's   :1
##  No.of.trades.opened.in.last.12.months
##  Min.   : 0.000
##  1st Qu.: 2.000
##  Median : 5.000
##  Mean   : 5.827
##  3rd Qu.: 9.000
##  Max.   :28.000
##
##  No.of.PL.trades.opened.in.last.6.months
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :1.000
##  Mean   :1.207
##  3rd Qu.:2.000
##  Max.   :6.000
##
##  No.of.PL.trades.opened.in.last.12.months
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 2.000
##  Mean   : 2.397
##  3rd Qu.: 4.000
##  Max.   :12.000
##
##  No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 1.000
```

```
##  Mean   : 1.764
##  3rd Qu.: 3.000
##  Max.   :10.000
##
##  No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 3.000
##  Mean   : 3.535
##  3rd Qu.: 5.000
##  Max.   :20.000
##
##  Presence.of.open.home.loan Outstanding.Balance Total.No.of.Trades
##  Min.   :0.0000             Min.   :      0     Min.   : 0.000
##  1st Qu.:0.0000             1st Qu.: 211532     1st Qu.: 3.000
##  Median :0.0000             Median : 774992     Median : 6.000
##  Mean   :0.2564             Mean   :1249163     Mean   : 8.187
##  3rd Qu.:1.0000             3rd Qu.:2920796     3rd Qu.:10.000
##  Max.   :1.0000             Max.   :5218801     Max.   :44.000
##  NA's   :272                NA's   :272
##  Presence.of.open.auto.loan Performance.Tag
##  Min.   :0.00000            Min.   :0.0000
##  1st Qu.:0.00000            1st Qu.:0.0000
##  Median :0.00000            Median :0.0000
##  Mean   :0.08462            Mean   :0.0422
##  3rd Qu.:0.00000            3rd Qu.:0.0000
##  Max.   :1.00000            Max.   :1.0000
##                             NA's   :1425

## The following objects are masked from dem:
##
##     Application.ID, Performance.Tag
```

We can see that there are empty labels in a lot of the factor type data. We will replace this label with NA.

```
levels(dem$Gender)[levels(dem$Gender)==""]<-"NA"
levels(dem$Marital.Status..at.the.time.of.application.)[levels(dem$Marital.Status..at.the.time.of.appli
levels(dem$Education)[levels(dem$Education)==""]<-"NA"
levels(dem$Profession)[levels(dem$Profession)==""]<-"NA"
levels(dem$Type.of.residence)[levels(dem$Type.of.residence)==""]<-"NA"
```

We now print any duplicate entries based on their Application.ID, and proceed to remove any such entries.

```
#checking for duplicates
dem %>%
  group_by(Application.ID)%>%
  filter(n()>1)
```

```
## # A tibble: 6 x 12
## # Groups:   Application.ID [3]
##   Application.ID   Age Gender Marital.Status.~ No.of.dependents Income
##            <int> <int> <fct>  <fct>                       <int>  <dbl>
## 1      653287861    26 M      Married                         3   25
```

```
## 2      765011468    57 M      Single                          4    4.5
## 3      765011468    38 M      Married                         4    4.5
## 4      653287861    40 M      Married                         5   32
## 5      671989187    27 M      Married                         2   35
## 6      671989187    57 M      Married                         4    7
## # ... with 6 more variables: Education <fct>, Profession <fct>,
## #   Type.of.residence <fct>, No.of.months.in.current.residence <int>,
## #   No.of.months.in.current.company <int>, Performance.Tag <int>
```

```
credit %>%
  group_by(Application.ID)%>%
  filter(n()>1)
```

```
## # A tibble: 6 x 19
## # Groups:   Application.ID [3]
##   Application.ID No.of.times.90.~ No.of.times.60.~ No.of.times.30.~
##            <int>            <int>            <int>            <int>
## 1      653287861                0                0                0
## 2      765011468                0                0                0
## 3      765011468                0                0                0
## 4      653287861                1                1                1
## 5      671989187                1                2                3
## 6      671989187                0                1                2
## # ... with 15 more variables:
## #   No.of.times.90.DPD.or.worse.in.last.12.months <int>,
## #   No.of.times.60.DPD.or.worse.in.last.12.months <int>,
## #   No.of.times.30.DPD.or.worse.in.last.12.months <int>,
## #   Avgas.CC.Utilization.in.last.12.months <int>,
## #   No.of.trades.opened.in.last.6.months <int>,
## #   No.of.trades.opened.in.last.12.months <int>,
## #   No.of.PL.trades.opened.in.last.6.months <int>,
## #   No.of.PL.trades.opened.in.last.12.months <int>,
## #   No.of.Inquiries.in.last.6.months..excluding.home...auto.loans. <int>,
## #   No.of.Inquiries.in.last.12.months..excluding.home...auto.loans. <int>,
## #   Presence.of.open.home.loan <int>, Outstanding.Balance <int>,
## #   Total.No.of.Trades <int>, Presence.of.open.auto.loan <int>,
## #   Performance.Tag <int>
```

```
#selecting only unique ID's
dem <- dem %>%
  group_by(Application.ID)%>%
  filter(n()==1)

credit <-  credit %>%
  group_by(Application.ID)%>%
  filter(n()==1)
```

We proceed to merge the two datasets, as we can operate on it in one go from now.

```
#Mergimg the datasets
merged_data <- merge(dem,credit,by=c("Application.ID"))
#removing performance variable obtained from dem. Performance.Tag.y is from credit. Same result
merged_data <- merged_data[,-12]
```

It will be useful to check the clas of our independent and dependent variables, making sure that the dependent variable is not a factor, while the independent variables are either factors or numeric.

```
#checking if categorical independent variables are factors
#also checking if dependent categorical variable is integer (not factor)
#both of these are done to use the woe and IV functions effectively
#dem_class contains the type of data of each coloumns
merged_data_class <- data.frame(colnames(merged_data))
colnames(merged_data_class) <- "Variable"
for (i in 1:ncol(dem)) {
  merged_data_class$Class[i] <- class(merged_data[,i])
}
merged_data_class
```

```
##                                                        Variable   Class
## 1                                                Application.ID integer
## 2                                                           Age integer
## 3                                                        Gender  factor
## 4          Marital.Status..at.the.time.of.application.  factor
## 5                                            No.of.dependents integer
## 6                                                        Income numeric
## 7                                                     Education  factor
## 8                                                    Profession  factor
## 9                                                Type.of.residence  factor
## 10                                 No.of.months.in.current.residence integer
## 11                                   No.of.months.in.current.company integer
## 12                  No.of.times.90.DPD.or.worse.in.last.6.months integer
## 13                  No.of.times.60.DPD.or.worse.in.last.6.months integer
## 14                  No.of.times.30.DPD.or.worse.in.last.6.months integer
## 15                 No.of.times.90.DPD.or.worse.in.last.12.months integer
## 16                 No.of.times.60.DPD.or.worse.in.last.12.months integer
## 17                 No.of.times.30.DPD.or.worse.in.last.12.months integer
## 18                        Avgas.CC.Utilization.in.last.12.months integer
## 19                          No.of.trades.opened.in.last.6.months integer
## 20                         No.of.trades.opened.in.last.12.months integer
## 21                       No.of.PL.trades.opened.in.last.6.months integer
## 22                      No.of.PL.trades.opened.in.last.12.months integer
## 23  No.of.Inquiries.in.last.6.months..excluding.home...auto.loans. integer
## 24 No.of.Inquiries.in.last.12.months..excluding.home...auto.loans. integer
## 25                                    Presence.of.open.home.loan integer
## 26                                           Outstanding.Balance integer
## 27                                            Total.No.of.Trades integer
## 28                                    Presence.of.open.auto.loan integer
## 29                                             Performance.Tag.y integer
```

It is also important to remove any entries with no target entry. This step is required for the WOE calculations in the next section.

```
#checking for NA in Performance.Tag i.e dependent categorical variable
#missing values in dependent variable cannot be practically solved
#thus we resort to removing them from the dataset
merged_data$Performance.Tag.y %>%
  is.na()%>%
  sum()
```

```
## [1] 1425
```

```
merged_data<- merged_data %>%
  filter(!is.na(Performance.Tag.y))
```

Data Cleaning and Preparation:

As we can see from above, there are a number of unavailable entries (NA) and outlier values within the data. We will replace these values with the respective WOE values, indicating a relationship with the respective target value. For this, we make use of the *scorecard* package:

```
#computing IV and WOE
bins <- woebin(merged_data[,-1],"Performance.Tag.y")
```

```
## [INFO] creating woe binning ...
```

```
bins$Gender
```

```
##    variable    bin count count_distr  good  bad    badprob         woe
## 1:   Gender NA%,%F 16508   0.2362876 15790  718 0.04349406  0.03200280
## 2:   Gender      M 53356   0.7637124 51127 2229 0.04177600 -0.01009434
##          bin_iv     total_iv breaks is_special_values
## 1: 2.455781e-04 0.0003230384 NA%,%F             FALSE
## 2: 7.746033e-05 0.0003230384      M             FALSE
```

As an example, we have printed the resulting statistics for the Gender variable. The function has binned NA and $Female$ categories as one bin (most likely due to similar WOE values) and $Male$ as a seperate bin. The WOE coloumn shows their relationship score with the target variable $Performance.Tag$ along with the IV value that indicates the strength of this relationship. Note that the $total_iv$ is representative of the relationship strength for the whole variable.
We will now create the woe_data by replacing the merged_data dataset with the respective woe values. We also present the summary statistics of the woe_data:

```
#woe data
woe_data<-woebin_ply(merged_data[,-1],bins)
```

```
## [INFO] converting into woe values ...
```

```
#changing dependent coloumn
woe_data$Performance.Tag <- woe_data$Performance.Tag.y
#merging application id data
colnames(woe_data)[1]<-"Application.ID"
woe_data$Application.ID <- merged_data$Application.ID
#writing file
write.csv(woe_data,file="woe_data.csv")
#summary
summary(woe_data)
```

```
##  Application.ID         Age_woe           Gender_woe
##  Min.   :1.004e+05   Min.   :-0.12961   Min.   :-0.0100943
```

```
##  1st Qu.:2.486e+08   1st Qu.:-0.11199   1st Qu.:-0.0100943
##  Median :4.980e+08   Median : 0.02452   Median :-0.0100943
##  Mean   :4.992e+08   Mean   :-0.00314   Mean   :-0.0001473
##  3rd Qu.:7.499e+08   3rd Qu.: 0.07415   3rd Qu.:-0.0100943
##  Max.   :1.000e+09   Max.   : 0.11359   Max.   : 0.0320028
##  Marital.Status..at.the.time.of.application._woe No.of.dependents_woe
##  Min.   :-4.102e-03                              Min.   :-0.085534
##  1st Qu.:-4.102e-03                              1st Qu.:-0.025498
##  Median :-4.102e-03                              Median : 0.004010
##  Mean   :-4.376e-05                              Mean   :-0.001425
##  3rd Qu.:-4.102e-03                              3rd Qu.: 0.039704
##  Max.   : 2.338e-02                              Max.   : 2.013667
##    Income_woe       Education_woe        Profession_woe
##  Min.   :-0.38554   Min.   :-0.0295568   Min.   :-0.028375
##  1st Qu.:-0.19632   1st Qu.:-0.0179334   1st Qu.:-0.028375
##  Median : 0.06925   Median : 0.0136604   Median :-0.028375
##  Mean   :-0.01973   Mean   :-0.0001268   Mean   :-0.001008
##  3rd Qu.: 0.06925   3rd Qu.: 0.0136604   3rd Qu.:-0.013343
##  Max.   : 0.32547   Max.   : 0.0136604   Max.   : 0.091379
##  Type.of.residence_woe No.of.months.in.current.residence_woe
##  Min.   :-4.319e-03    Min.   :-0.30230
##  1st Qu.:-4.319e-03    1st Qu.:-0.30230
##  Median :-4.319e-03    Median : 0.03168
##  Mean   :-5.788e-05    Mean   :-0.04085
##  3rd Qu.: 4.104e-03    3rd Qu.: 0.03168
##  Max.   : 4.582e-02    Max.   : 0.50236
##  No.of.months.in.current.company_woe
##  Min.   :-0.39973
##  1st Qu.:-0.10016
##  Median :-0.05894
##  Mean   :-0.01399
##  3rd Qu.: 0.23259
##  Max.   : 0.23259
##  No.of.times.90.DPD.or.worse.in.last.6.months_woe
##  Min.   :-0.26069
##  1st Qu.:-0.26069
##  Median :-0.26069
##  Mean   :-0.06851
##  3rd Qu.:-0.26069
##  Max.   : 0.62248
##  No.of.times.60.DPD.or.worse.in.last.6.months_woe
##  Min.   :-0.33637
##  1st Qu.:-0.33637
##  Median :-0.33637
##  Mean   :-0.09043
##  3rd Qu.: 0.54135
##  Max.   : 0.74337
##  No.of.times.30.DPD.or.worse.in.last.6.months_woe
##  Min.   :-0.3868
##  1st Qu.:-0.3868
##  Median :-0.3868
##  Mean   :-0.1050
##  3rd Qu.: 0.4643
##  Max.   : 0.7429
```

```
##  No.of.times.90.DPD.or.worse.in.last.12.months_woe
##  Min.   :-0.35664
##  1st Qu.:-0.35664
##  Median :-0.35664
##  Mean   :-0.09311
##  3rd Qu.: 0.50878
##  Max.   : 0.72208
##  No.of.times.60.DPD.or.worse.in.last.12.months_woe
##  Min.   :-0.35192
##  1st Qu.:-0.35192
##  Median :-0.35192
##  Mean   :-0.08121
##  3rd Qu.: 0.21411
##  Max.   : 0.79562
##  No.of.times.30.DPD.or.worse.in.last.12.months_woe
##  Min.   :-0.37639
##  1st Qu.:-0.37639
##  Median :-0.37639
##  Mean   :-0.09326
##  3rd Qu.: 0.07100
##  Max.   : 0.79960
##  Avgas.CC.Utilization.in.last.12.months_woe
##  Min.   :-0.6812
##  1st Qu.:-0.6812
##  Median :-0.6812
##  Mean   :-0.1432
##  3rd Qu.: 0.5125
##  Max.   : 0.5125
##  No.of.trades.opened.in.last.6.months_woe
##  Min.   :-0.5435518
##  1st Qu.:-0.5435518
##  Median :-0.0006892
##  Mean   :-0.0852944
##  3rd Qu.: 0.3686856
##  Max.   : 3.1122791
##  No.of.trades.opened.in.last.12.months_woe
##  Min.   :-0.86509
##  1st Qu.:-0.86509
##  Median : 0.07524
##  Mean   :-0.13389
##  3rd Qu.: 0.39090
##  Max.   : 0.39090
##  No.of.PL.trades.opened.in.last.6.months_woe
##  Min.   :-0.6492
##  1st Qu.:-0.6492
##  Median : 0.1994
##  Mean   :-0.1054
##  3rd Qu.: 0.4006
##  Max.   : 0.4006
##  No.of.PL.trades.opened.in.last.12.months_woe
##  Min.   :-0.8938
##  1st Qu.:-0.8938
##  Median : 0.3706
##  Mean   :-0.1444
```

```
##  3rd Qu.: 0.3706
##  Max.   : 0.3706
##  No.of.Inquiries.in.last.6.months..excluding.home...auto.loans._woe
##  Min.   :-0.71823
##  1st Qu.:-0.71823
##  Median : 0.29292
##  Mean   :-0.09488
##  3rd Qu.: 0.29292
##  Max.   : 0.29292
##  No.of.Inquiries.in.last.12.months..excluding.home...auto.loans._woe
##  Min.   :-1.0675
##  1st Qu.:-1.0675
##  Median : 0.1166
##  Mean   :-0.1413
##  3rd Qu.: 0.3396
##  Max.   : 0.3396
##  Presence.of.open.home.loan_woe Outstanding.Balance_woe
##  Min.   :-0.373842              Min.   :-0.7735
##  1st Qu.:-0.236703              1st Qu.:-0.6965
##  Median : 0.073722              Median : 0.2645
##  Mean   :-0.008315              Mean   :-0.1178
##  3rd Qu.: 0.073722              3rd Qu.: 0.3857
##  Max.   : 0.073722              Max.   : 0.3857
##  Total.No.of.Trades_woe Presence.of.open.auto.loan_woe Performance.Tag
##  Min.   :-0.7956        Min.   :-0.138237              Min.   :0.00000
##  1st Qu.:-0.7956        1st Qu.: 0.011973              1st Qu.:0.00000
##  Median :-0.1004        Median : 0.011973              Median :0.00000
##  Mean   :-0.1145        Mean   :-0.000775              Mean   :0.04218
##  3rd Qu.: 0.5290        3rd Qu.: 0.011973              3rd Qu.:0.00000
##  Max.   : 0.5290        Max.   : 0.011973              Max.   :1.00000
```

Now we print the IV values of the independent variables in descending order, to remove those variables with a weak relationship with the dependent variable:

```
#obtaining all the IV values from bins
IV_temp <- round(c(bins$Age$total_iv[1],bins$Gender$total_iv[1],
                   bins$Marital.Status..at.the.time.of.application.$total_iv[1],
      bins$No.of.dependents$total_iv[1],bins$Income$total_iv[1],bins$Education$total_iv[1],
      bins$Profession$total_iv[1],bins$Type.of.residence$total_iv[1],bins$No.of.months.in.current.reside
              bins$No.of.months.in.current.company$total_iv[1],bins$No.of.times.90.DPD.or.worse.in.last
              bins$No.of.times.30.DPD.or.worse.in.last.6.months$total_iv[1],bins$No.of.times.90.DPD.or.
              bins$No.of.times.30.DPD.or.worse.in.last.12.months$total_iv[1],bins$Avgas.CC.Utilization.
              bins$No.of.trades.opened.in.last.12.months$total_iv[1],bins$No.of.PL.trades.opened.in.las
              bins$No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.$total_iv[1],bins$No.o
              bins$Outstanding.Balance$total_iv[1],bins$Total.No.of.Trades$total_iv[1],bins$Presence.of

#creating a table of IV values
IV<- data.frame("Variable"=colnames(merged_data)[-c(1,ncol(merged_data))],
                "IV"=IV_temp)
IV$Variable<-as.character(IV$Variable)
IV_desc <- IV[order(-IV$IV),]
IV_cut <- subset(IV_desc,IV_desc$IV>0.1)
rownames(IV_cut)<-1:nrow(IV_cut)
print(IV_desc)
```

```
##                                                          Variable   IV
## 17                          Avgas.CC.Utilization.in.last.12.months 0.30
## 21                        No.of.PL.trades.opened.in.last.12.months 0.29
## 19                           No.of.trades.opened.in.last.12.months 0.27
## 23 No.of.Inquiries.in.last.12.months..excluding.home...auto.loans. 0.27
## 13                     No.of.times.30.DPD.or.worse.in.last.6.months 0.24
## 25                                              Outstanding.Balance 0.24
## 26                                               Total.No.of.Trades 0.24
## 16                    No.of.times.30.DPD.or.worse.in.last.12.months 0.22
## 20                          No.of.PL.trades.opened.in.last.6.months 0.22
## 12                     No.of.times.60.DPD.or.worse.in.last.6.months 0.21
## 14                    No.of.times.90.DPD.or.worse.in.last.12.months 0.21
## 15                    No.of.times.60.DPD.or.worse.in.last.12.months 0.19
## 22  No.of.Inquiries.in.last.6.months..excluding.home...auto.loans. 0.19
## 18                             No.of.trades.opened.in.last.6.months 0.18
## 11                     No.of.times.90.DPD.or.worse.in.last.6.months 0.16
## 9                                 No.of.months.in.current.residence 0.09
## 5                                                           Income 0.04
## 10                                No.of.months.in.current.company 0.03
## 24                                     Presence.of.open.home.loan 0.02
## 1                                                             Age 0.01
## 2                                                          Gender 0.00
## 3                       Marital.Status..at.the.time.of.application. 0.00
## 4                                              No.of.dependents 0.00
## 6                                                     Education 0.00
## 7                                                    Profession 0.00
## 8                                             Type.of.residence 0.00
## 27                                    Presence.of.open.auto.loan 0.00
```

The independent variables that hold a medium level of predictive power to the target variable, $Performance.Tag$ is:

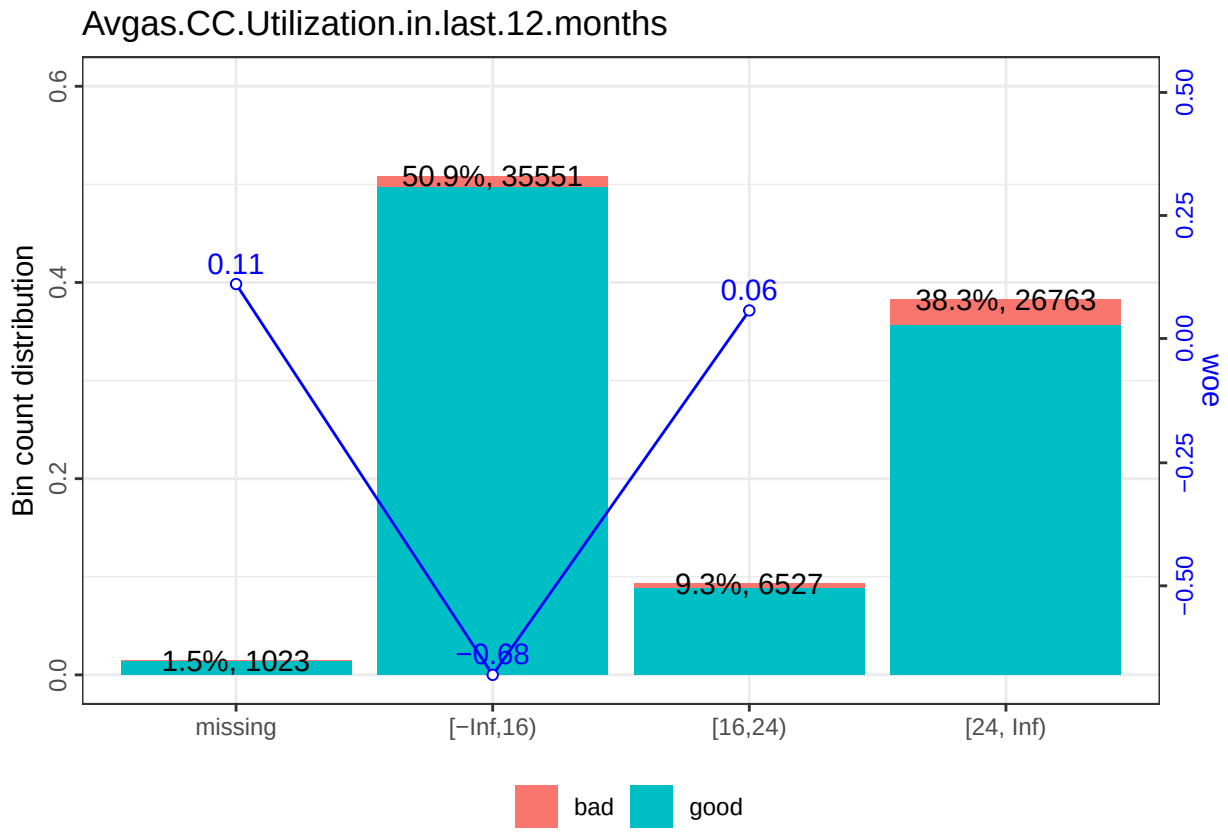
```
print(IV_cut)
```

```
##                                                          Variable   IV
## 1                           Avgas.CC.Utilization.in.last.12.months 0.30
## 2                         No.of.PL.trades.opened.in.last.12.months 0.29
## 3                            No.of.trades.opened.in.last.12.months 0.27
## 4  No.of.Inquiries.in.last.12.months..excluding.home...auto.loans. 0.27
## 5                      No.of.times.30.DPD.or.worse.in.last.6.months 0.24
## 6                                               Outstanding.Balance 0.24
## 7                                                Total.No.of.Trades 0.24
## 8                     No.of.times.30.DPD.or.worse.in.last.12.months 0.22
## 9                           No.of.PL.trades.opened.in.last.6.months 0.22
## 10                     No.of.times.60.DPD.or.worse.in.last.6.months 0.21
## 11                    No.of.times.90.DPD.or.worse.in.last.12.months 0.21
## 12                    No.of.times.60.DPD.or.worse.in.last.12.months 0.19
## 13  No.of.Inquiries.in.last.6.months..excluding.home...auto.loans. 0.19
## 14                             No.of.trades.opened.in.last.6.months 0.18
## 15                     No.of.times.90.DPD.or.worse.in.last.6.months 0.16
```

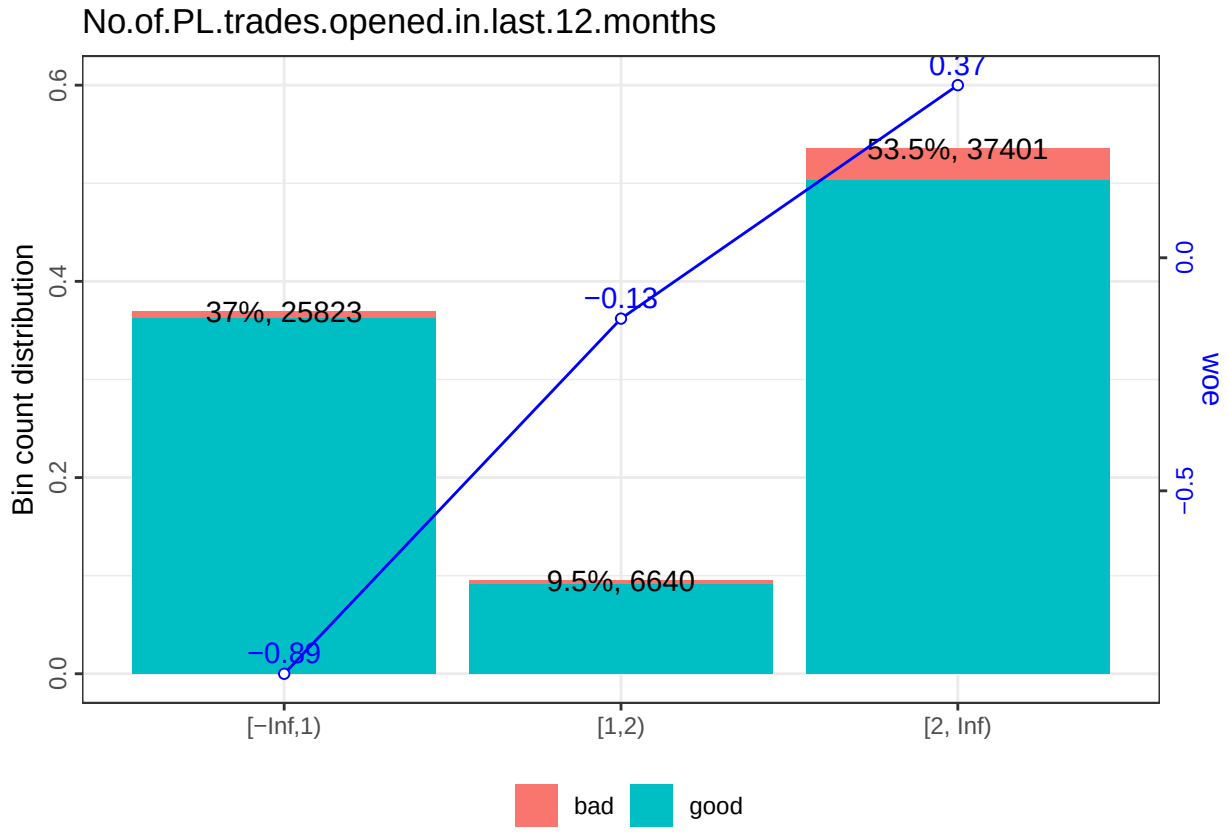
We will now plot the count distribution of the variables that show a relatively strong level of predictive power, along with a line plot of their WOE values:

```
par(mfrow=c(2,2))
woebin_plot(bins[IV_cut$Variable],show_iv = F,line_value = c("woe"))
```

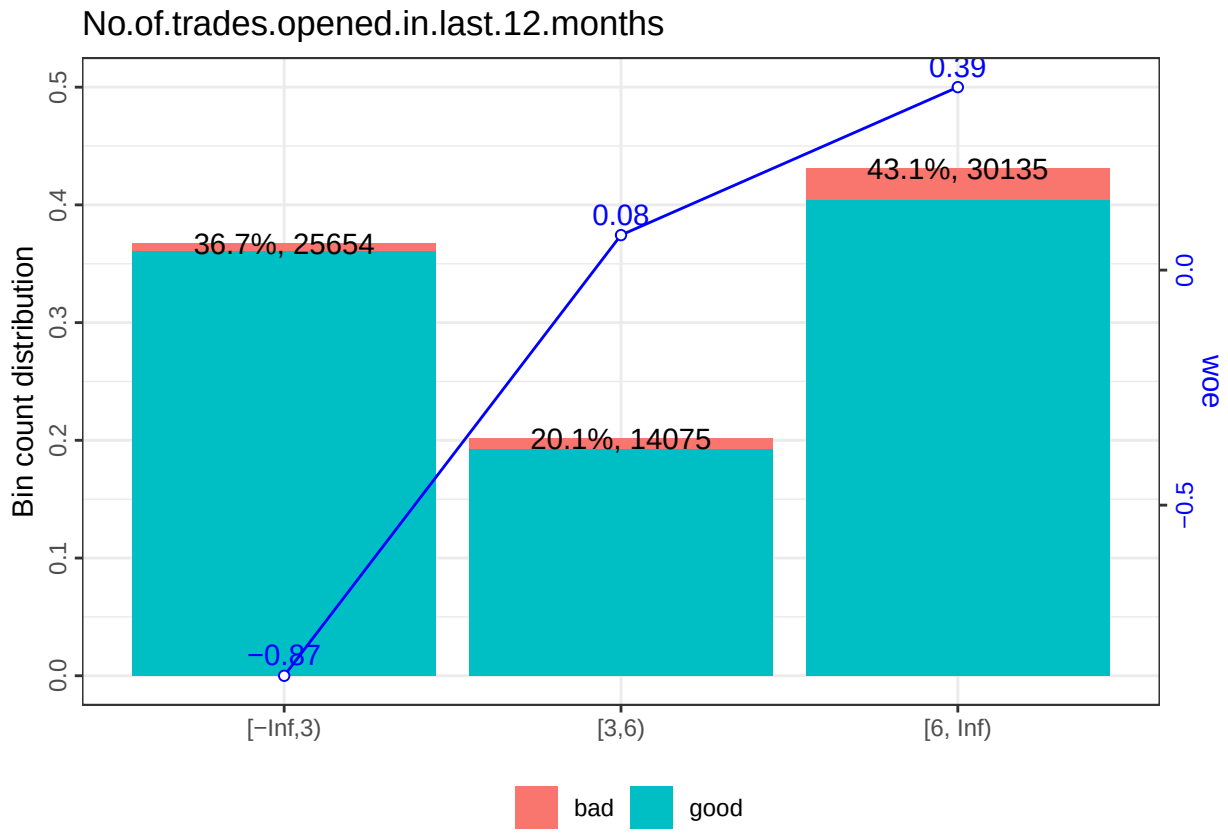
```
## $Avgas.CC.Utilization.in.last.12.months
```



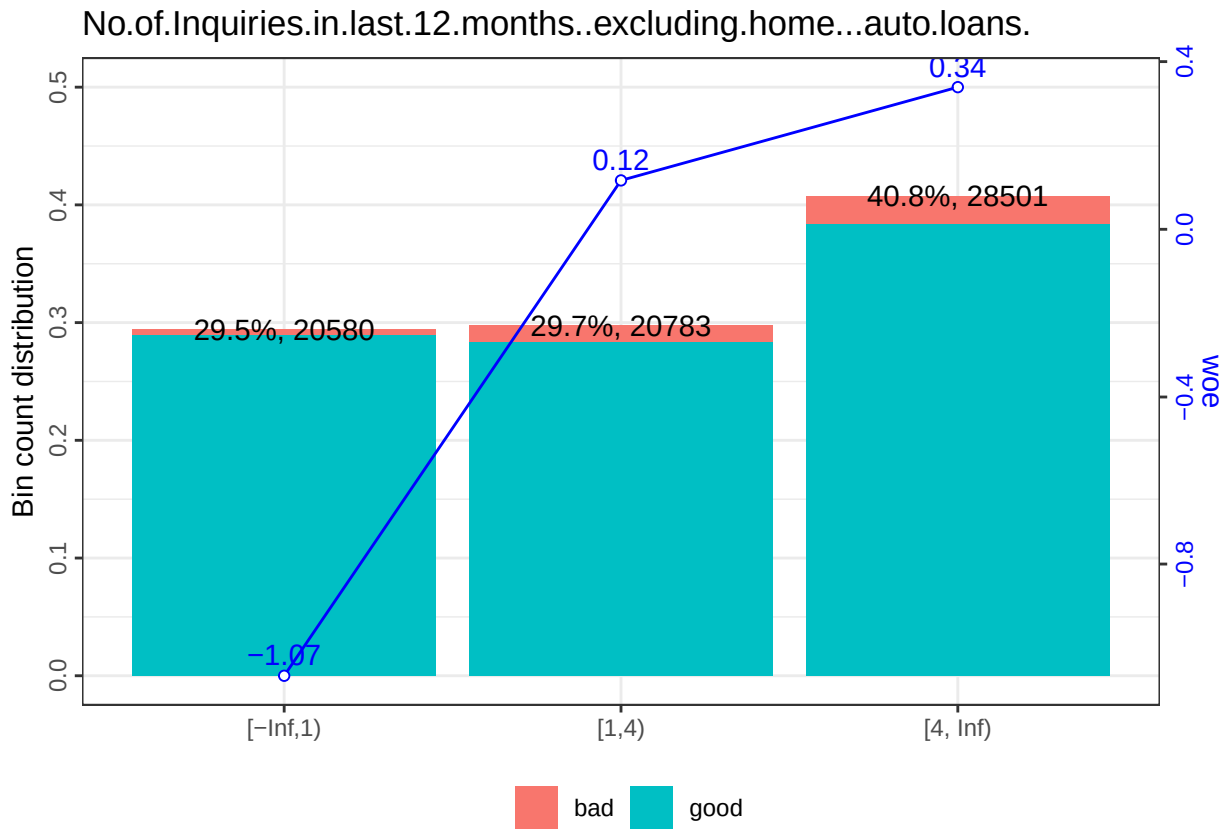

```
##
## $No.of.PL.trades.opened.in.last.12.months
```



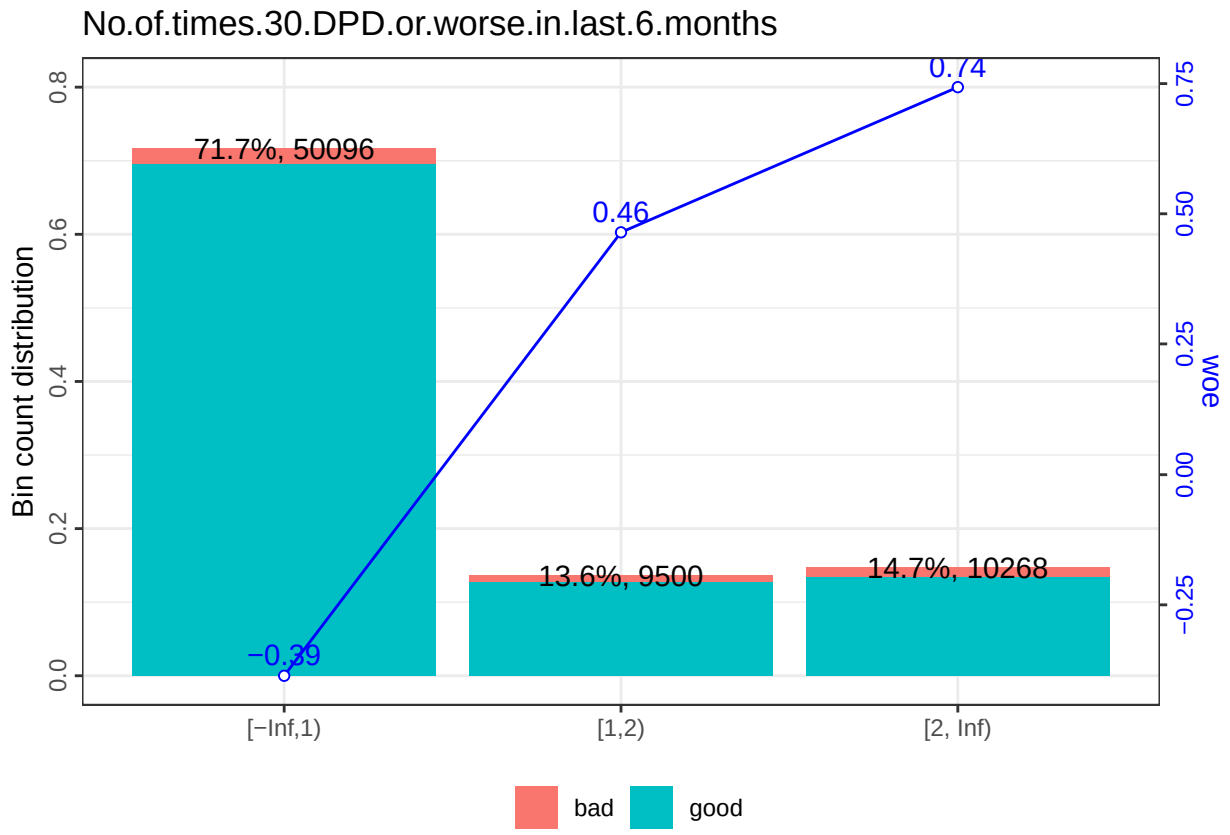

```
##
## $No.of.trades.opened.in.last.12.months
```



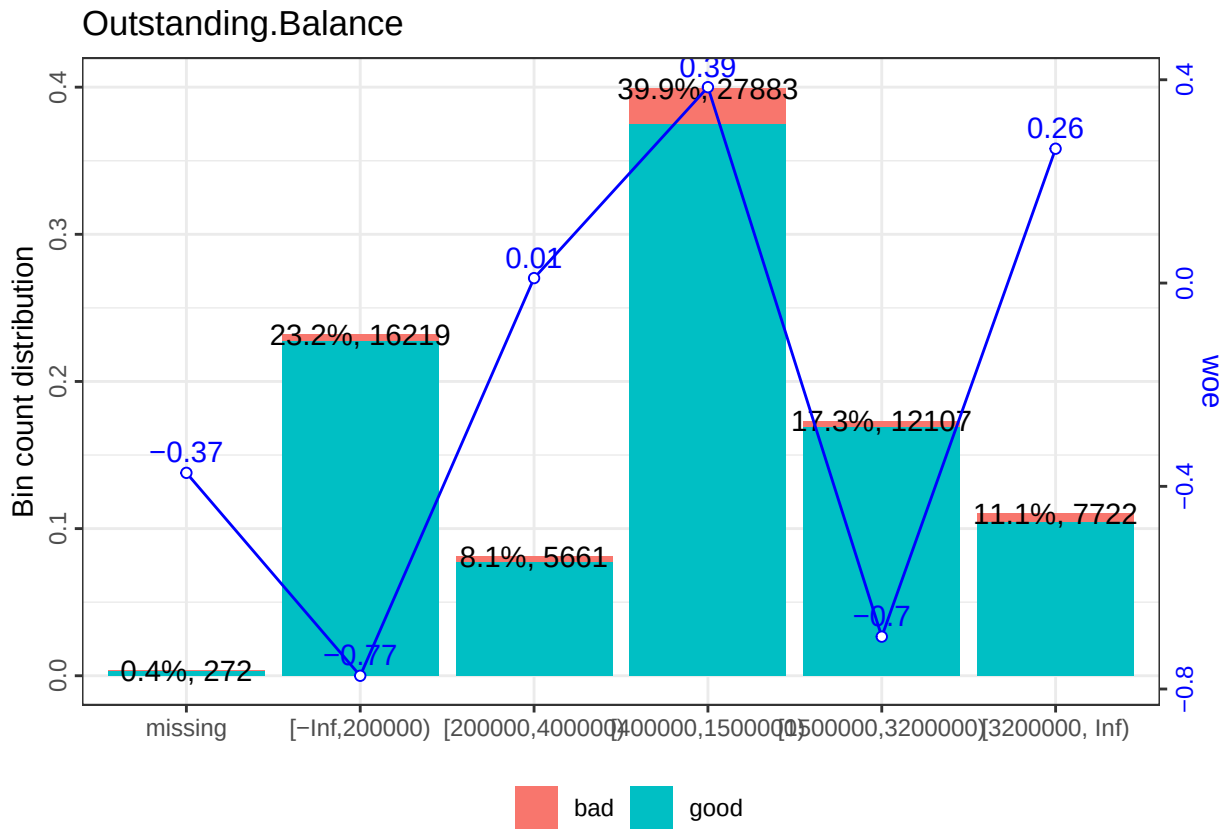

```
##
## $No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.
```



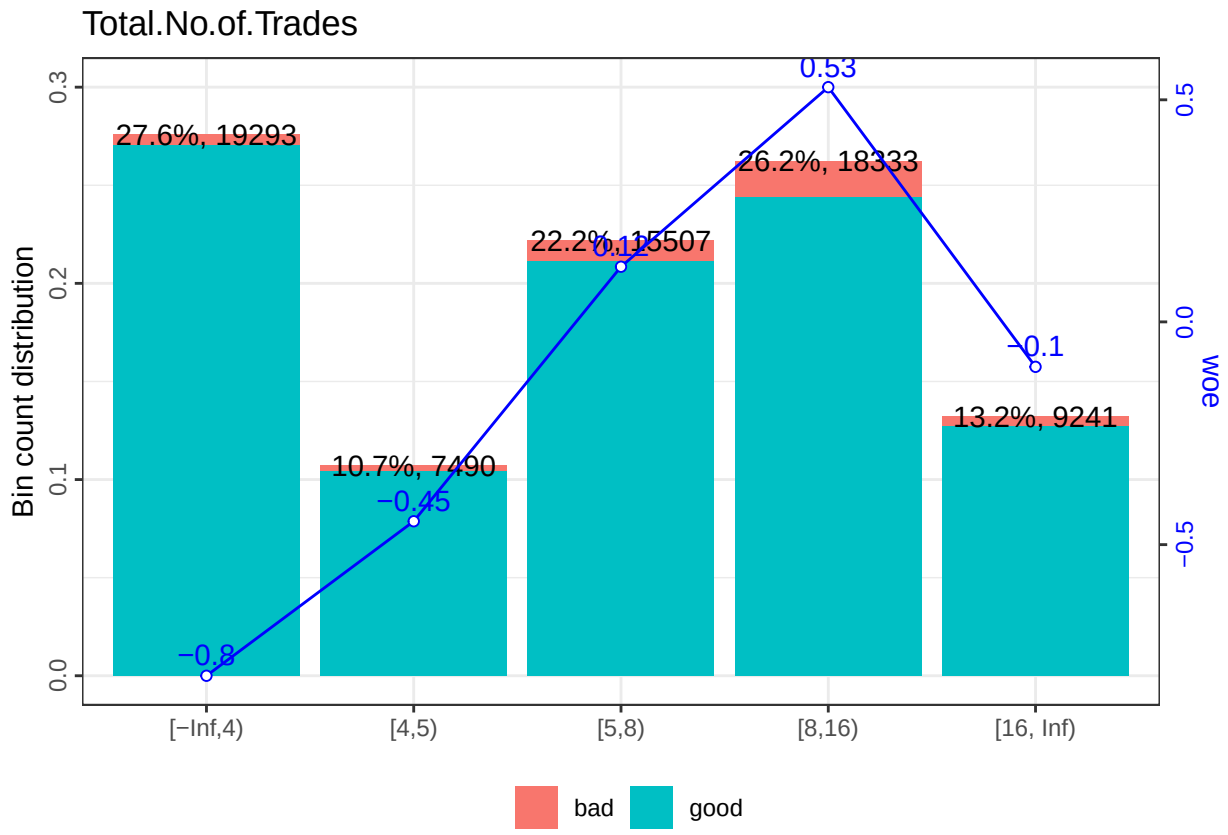

```
##
## $No.of.times.30.DPD.or.worse.in.last.6.months
```



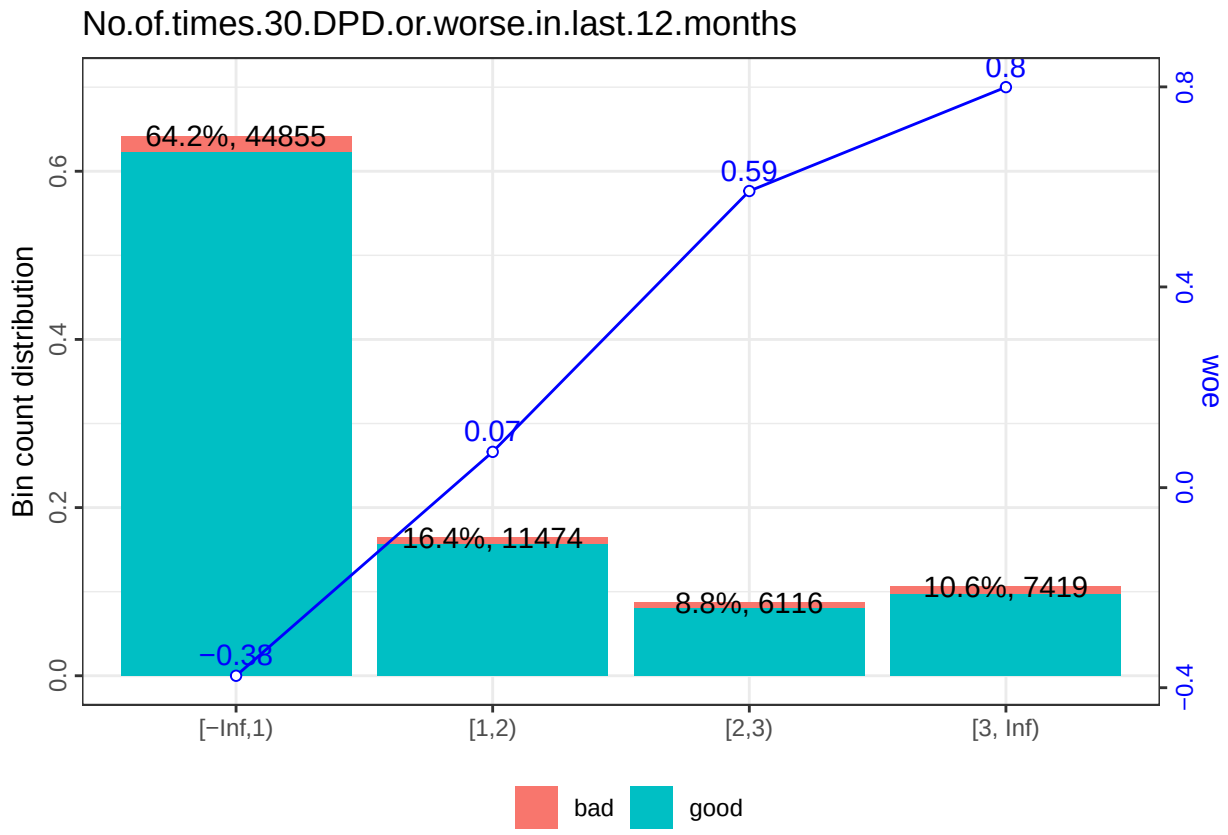

```
##
## $Outstanding.Balance
```



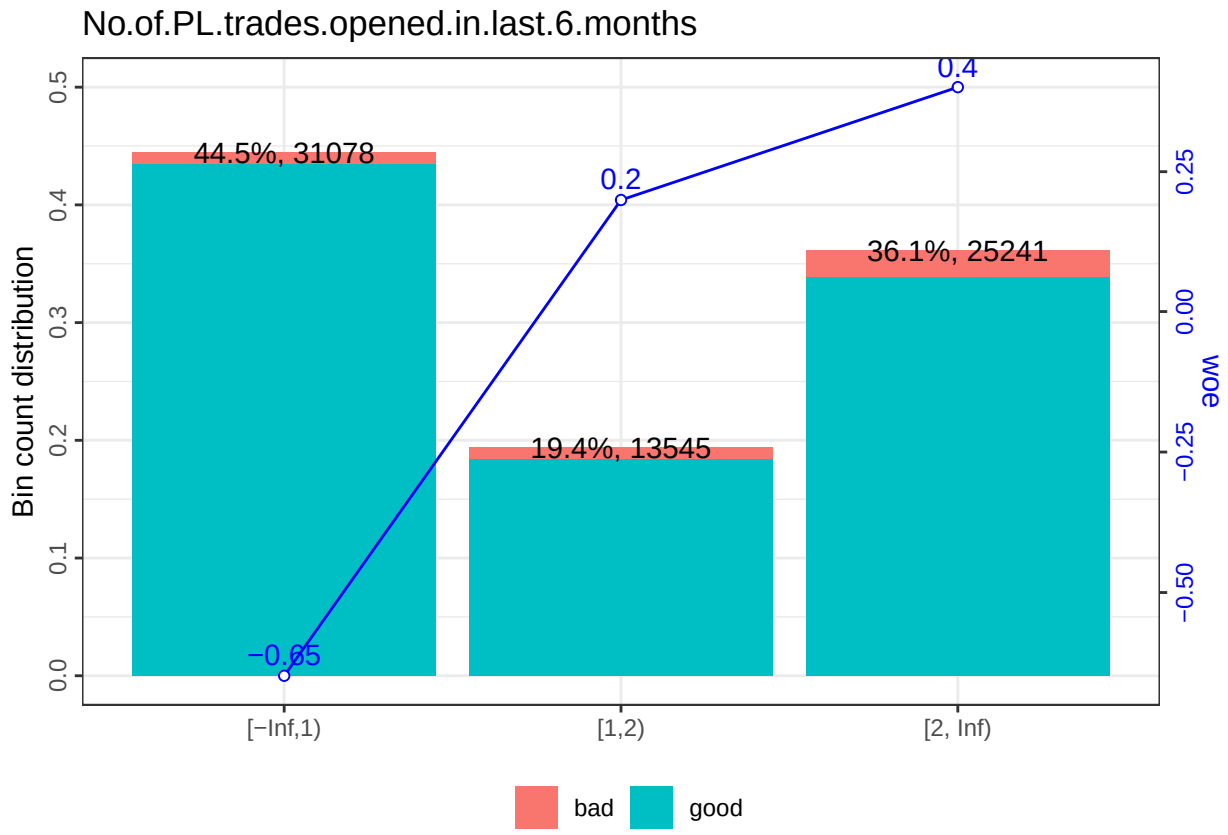

```
##
## $Total.No.of.Trades
```



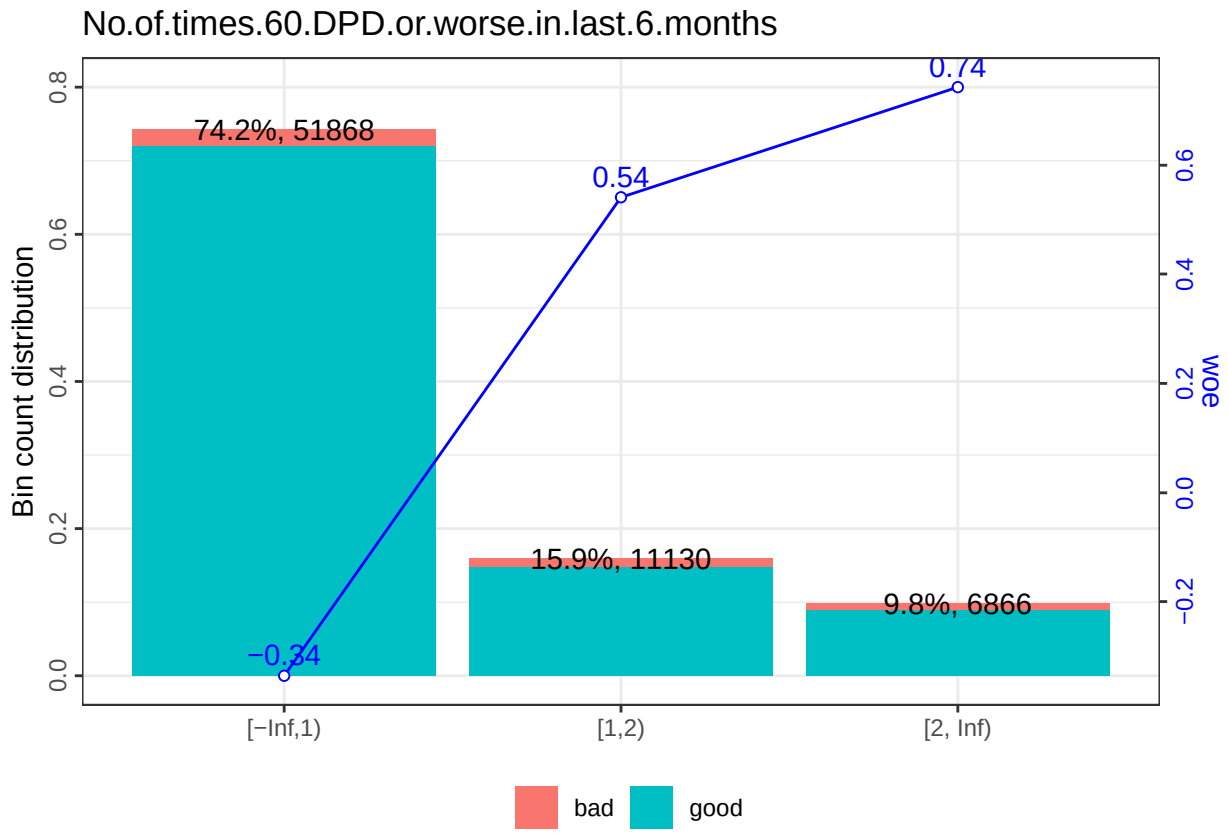

```
##
## $No.of.times.30.DPD.or.worse.in.last.12.months
```



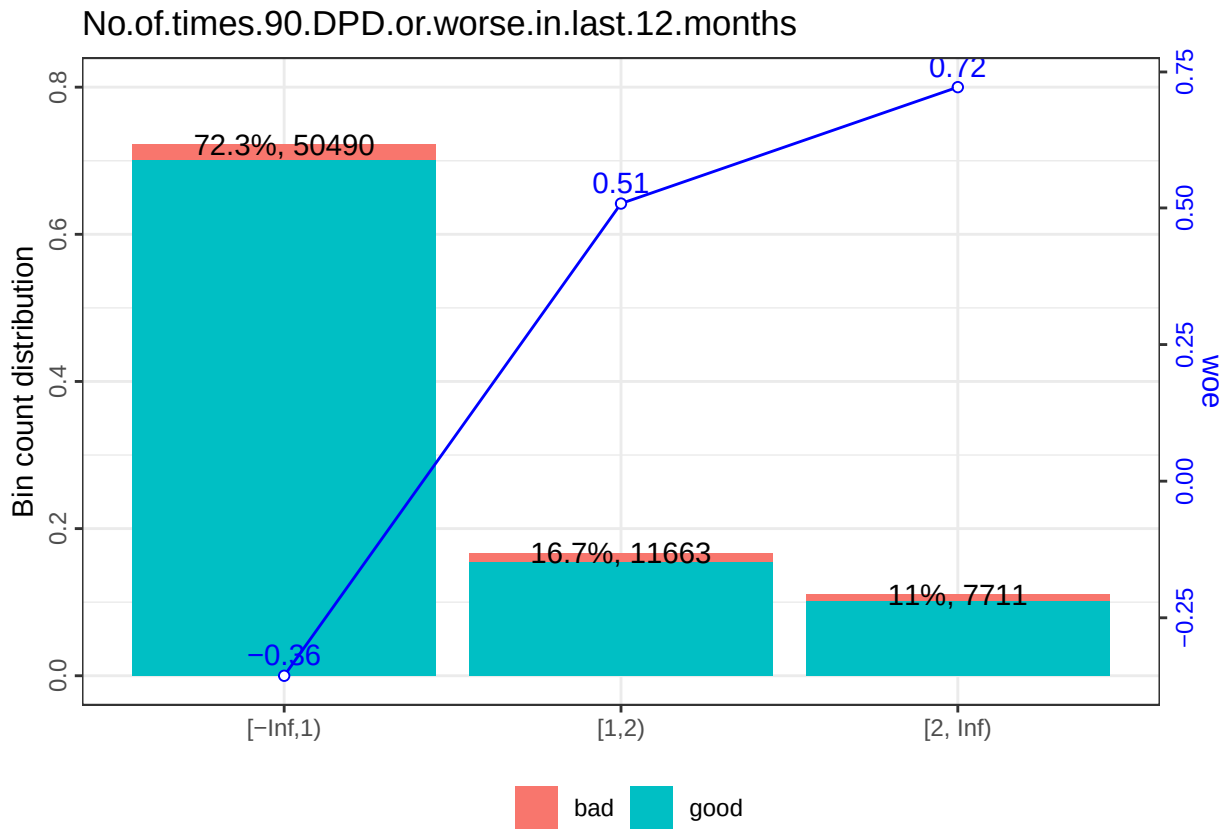

```
##
## $No.of.PL.trades.opened.in.last.6.months
```



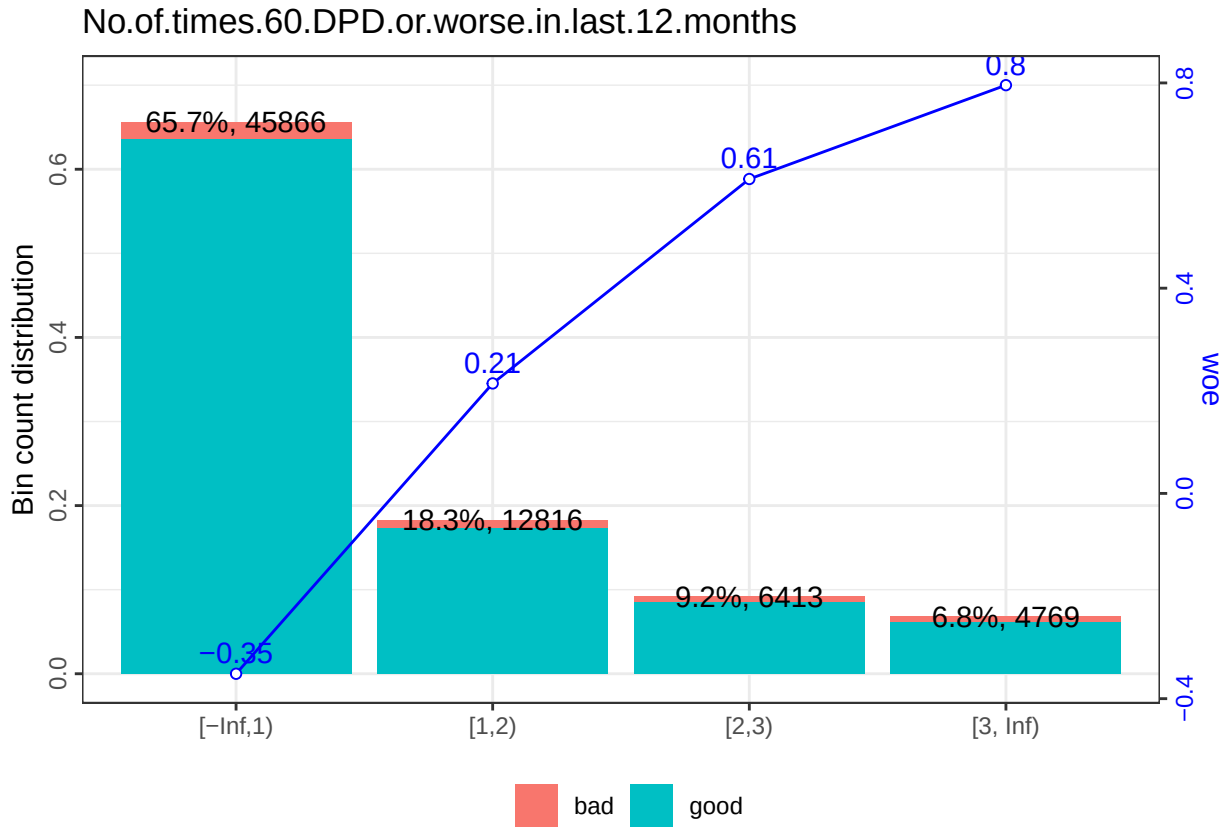

```
##
## $No.of.times.60.DPD.or.worse.in.last.6.months
```



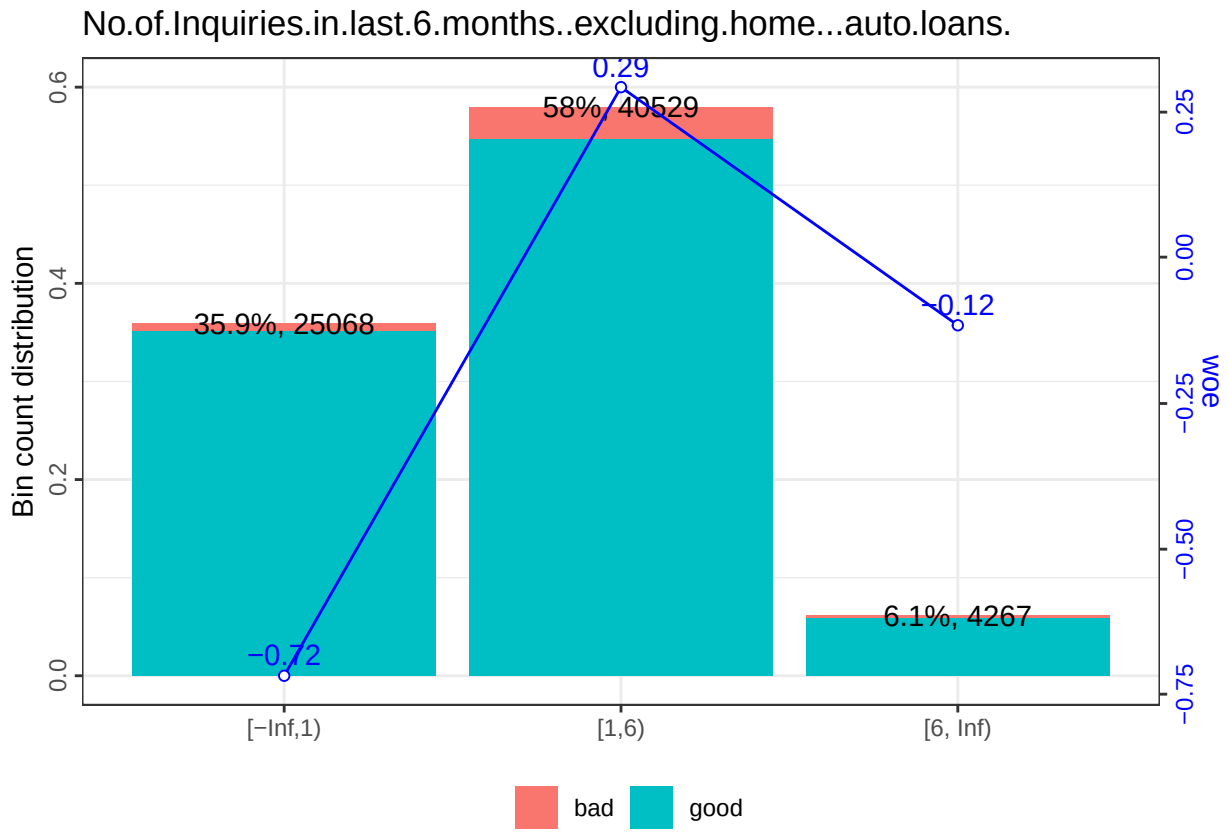

```
##
## $No.of.times.90.DPD.or.worse.in.last.12.months
```



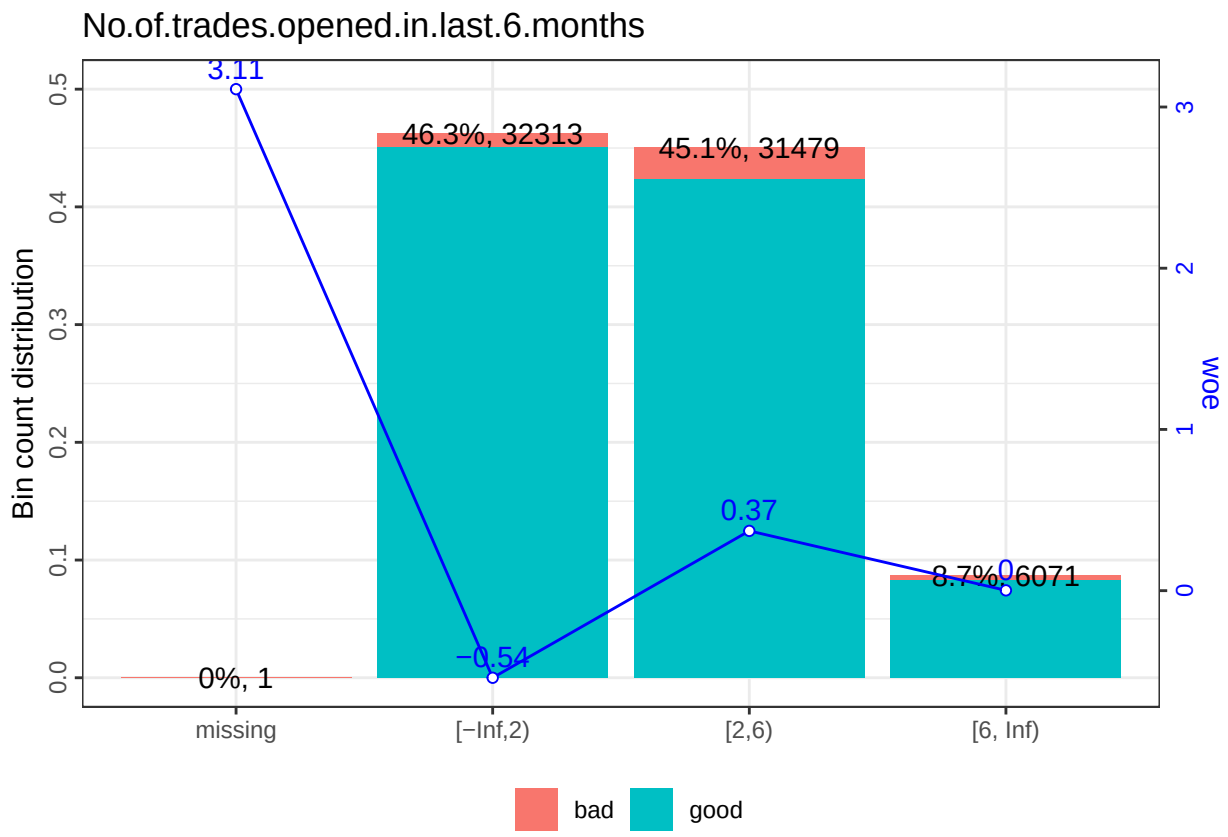

```
##
## $No.of.times.60.DPD.or.worse.in.last.12.months
```



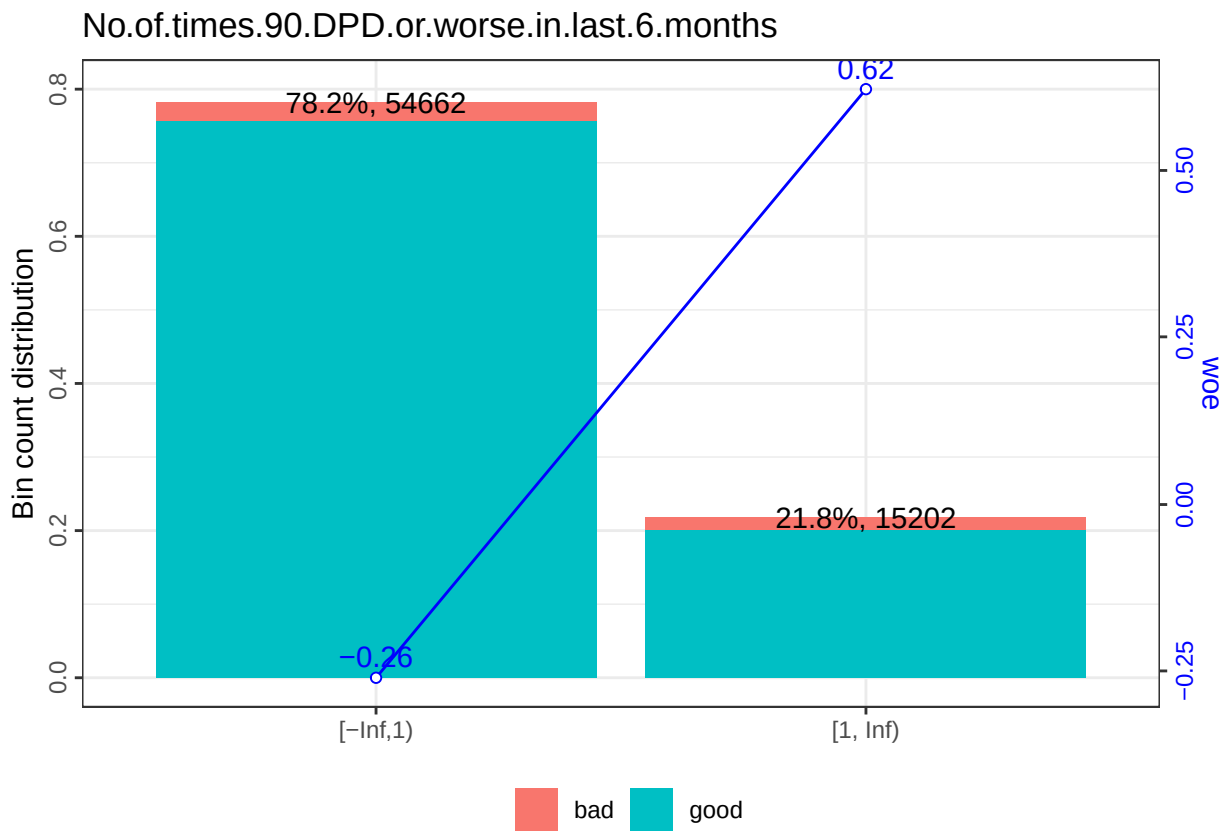

```
##
## $No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.
```




```
##
## $No.of.trades.opened.in.last.6.months
```

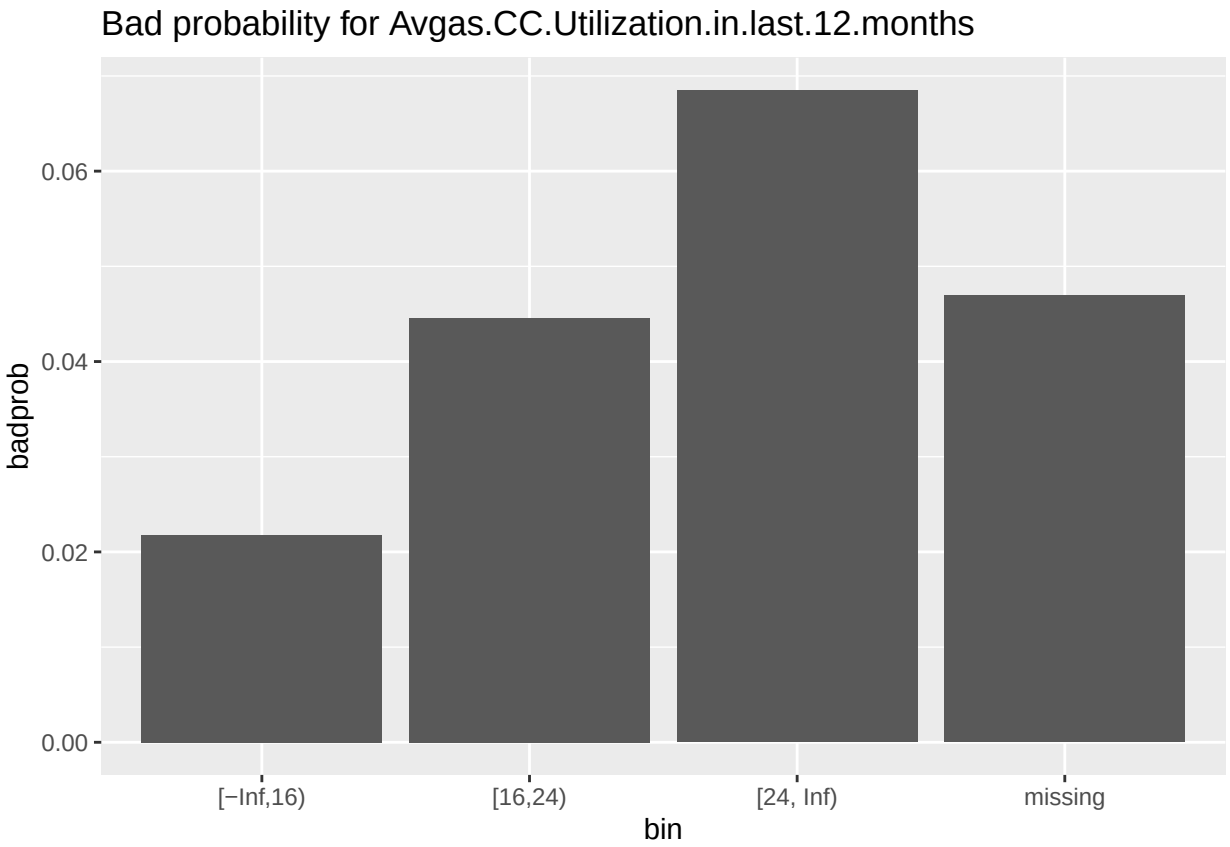



```
##
## $No.of.times.90.DPD.or.worse.in.last.6.months
```

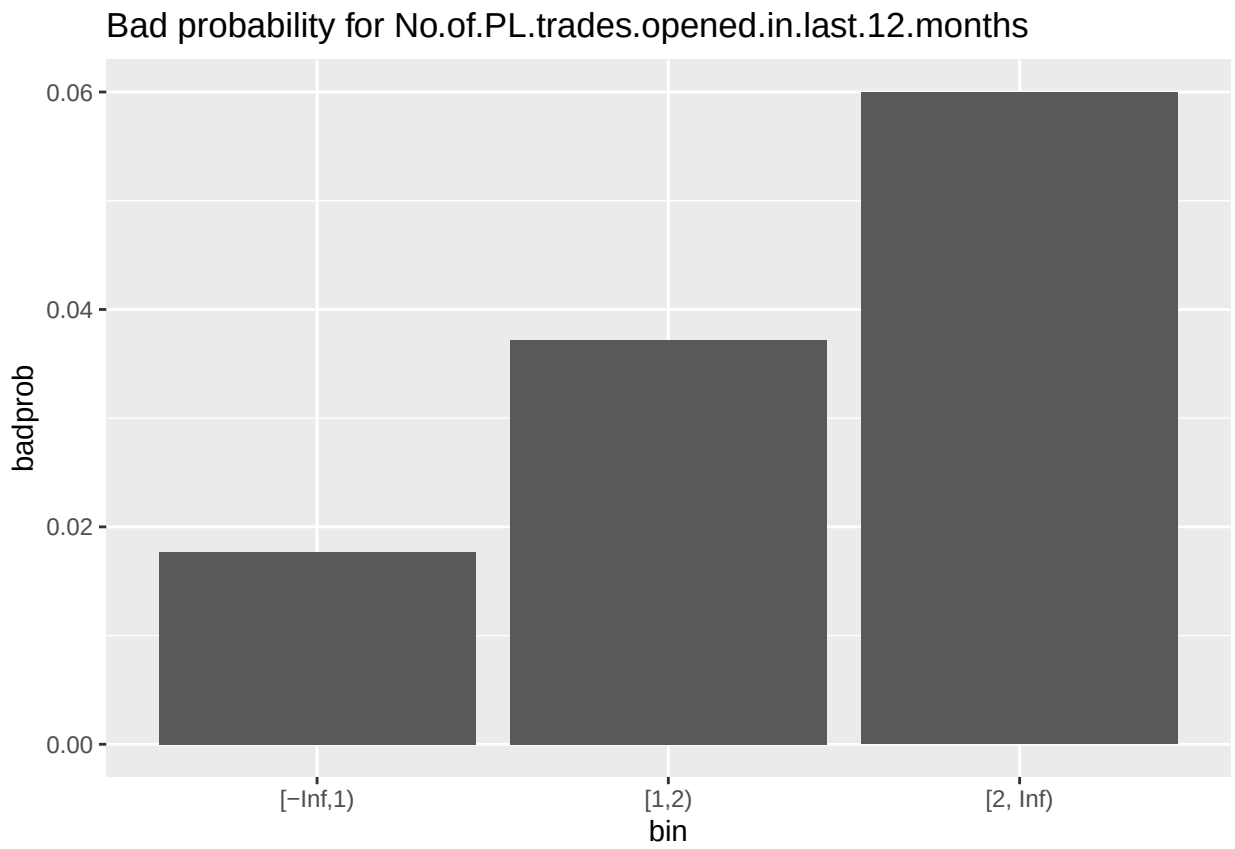


We further include plots describing the probability of bad customers for each independent variable in each category:

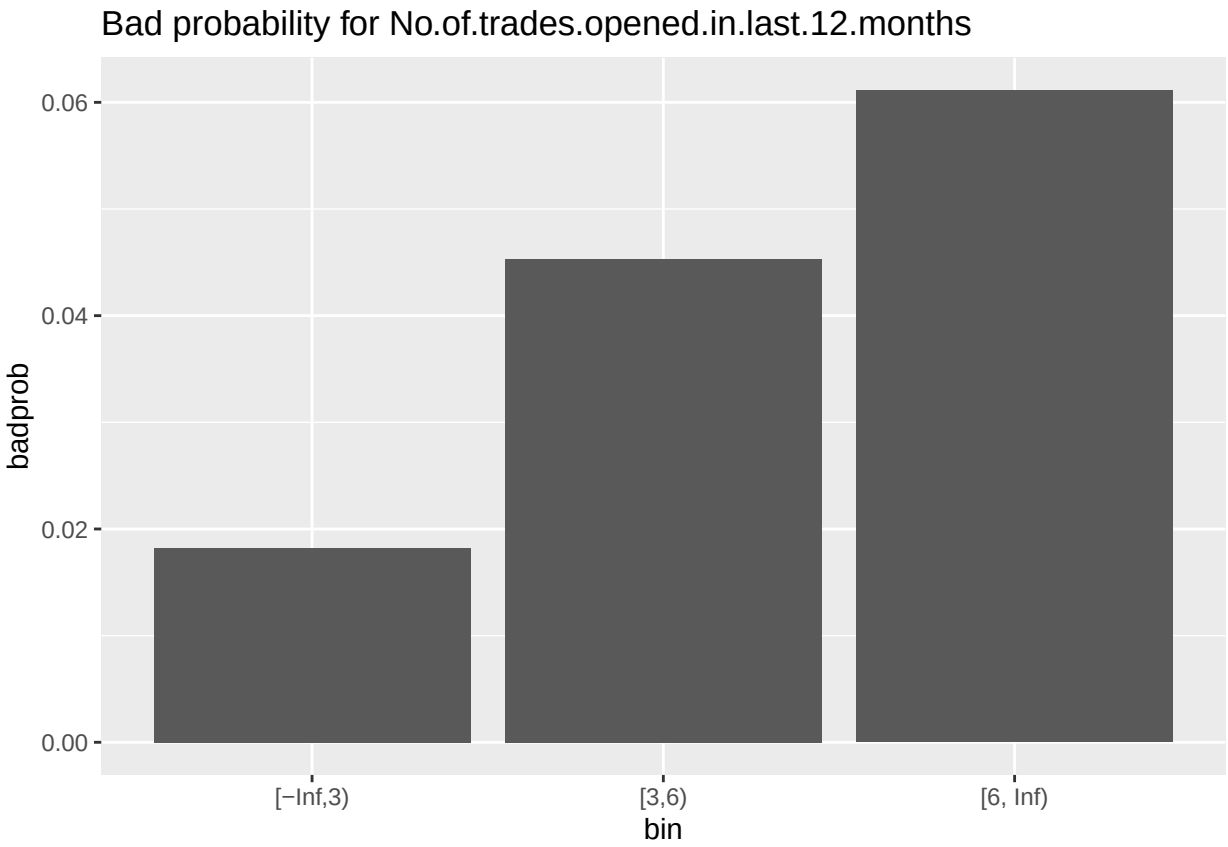
```
ggplot(bins$Avgas.CC.Utilization.in.last.12.months,aes(x=bin,y=badprob))+
  geom_bar(stat="identity")+ labs(title = "Bad probability for Avgas.CC.Utilization.in.last.12.months")
```



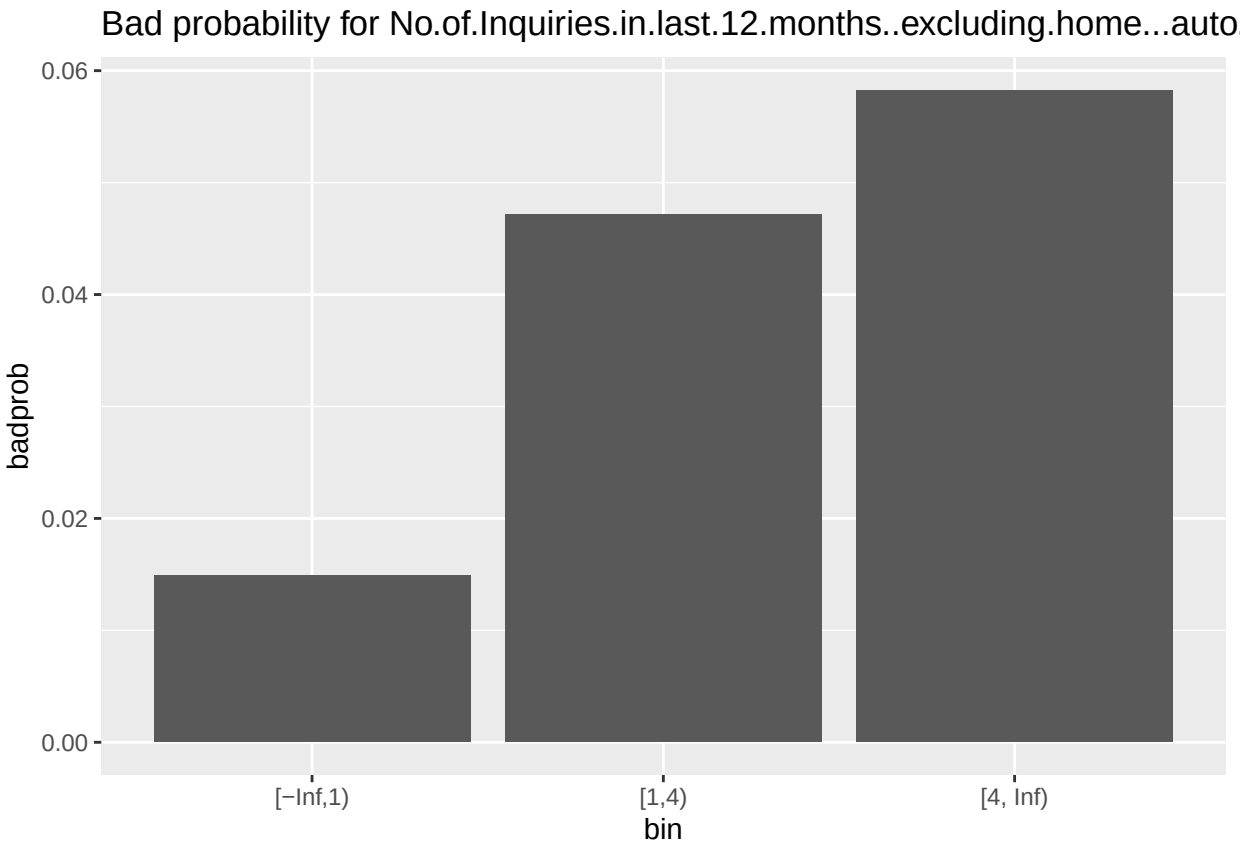

```
ggplot(bins$No.of.PL.trades.opened.in.last.12.months,aes(x=bin,y=badprob))+
  geom_bar(stat="identity")+labs(title = "Bad probability for No.of.PL.trades.opened.in.last.12.months")
```



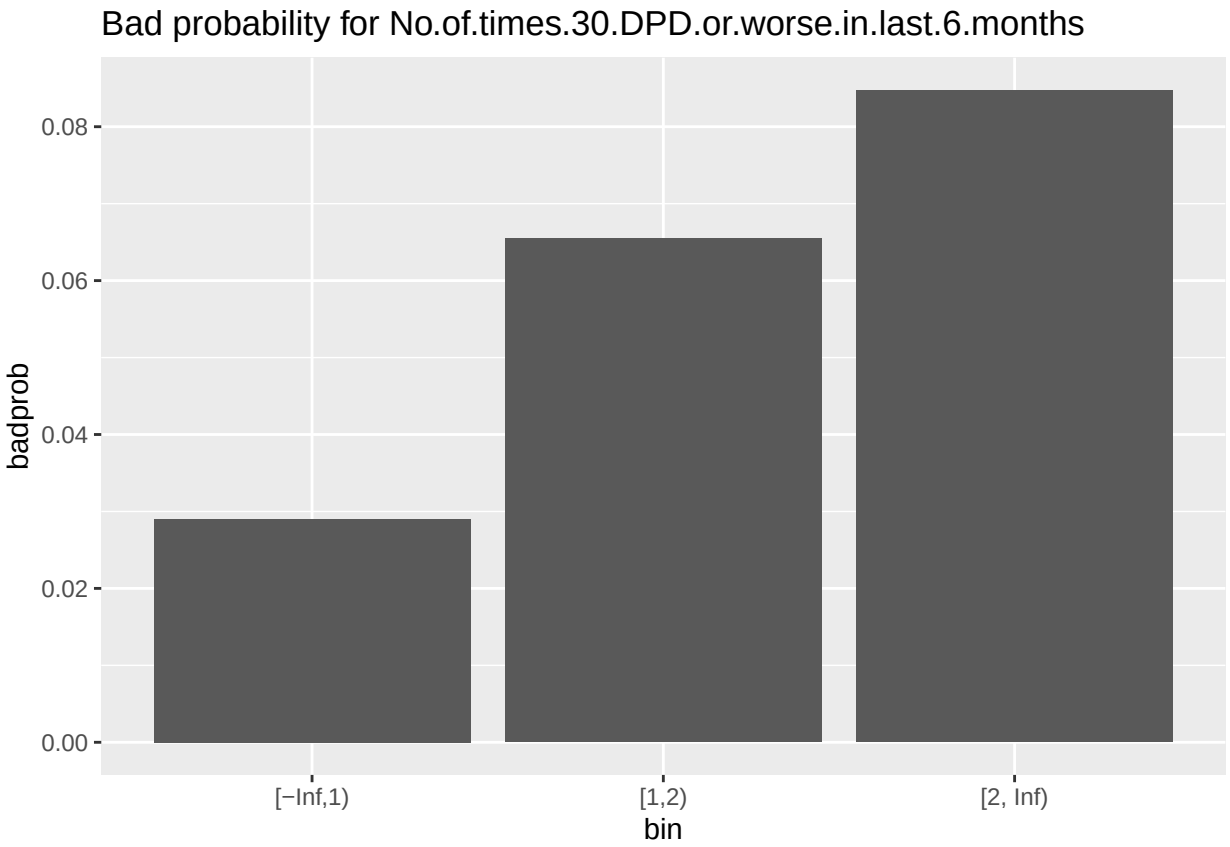

```
ggplot(bins$No.of.trades.opened.in.last.12.months,aes(x=bin,y=badprob))+
  geom_bar(stat="identity")+labs(title = "Bad probability for No.of.trades.opened.in.last.12.months")
```



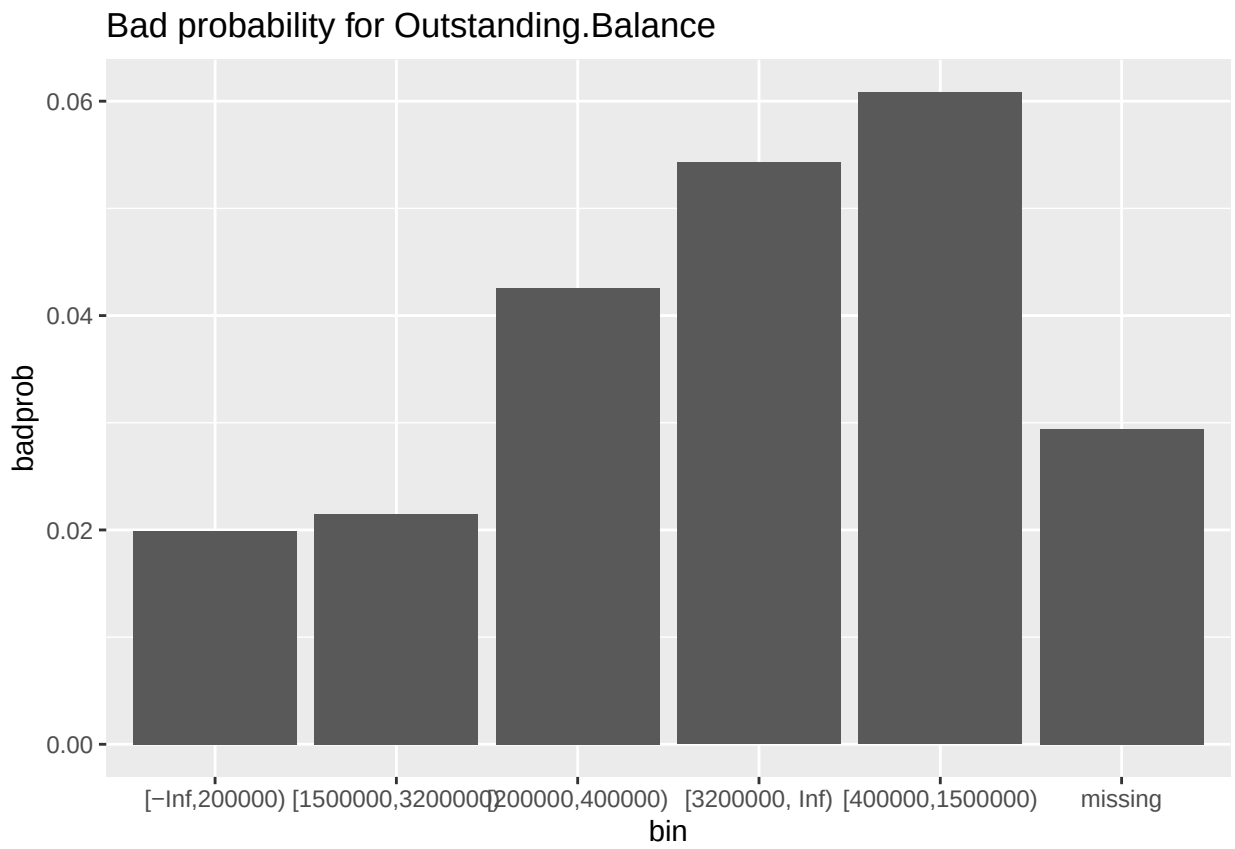

```
ggplot(bins$No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.,aes(x=bin,y=badprob))+
  geom_bar(stat="identity")+labs(title = "Bad probability for No.of.Inquiries.in.last.12.months..exclud
```



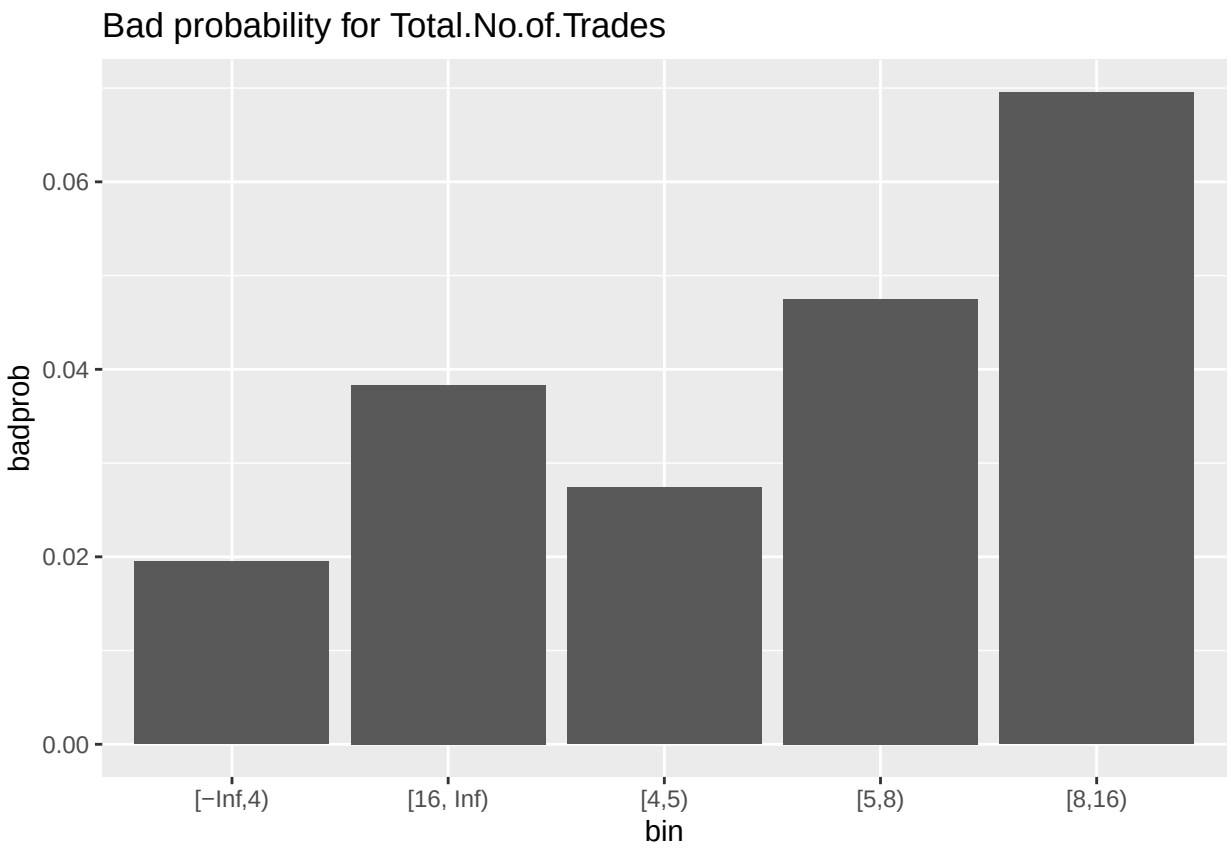

```
ggplot(bins$No.of.times.30.DPD.or.worse.in.last.6.months,aes(x=bin,y=badprob))+
  geom_bar(stat="identity")+labs(title = "Bad probability for No.of.times.30.DPD.or.worse.in.last.6.mon
```



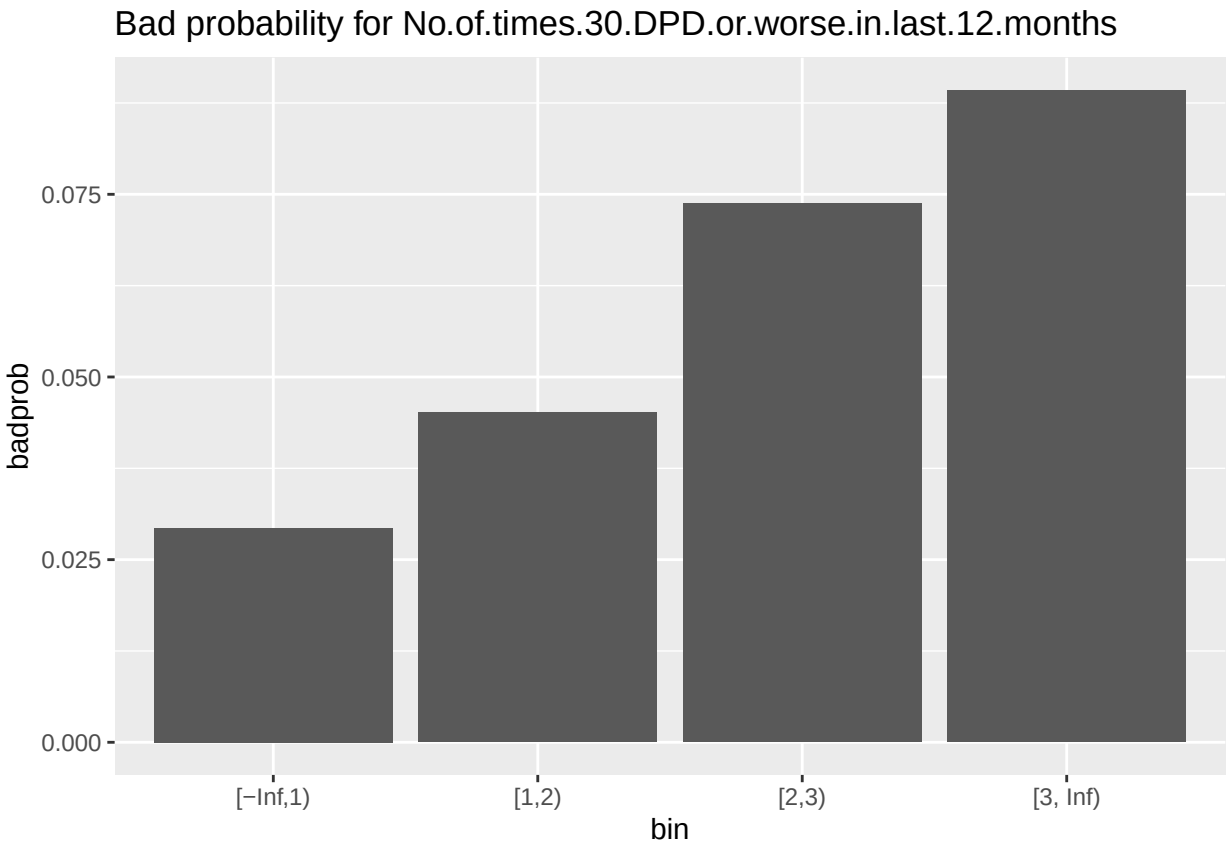

```
ggplot(bins$Outstanding.Balance,aes(x=bin,y=badprob))+
  geom_bar(stat="identity")+labs(title = "Bad probability for Outstanding.Balance")
```



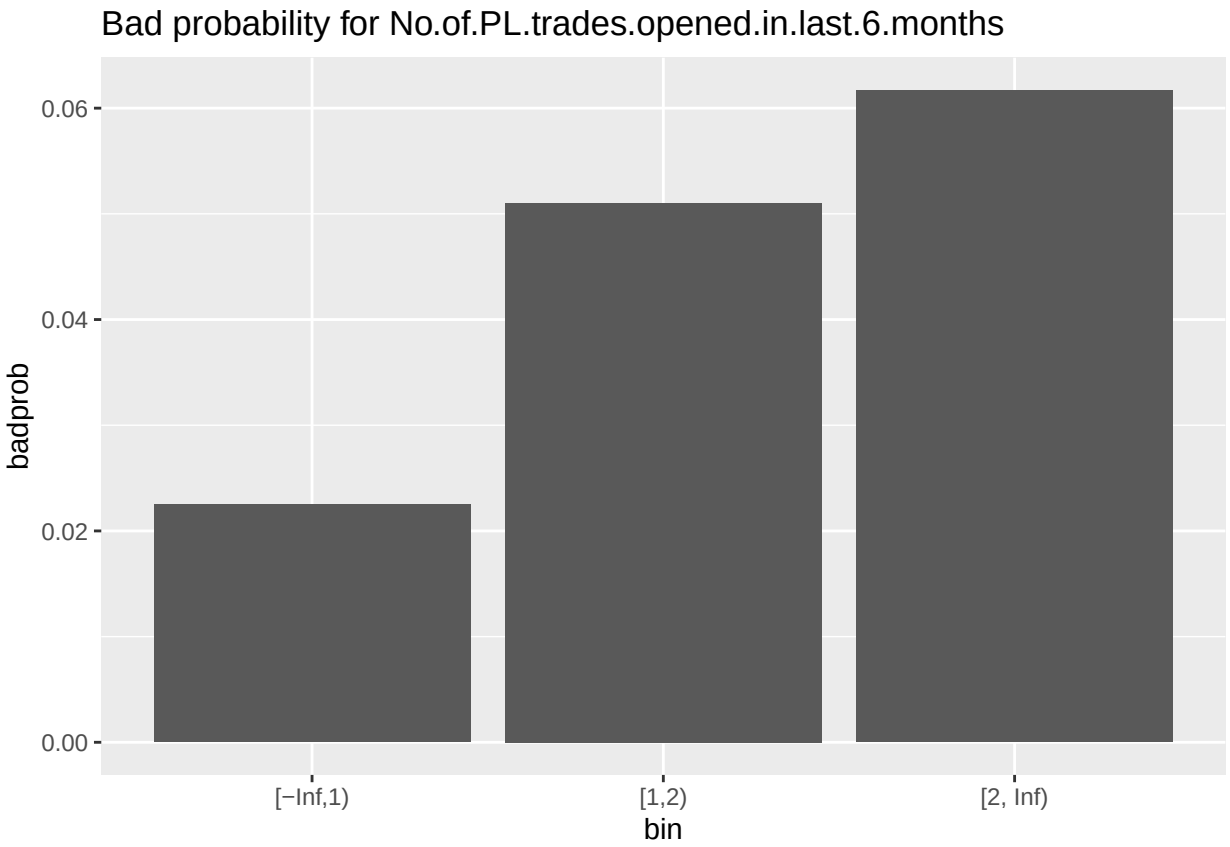

```
ggplot(bins$Total.No.of.Trades,aes(x=bin,y=badprob))+
  geom_bar(stat="identity")+labs(title = "Bad probability for Total.No.of.Trades")
```



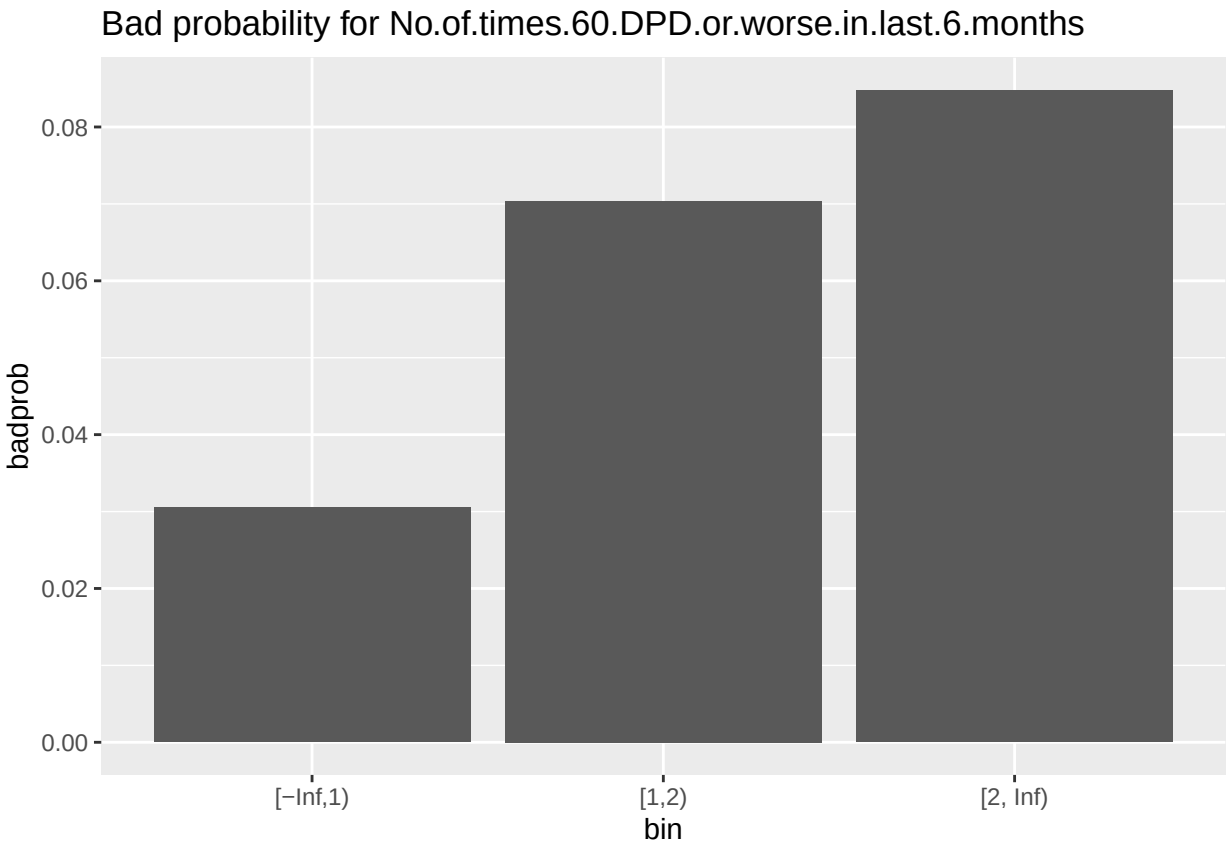

```
ggplot(bins$No.of.times.30.DPD.or.worse.in.last.12.months,aes(x=bin,y=badprob))+
  geom_bar(stat="identity")+labs(title = "Bad probability for No.of.times.30.DPD.or.worse.in.last.12.mo
```



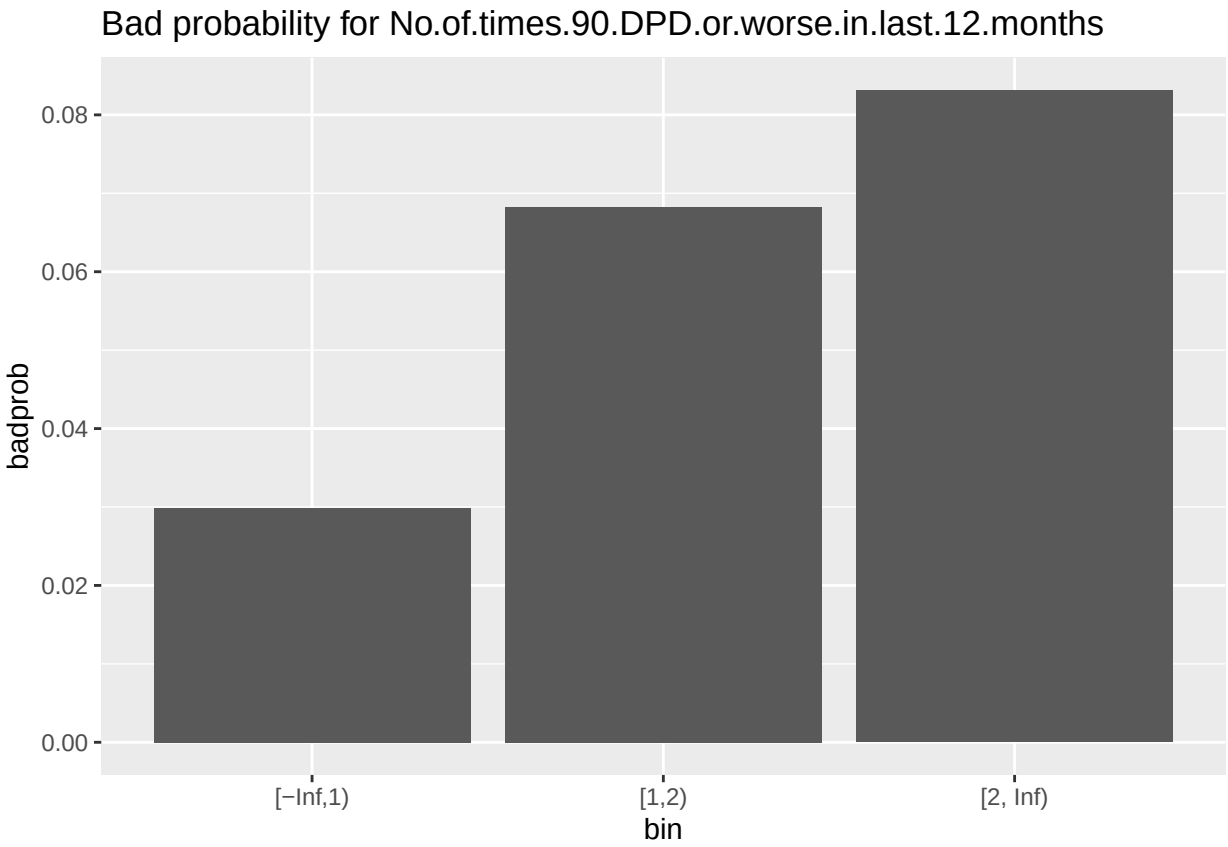

```
ggplot(bins$No.of.PL.trades.opened.in.last.6.months,aes(x=bin,y=badprob))+
  geom_bar(stat="identity")+labs(title = "Bad probability for No.of.PL.trades.opened.in.last.6.months")
```



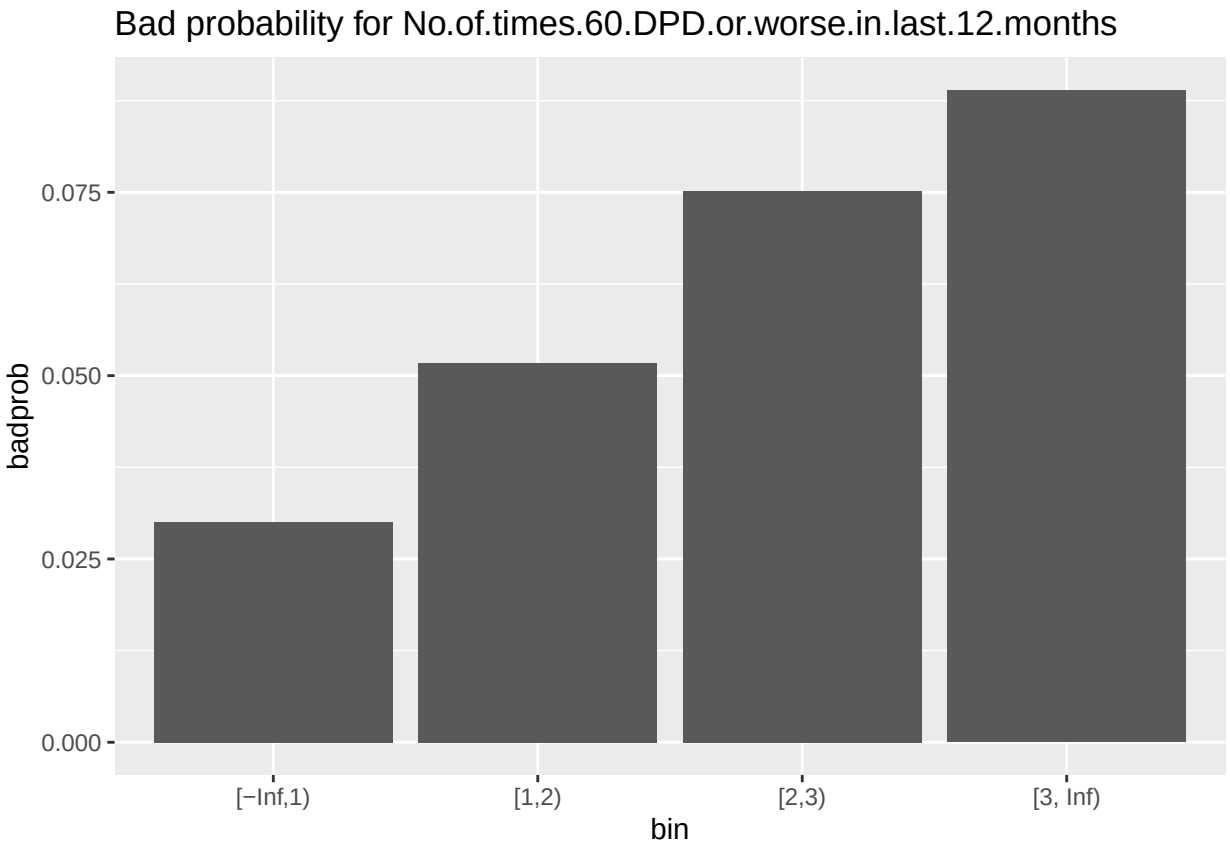

```
ggplot(bins$No.of.times.60.DPD.or.worse.in.last.6.months,aes(x=bin,y=badprob))+
  geom_bar(stat="identity")+labs(title = "Bad probability for No.of.times.60.DPD.or.worse.in.last.6.mon
```



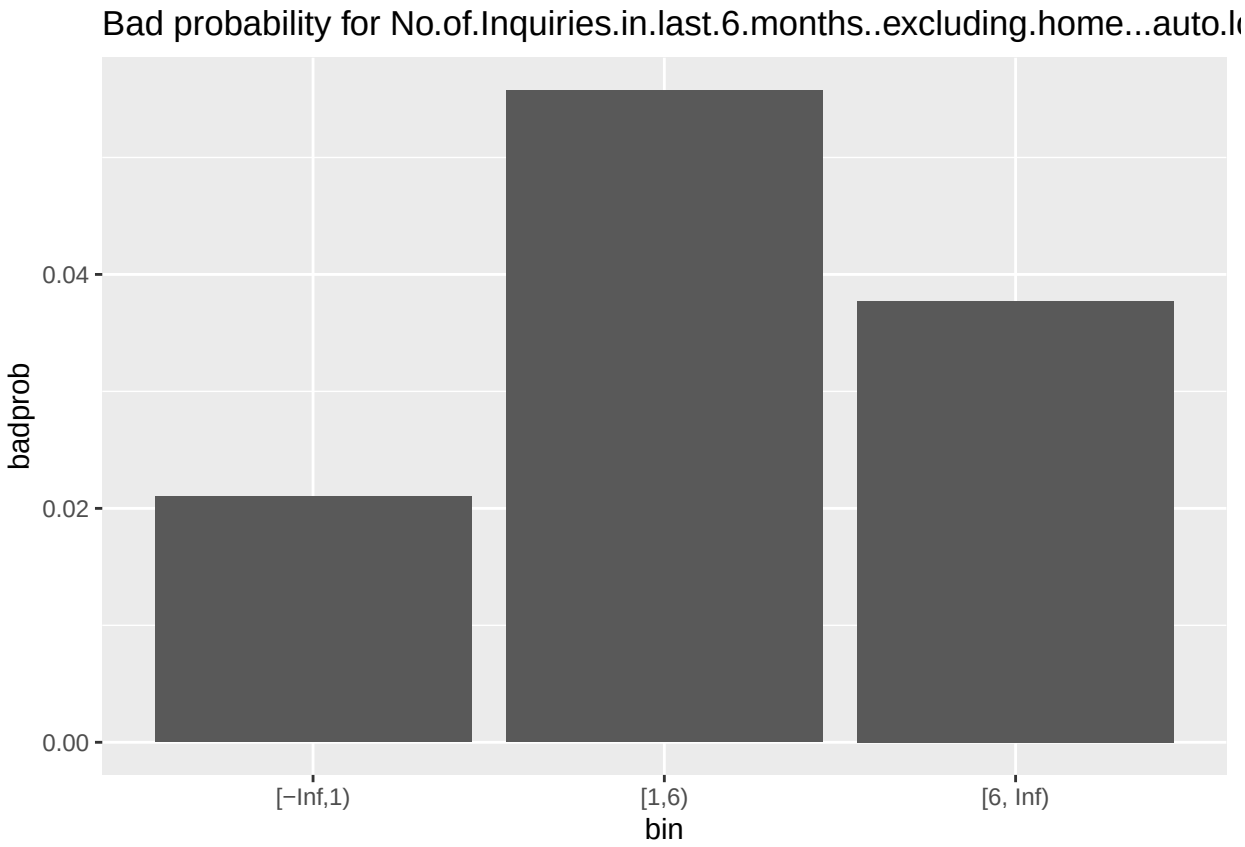

```
ggplot(bins$No.of.times.90.DPD.or.worse.in.last.12.months,aes(x=bin,y=badprob))+
  geom_bar(stat="identity")+labs(title = "Bad probability for No.of.times.90.DPD.or.worse.in.last.12.mo
```



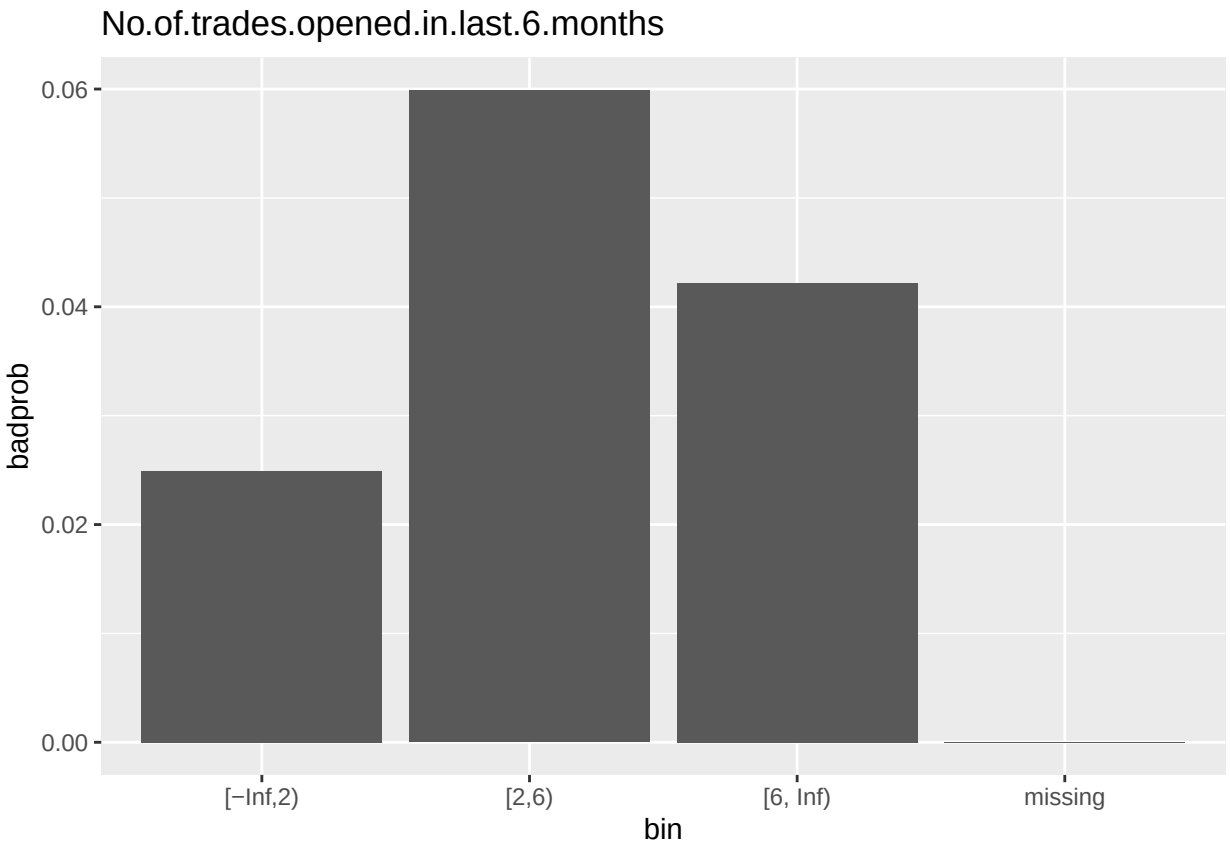

```
ggplot(bins$No.of.times.60.DPD.or.worse.in.last.12.months,aes(x=bin,y=badprob))+
  geom_bar(stat="identity")+labs(title = "Bad probability for No.of.times.60.DPD.or.worse.in.last.12.mo
```



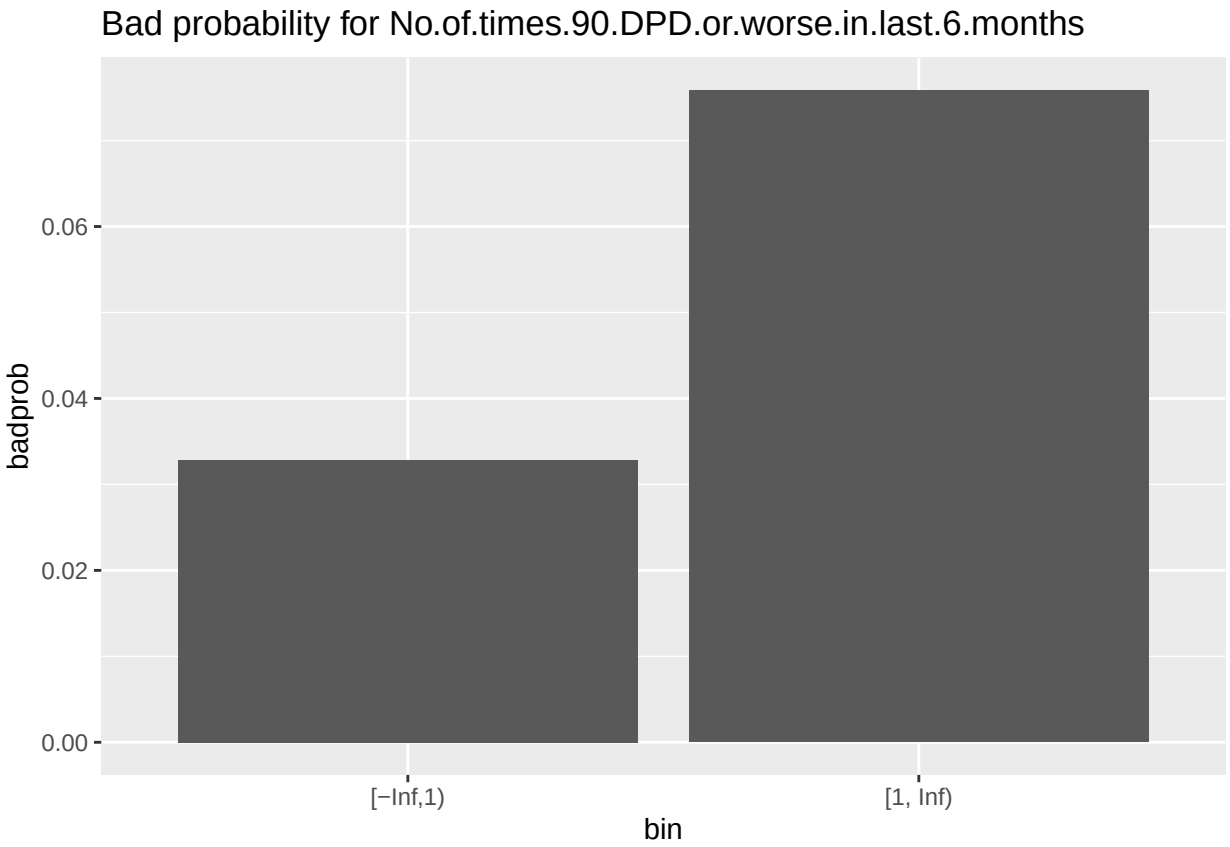

```
ggplot(bins$No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.,aes(x=bin,y=badprob))+
  geom_bar(stat="identity")+labs(title = "Bad probability for No.of.Inquiries.in.last.6.months..excludi
```




```
ggplot(bins$No.of.trades.opened.in.last.6.months,aes(x=bin,y=badprob))+
  geom_bar(stat="identity")+labs(title = "No.of.trades.opened.in.last.6.months")
```




```
ggplot(bins$No.of.times.90.DPD.or.worse.in.last.6.months,aes(x=bin,y=badprob))+
  geom_bar(stat="identity")+labs(title = "Bad probability for No.of.times.90.DPD.or.worse.in.last.6.mon
```


Bad probability for No.of.times.90.DPD.or.worse.in.last.6.months
badprob
0.06
0.04
0.02
0.00
[−Inf,1)
[1, Inf)
bin