BCG - Task 2

Lingyu Tan

1 Data loading

First, we need to read all the data we got. We notice that "ml_case_training_output.csv" is the churn status for the clients in "ml_case_training_data.csv", so we also need to merge them into one data frame for convenience.

```
### Read all csv files
train_data <- read.csv("ml_case_training_data.csv")
train_data_op <- read.csv("ml_case_training_output.csv")
hist <- read.csv("ml_case_training_hist_data.csv")

### Quick look at the data
head(train_data)
```

```
##                                 id                     activity_new
## 1 48ada52261e7cf58715202705a0451c9 esoiiifxdlbkcsluxmfuacbdckommixw
## 2 24011ae4ebbe3035111d65fa7c15bc57
## 3 d29c2c54acc38ff3c0614d0a653813dd
## 4 764c75f661154dac3a6c254cd082ea7d
## 5 bba03439a292a1e166f80264c16191cb
## 6 568bb38a1afd7c0fc49c77b3789b59a3 sfisfxfcocfpcmckuekokxuseixdaoeu
##   campaign_disc_ele                    channel_sales cons_12m cons_gas_12m
## 1                NA lmkebamcaaclubfxadlmueccxoimlema   309275            0
## 2                NA foosdfpfkusacimwkcsosbicdxkicaua        0        54946
## 3                NA                                      4660            0
## 4                NA foosdfpfkusacimwkcsosbicdxkicaua      544            0
## 5                NA lmkebamcaaclubfxadlmueccxoimlema     1584            0
## 6                NA foosdfpfkusacimwkcsosbicdxkicaua   121335            0
##   cons_last_month date_activ  date_end date_first_activ date_modif_prod
## 1           10025  2012/11/7 2016/11/6                        2012/11/7
## 2               0  2013/6/15 2016/6/15
## 3               0  2009/8/21 2016/8/30                        2009/8/21
## 4               0  2010/4/16 2016/4/16                        2010/4/16
## 5               0  2010/3/30 2016/3/30                        2010/3/30
## 6           12400   2010/4/8  2016/4/8         2010/4/8        2010/4/8
##   date_renewal forecast_base_bill_ele forecast_base_bill_year forecast_bill_12m
## 1    2015/11/9                     NA                      NA                NA
## 2    2015/6/23                     NA                      NA                NA
## 3    2015/8/31                     NA                      NA                NA
## 4    2015/4/17                     NA                      NA                NA
## 5    2015/3/31                     NA                      NA                NA
## 6    2015/4/12                1399.83                 1399.83          14559.74
##   forecast_cons forecast_cons_12m forecast_cons_year forecast_discount_energy
```

```
## 1          NA         26520.30            10025                      0
## 2          NA             0.00                0                      0
## 3          NA           189.95                0                      0
## 4          NA            47.96                0                      0
## 5          NA           240.04                0                      0
## 6     1052.37         10865.02            12400                      0
##   forecast_meter_rent_12m forecast_price_energy_p1 forecast_price_energy_p2
## 1                  359.29                 0.095919                 0.088347
## 2                    1.78                 0.114481                 0.098142
## 3                   16.27                 0.145711                 0.000000
## 4                   38.72                 0.165794                 0.087899
## 5                   19.83                 0.146694                 0.000000
## 6                  170.74                 0.110083                 0.093746
##   forecast_price_pow_p1 has_gas imp_cons margin_gross_pow_ele
## 1              58.99595       f   831.80               -41.76
## 2              40.60670       t     0.00                25.44
## 3              44.31138       f     0.00                16.38
## 4              44.31138       f     0.00                28.60
## 5              44.31138       f     0.00                30.22
## 6              40.60670       f  1052.37                -3.18
##   margin_net_pow_ele nb_prod_act net_margin num_years_antig
## 1             -41.76           1    1732.36               3
## 2              25.44           2     678.99               3
## 3              16.38           1      18.89               6
## 4              28.60           1       6.60               6
## 5              30.22           1      25.46               6
## 6              -3.18           1     823.18               6
##                          origin_up pow_max
## 1 ldkssxwpmemidmecebumciepifcamkci 180.000
## 2 lxidpiddsbxsbosboudacockeimpuepw  43.648
## 3 kamkkxfxxuwbdslkwifmmcsiusiuosws  13.800
## 4 kamkkxfxxuwbdslkwifmmcsiusiuosws  13.856
## 5 kamkkxfxxuwbdslkwifmmcsiusiuosws  13.200
## 6 lxidpiddsbxsbosboudacockeimpuepw  75.000
```

```
head(train_data_op)
```

```
##                                 id churn
## 1 48ada52261e7cf58715202705a0451c9     0
## 2 24011ae4ebbe3035111d65fa7c15bc57     1
## 3 d29c2c54acc38ff3c0614d0a653813dd     0
## 4 764c75f661154dac3a6c254cd082ea7d     0
## 5 bba03439a292a1e166f80264c16191cb     0
## 6 568bb38a1afd7c0fc49c77b3789b59a3     0
```

```
head(hist)
```

```
##                                 id price_date price_p1_var price_p2_var
## 1 038af19179925da21a25619c5a24b745   2015/1/1     0.151367            0
## 2 038af19179925da21a25619c5a24b745   2015/2/1     0.151367            0
## 3 038af19179925da21a25619c5a24b745   2015/3/1     0.151367            0
## 4 038af19179925da21a25619c5a24b745   2015/4/1     0.149626            0
## 5 038af19179925da21a25619c5a24b745   2015/5/1     0.149626            0
```

```
## 6 038af19179925da21a25619c5a24b745   2015/6/1     0.149626            0
##   price_p3_var price_p1_fix price_p2_fix price_p3_fix
## 1            0     44.26693            0            0
## 2            0     44.26693            0            0
## 3            0     44.26693            0            0
## 4            0     44.26693            0            0
## 5            0     44.26693            0            0
## 6            0     44.26693            0            0
```

```
### Merge dataset
train <- merge(train_data, train_data_op, by = "id")
```

2 Data type Conversions

By looking at the name and class of each variable in the dataset, we then convert their types as follows (mainly convert characters to their corresponding types such as dates, factors or logical):

```
### List names and types of variables
names(train)
```

```
##  [1] "id"                       "activity_new"
##  [3] "campaign_disc_ele"        "channel_sales"
##  [5] "cons_12m"                 "cons_gas_12m"
##  [7] "cons_last_month"          "date_activ"
##  [9] "date_end"                 "date_first_activ"
## [11] "date_modif_prod"          "date_renewal"
## [13] "forecast_base_bill_ele"   "forecast_base_bill_year"
## [15] "forecast_bill_12m"        "forecast_cons"
## [17] "forecast_cons_12m"        "forecast_cons_year"
## [19] "forecast_discount_energy" "forecast_meter_rent_12m"
## [21] "forecast_price_energy_p1" "forecast_price_energy_p2"
## [23] "forecast_price_pow_p1"    "has_gas"
## [25] "imp_cons"                 "margin_gross_pow_ele"
## [27] "margin_net_pow_ele"       "nb_prod_act"
## [29] "net_margin"               "num_years_antig"
## [31] "origin_up"                "pow_max"
## [33] "churn"
```

```
lapply(train, class)
```

```
## $id
## [1] "character"
##
## $activity_new
## [1] "character"
##
## $campaign_disc_ele
## [1] "logical"
##
## $channel_sales
## [1] "character"
```

```
##
## $cons_12m
## [1] "integer"
##
## $cons_gas_12m
## [1] "integer"
##
## $cons_last_month
## [1] "integer"
##
## $date_activ
## [1] "character"
##
## $date_end
## [1] "character"
##
## $date_first_activ
## [1] "character"
##
## $date_modif_prod
## [1] "character"
##
## $date_renewal
## [1] "character"
##
## $forecast_base_bill_ele
## [1] "numeric"
##
## $forecast_base_bill_year
## [1] "numeric"
##
## $forecast_bill_12m
## [1] "numeric"
##
## $forecast_cons
## [1] "numeric"
##
## $forecast_cons_12m
## [1] "numeric"
##
## $forecast_cons_year
## [1] "integer"
##
## $forecast_discount_energy
## [1] "integer"
##
## $forecast_meter_rent_12m
## [1] "numeric"
##
## $forecast_price_energy_p1
## [1] "numeric"
##
## $forecast_price_energy_p2
## [1] "numeric"
```

```
##
## $forecast_price_pow_p1
## [1] "numeric"
##
## $has_gas
## [1] "character"
##
## $imp_cons
## [1] "numeric"
##
## $margin_gross_pow_ele
## [1] "numeric"
##
## $margin_net_pow_ele
## [1] "numeric"
##
## $nb_prod_act
## [1] "integer"
##
## $net_margin
## [1] "numeric"
##
## $num_years_antig
## [1] "integer"
##
## $origin_up
## [1] "character"
##
## $pow_max
## [1] "numeric"
##
## $churn
## [1] "integer"
```

```
names(hist)
```

```
## [1] "id"           "price_date"   "price_p1_var" "price_p2_var" "price_p3_var"
## [6] "price_p1_fix" "price_p2_fix" "price_p3_fix"
```

```
lapply(hist, class)
```

```
## $id
## [1] "character"
##
## $price_date
## [1] "character"
##
## $price_p1_var
## [1] "numeric"
##
## $price_p2_var
## [1] "numeric"
##
```

```
## $price_p3_var
## [1] "numeric"
##
## $price_p1_fix
## [1] "numeric"
##
## $price_p2_fix
## [1] "numeric"
##
## $price_p3_fix
## [1] "numeric"
```

```
### Convert Data Type
train$date_activ <- as.Date(train$date_activ, "%Y/%m/%d")
train$date_end <- as.Date(train$date_end, "%Y/%m/%d")
train$date_first_activ <- as.Date(train$date_first_activ, "%Y/%m/%d")
train$date_modif_prod <- as.Date(train$date_modif_prod, "%Y/%m/%d")
train$date_renewal <- as.Date(train$date_renewal, "%Y/%m/%d")
train$has_gas <- as.logical(toupper(train$has_gas))
train$churn <- as.logical(train$churn)

train$activity_new <- as.factor(train$activity_new)
train$channel_sales<- as.factor(train$channel_sales)
train$origin_up <- as.factor(train$origin_up)

hist$price_date <- as.Date(hist$price_date, "%Y/%m/%d")
```

3 Missing values disposal

Also, it is obvious that there are tons of missing values in the dataset. We will see how often NAs appear in a variable. If the proportion of NAs for a variable is way too large, then it is hard to fill them with estimates and we might need to delete them (we notice that the missing rates of some explanatory variables are over 78% which means they will contribute little to our prediction model thus ignoring).

```
colMeans(is.na(train))
```

```
##                       id             activity_new        campaign_disc_ele
##             0.0000000000             0.0000000000             1.0000000000
##            channel_sales                 cons_12m             cons_gas_12m
##             0.0000000000             0.0000000000             0.0000000000
##          cons_last_month               date_activ                 date_end
##             0.0000000000             0.0000000000             0.0001242545
##         date_first_activ          date_modif_prod             date_renewal
##             0.7820576541             0.0097539761             0.0024850895
##    forecast_base_bill_ele  forecast_base_bill_year        forecast_bill_12m
##             0.7820576541             0.7820576541             0.7820576541
##            forecast_cons        forecast_cons_12m       forecast_cons_year
##             0.7820576541             0.0000000000             0.0000000000
## forecast_discount_energy  forecast_meter_rent_12m forecast_price_energy_p1
##             0.0078280318             0.0000000000             0.0078280318
## forecast_price_energy_p2    forecast_price_pow_p1                  has_gas
##             0.0078280318             0.0078280318             0.0000000000
```

```
##                imp_cons     margin_gross_pow_ele       margin_net_pow_ele 
##            0.0000000000             0.0008076541             0.0008076541 
##             nb_prod_act               net_margin          num_years_antig 
##            0.0000000000             0.0009319085             0.0000000000 
##               origin_up                  pow_max                    churn 
##            0.0000000000             0.0001863817             0.0000000000
```

```r
colMeans(is.na(hist))
```

```
##          id   price_date price_p1_var price_p2_var price_p3_var price_p1_fix 
##  0.000000000  0.000000000  0.007041378  0.007041378  0.007041378  0.007041378 
## price_p2_fix price_p3_fix 
##  0.007041378  0.007041378
```

```r
train_rm <- which(colMeans(is.na(train)) > 0.5)
train <- train[, -train_rm]
```

After deleting those variables with too many NAs, there are still some NAs in our training dataset and they appear in the following explanatory variables:

```r
names(train)[unique(ceiling(which(is.na(train))/nrow(train)))]
```

```
##  [1] "date_end"                  "date_modif_prod"          
##  [3] "date_renewal"              "forecast_discount_energy" 
##  [5] "forecast_price_energy_p1" "forecast_price_energy_p2"
##  [7] "forecast_price_pow_p1"    "margin_gross_pow_ele"     
##  [9] "margin_net_pow_ele"       "net_margin"               
## [11] "pow_max"
```

We can replace the NAs with some specific values, for example, in ‘forecast_discount_energy’ we can replace all NAs with zeros. However, it can be extremely hard to do this kind of replacement if no other information is given, and there is only a fairly small proportion that have NAs, as a result we can simply ignore these items with NAs when building a regression model using the code below:

```r
### Inspect the id with NAs
id.rm <- train[rowSums(is.na(train)) > 0, 1]

### Delete the items with NAs in training dataset
# train[rowSums(is.na(train)) > 0,]
train_new <- na.omit(train)

### Delete the items with corresponding id in historical dataset
hist_new <- hist[!(hist$id %in% id.rm), ]
```

In the historical dataset we notice that some records for an id are incomplete as the number of items are not multiples of 12 (number of months) thus for some ids the historical data is missing and omitted in the dataset. We can scan the whole dataset and add the missing item or replace NAs using the nearest item without NAs under the same id. In this case, we assume that every id has at least one piece of data without NAs. The code is as follows:

```r
scan_i = 1
while (!is.na(hist_new$id[scan_i]))
{
  if (month(hist_new$price_date[scan_i]) %% 12 != scan_i %% 12){
    if (hist_new$id[scan_i-1] == hist_new$id[scan_i]){
      hist_new <- hist_new %>% add_row(hist_new[scan_i - 1, ], .before = scan_i)
    }
    else {
      hist_new <- hist_new %>% add_row(hist_new[scan_i, ], .before = scan_i)
    }
    hist_new$price_date[scan_i] <- hist_new$price_date[scan_i] %m+%
      months(scan_i %% 12 - month(hist_new$price_date[scan_i]))
  }

  if (sum(is.na(hist_new[scan_i, ])) > 0){
    if (hist_new$id[scan_i-1] == hist_new$id[scan_i]){
      hist_new[scan_i, 3:8] <- hist_new[scan_i-1, 3:8]
    }
    else {
      for (k in 1:12) {
        if (sum(is.na(hist_new[scan_i+k, 3:8])) == 0){
          hist_new[scan_i, 3:8] <- hist_new[scan_i+k, 3:8]
          break
        }
      }
    }
  }

  scan_i = scan_i + 1
}
```

Then we can check if there is still NAs in our dataset as below:

```r
sum(is.na(hist_new))
```

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

```r
sum(is.na(train_new))
```

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

We see that both training and historical dataset have no NAs now. Next, to test multicollinearity, we can first have a look at the correlation matrix of numeric variables.

```r
train_num <- unlist(lapply(train_new, is.numeric))
X <- train[, train_num]
cor(X, use = "complete.obs")
```

```
##                            cons_12m cons_gas_12m cons_last_month
## cons_12m                1.000000000  0.488639230     0.923357459
## cons_gas_12m            0.488639230  1.000000000     0.464456231
```

```
## cons_last_month           0.923357459  0.464456231     1.000000000
## forecast_cons_12m         0.164916093  0.061428111     0.129685057
## forecast_cons_year        0.138254239  0.059929319     0.150981555
## forecast_discount_energy -0.043470661 -0.014908903    -0.037684947
## forecast_meter_rent_12m   0.086705187  0.040539468     0.076871368
## forecast_price_energy_p1 -0.032599995 -0.022369658    -0.024371806
## forecast_price_energy_p2  0.145845865  0.078459832     0.122989335
## forecast_price_pow_p1    -0.024630712 -0.027108306    -0.019935599
## imp_cons                  0.137844493  0.062968115     0.153213034
## margin_gross_pow_ele     -0.065145711 -0.016609323    -0.053945957
## margin_net_pow_ele       -0.045344236 -0.007848937    -0.037441976
## nb_prod_act               0.310708444  0.280249558     0.351882086
## net_margin                0.120881235  0.060649042     0.096480195
## num_years_antig           0.008039043 -0.009534517     0.004882561
## pow_max                   0.105807819  0.055446073     0.092438791
##                          forecast_cons_12m forecast_cons_year
## cons_12m                        0.16491609        0.138254239
## cons_gas_12m                    0.06142811        0.059929319
## cons_last_month                 0.12968506        0.150981555
## forecast_cons_12m               1.00000000        0.743786488
## forecast_cons_year              0.74378649        1.000000000
## forecast_discount_energy        0.01506816       -0.008918112
## forecast_meter_rent_12m         0.38712189        0.325051334
## forecast_price_energy_p1       -0.21741232       -0.206128257
## forecast_price_energy_p2        0.24589115        0.225683161
## forecast_price_pow_p1           0.05851540        0.053956333
## imp_cons                        0.72352956        0.981773589
## margin_gross_pow_ele           -0.18472997       -0.138633105
## margin_net_pow_ele             -0.14173228       -0.105867732
## nb_prod_act                     0.01297959        0.014179346
## net_margin                      0.76900748        0.536095617
## num_years_antig                 0.06095507        0.062367444
## pow_max                         0.58669180        0.443257714
##                          forecast_discount_energy forecast_meter_rent_12m
## cons_12m                             -0.043470661            0.0867051875
## cons_gas_12m                         -0.014908903            0.0405394680
## cons_last_month                      -0.037684947            0.0768713677
## forecast_cons_12m                     0.015068163            0.3871218890
## forecast_cons_year                   -0.008918112            0.3250513337
## forecast_discount_energy              1.000000000           -0.0195882411
## forecast_meter_rent_12m              -0.019588241            1.0000000000
## forecast_price_energy_p1              0.319305487           -0.5584395812
## forecast_price_energy_p2              0.048915407            0.6368692475
## forecast_price_pow_p1                 0.024658837            0.0126578433
## imp_cons                              0.011473681            0.2926968549
## margin_gross_pow_ele                  0.199597850           -0.0171639123
## margin_net_pow_ele                    0.151127109            0.0025813212
## nb_prod_act                           0.055249982           -0.0001314331
## net_margin                            0.013499604            0.3334363555
## num_years_antig                      -0.071535467            0.1090391736
## pow_max                              -0.022635903            0.6074057450
##                          forecast_price_energy_p1 forecast_price_energy_p2
## cons_12m                              -0.03260000               0.14584586
## cons_gas_12m                          -0.02236966               0.07845983
```

```
## cons_last_month                    -0.02437181            0.12298934
## forecast_cons_12m                  -0.21741232            0.24589115
## forecast_cons_year                 -0.20612826            0.22568316
## forecast_discount_energy            0.31930549            0.04891541
## forecast_meter_rent_12m            -0.55843958            0.63686925
## forecast_price_energy_p1            1.00000000           -0.36471981
## forecast_price_energy_p2           -0.36471981            1.00000000
## forecast_price_pow_p1               0.39002395           -0.13738577
## imp_cons                           -0.16475737            0.21108625
## margin_gross_pow_ele                0.18462838            0.06350806
## margin_net_pow_ele                  0.02896633            0.07414584
## nb_prod_act                         0.02587751            0.02601578
## net_margin                         -0.18522125            0.25176133
## num_years_antig                    -0.19960786            0.10270809
## pow_max                            -0.35259508            0.33936398
##                          forecast_price_pow_p1    imp_cons margin_gross_pow_ele
## cons_12m                          -0.024630712  0.13784449          -0.06514571
## cons_gas_12m                      -0.027108306  0.06296811          -0.01660932
## cons_last_month                   -0.019935599  0.15321303          -0.05394596
## forecast_cons_12m                  0.058515397  0.72352956          -0.18472997
## forecast_cons_year                 0.053956333  0.98177359          -0.13863311
## forecast_discount_energy           0.024658837  0.01147368           0.19959785
## forecast_meter_rent_12m            0.012657843  0.29269685          -0.01716391
## forecast_price_energy_p1           0.390023949 -0.16475737           0.18462838
## forecast_price_energy_p2          -0.137385771  0.21108625           0.06350806
## forecast_price_pow_p1              1.000000000  0.05178806          -0.11453867
## imp_cons                           0.051788060  1.00000000          -0.12163984
## margin_gross_pow_ele              -0.114538672 -0.12163984           1.00000000
## margin_net_pow_ele                -0.133985399 -0.09164210           0.76459102
## nb_prod_act                       -0.011325194  0.01947712          -0.04407760
## net_margin                        -0.005512614  0.53543201          -0.09958930
## num_years_antig                   -0.038312049  0.04768462          -0.07948924
## pow_max                            0.051587889  0.40925151          -0.01121888
##                          margin_net_pow_ele   nb_prod_act   net_margin
## cons_12m                       -0.045344236  0.3107084436  0.120881235
## cons_gas_12m                   -0.007848937  0.2802495579  0.060649042
## cons_last_month                -0.037441976  0.3518820857  0.096480195
## forecast_cons_12m              -0.141732276  0.0129795855  0.769007476
## forecast_cons_year             -0.105867732  0.0141793459  0.536095617
## forecast_discount_energy        0.151127109  0.0552499816  0.013499604
## forecast_meter_rent_12m         0.002581321 -0.0001314331  0.333436355
## forecast_price_energy_p1        0.028966335  0.0258775054 -0.185221248
## forecast_price_energy_p2        0.074145839  0.0260157791  0.251761334
## forecast_price_pow_p1          -0.133985399 -0.0113251941 -0.005512614
## imp_cons                       -0.091642098  0.0194771188  0.535432012
## margin_gross_pow_ele            0.764591018 -0.0440775986 -0.099589296
## margin_net_pow_ele              1.000000000 -0.0323571285 -0.087147864
## nb_prod_act                    -0.032357129  1.0000000000  0.004143330
## net_margin                     -0.087147864  0.0041433302  1.000000000
## num_years_antig                -0.035800707  0.0094058034  0.033355848
## pow_max                         0.000969799  0.0187447972  0.452370443
##                          num_years_antig      pow_max
## cons_12m                     0.008039043  0.105807819
## cons_gas_12m                -0.009534517  0.055446073
```

```
## cons_last_month             0.004882561  0.092438791
## forecast_cons_12m           0.060955073  0.586691797
## forecast_cons_year          0.062367444  0.443257714
## forecast_discount_energy   -0.071535467 -0.022635903
## forecast_meter_rent_12m     0.109039174  0.607405745
## forecast_price_energy_p1   -0.199607862 -0.352595076
## forecast_price_energy_p2    0.102708095  0.339363985
## forecast_price_pow_p1      -0.038312049  0.051587889
## imp_cons                    0.047684617  0.409251515
## margin_gross_pow_ele       -0.079489241 -0.011218882
## margin_net_pow_ele         -0.035800707  0.000969799
## nb_prod_act                 0.009405803  0.018744797
## net_margin                  0.033355848  0.452370443
## num_years_antig             1.000000000  0.079751325
## pow_max                     0.079751325  1.000000000
```

We see that there are quite a few large correlation coefficients greater than 0.9, for example, 'cons_12m' seems to be highly positive correlated with 'cons_last_month' with r = 0.9713. In our further regression models we need to take this into consideration and run a VIF test to confirm multicollinearity, then remove variables until all VIF scores are relatively low (e.g. < 4).

And here we have the summary of our pretreated dataset:

```
summary(hist_new)
```

```
##       id               price_date          price_p1_var     price_p2_var
##  Length:189132      Min.   :2014-12-01   Min.   :0.0000   Min.   :0.00000
##  Class :character   1st Qu.:2015-03-01   1st Qu.:0.1260   1st Qu.:0.00000
##  Mode  :character   Median :2015-06-01   Median :0.1460   Median :0.08547
##                     Mean   :2015-06-16   Mean   :0.1410   Mean   :0.05438
##                     3rd Qu.:2015-09-01   3rd Qu.:0.1516   3rd Qu.:0.10178
##                     Max.   :2015-12-01   Max.   :0.2807   Max.   :0.22979
##   price_p3_var      price_p1_fix      price_p2_fix       price_p3_fix
##  Min.   :0.00000   Min.   :-0.1778   Min.   :-0.09775   Min.   :-0.06517
##  1st Qu.:0.00000   1st Qu.:40.7289   1st Qu.: 0.00000   1st Qu.: 0.00000
##  Median :0.00000   Median :44.2669   Median : 0.00000   Median : 0.00000
##  Mean   :0.03071   Mean   :43.3258   Mean   :10.69347   Mean   : 6.45457
##  3rd Qu.:0.07256   3rd Qu.:44.4447   3rd Qu.:24.33958   3rd Qu.:16.22639
##  Max.   :0.11410   Max.   :59.4447   Max.   :36.49069   Max.   :17.45822
```

```
summary(train_new)
```

```
##       id                                       activity_new
##  Length:15761                                        :9360
##  Class :character   apdekpcbwosbxepsfxclislboipuxpop:1532
##  Mode  :character   kkklcdamwfafdcfwofuscwfwadblfmce: 420
##                     kwuslieomapmswolewpobpplkaooaaew: 226
##                     fmwdwsxillemwbbwelxsampiuwwpcdcb: 216
##                     ckfxocssowaeipxueikxcmaxdmcduxsa: 187
##                     (Other)                         :3820
##                           channel_sales     cons_12m         cons_gas_12m
##  foosdfpfkusacimwkcsosbicdxkicaua:7151   Min.   : -125276   Min.   :  -3037
##                                  :4177   1st Qu.:    5886   1st Qu.:      0
```

```
##  lmkebamcaaclubfxadlmueccxoimlema:2052   Median :   15215   Median :      0
##  usilxuppasemubllopkaafesmlibmsdf:1418   Mean   :  191318   Mean   :  31375
##  ewpakwlliwisiwduibdlfmalxowmwpci: 949   3rd Qu.:   49524   3rd Qu.:      0
##  sddiedcslfslkckwlfkdpoeeailfpeds:  10   Max.   :16097108   Max.   :4154590
##  (Other)                         :   4
##  cons_last_month     date_activ            date_end
##  Min.   : -91386   Min.   :2000-07-25   Min.   :2013-05-06
##  1st Qu.:      0   1st Qu.:2010-01-11   1st Qu.:2016-04-27
##  Median :    896   Median :2011-02-21   Median :2016-08-01
##  Mean   :  19263   Mean   :2011-01-08   Mean   :2016-07-28
##  3rd Qu.:   4104   3rd Qu.:2012-04-17   3rd Qu.:2016-11-01
##  Max.   :4538720   Max.   :2014-09-01   Max.   :2017-06-13
##
##  date_modif_prod       date_renewal        forecast_cons_12m
##  Min.   :2000-07-25   Min.   :2013-06-26   Min.   :-16689.3
##  1st Qu.:2010-08-05   1st Qu.:2015-04-17   1st Qu.:   512.4
##  Median :2013-04-25   Median :2015-07-27   Median :  1177.4
##  Mean   :2012-12-11   Mean   :2015-07-21   Mean   :  2354.9
##  3rd Qu.:2015-05-24   3rd Qu.:2015-10-30   3rd Qu.:  2680.3
##  Max.   :2016-01-29   Max.   :2016-01-28   Max.   :103801.9
##
##  forecast_cons_year forecast_discount_energy forecast_meter_rent_12m
##  Min.   :-85627     Min.   : 0.0000          Min.   :-242.96
##  1st Qu.:     0     1st Qu.: 0.0000          1st Qu.:  16.23
##  Median :   376     Median : 0.0000          Median :  19.44
##  Mean   :  1895     Mean   : 0.9792          Mean   :  70.34
##  3rd Qu.:  1993     3rd Qu.: 0.0000          3rd Qu.: 131.51
##  Max.   :175375     Max.   :50.0000          Max.   :2411.69
##
##  forecast_price_energy_p1 forecast_price_energy_p2 forecast_price_pow_p1
##  Min.   :0.0000           Min.   :0.00000          Min.   :-0.1222
##  1st Qu.:0.1152           1st Qu.:0.00000          1st Qu.:40.6067
##  Median :0.1429           Median :0.08616          Median :44.3114
##  Mean   :0.1359           Mean   :0.05291          Mean   :43.5334
##  3rd Qu.:0.1463           3rd Qu.:0.09884          3rd Qu.:44.3114
##  Max.   :0.2740           Max.   :0.19598          Max.   :59.4447
##
##   has_gas           imp_cons        margin_gross_pow_ele margin_net_pow_ele
##  Mode :logical   Min.   :-9038.21   Min.   :-525.54      Min.   :-615.66
##  FALSE:12864     1st Qu.:    0.00   1st Qu.:  11.95      1st Qu.:  11.88
##  TRUE :2897      Median :   44.04   Median :  20.95      Median :  20.80
##                  Mean   :  194.80   Mean   :  22.41      Mean   :  21.39
##                  3rd Qu.:  217.59   3rd Qu.:  29.64      3rd Qu.:  29.64
##                  Max.   :15042.79   Max.   : 374.64      Max.   : 374.64
##
##   nb_prod_act       net_margin       num_years_antig
##  Min.   : 1.000   Min.   :-4148.99   Min.   : 1.000
##  1st Qu.: 1.000   1st Qu.:   51.97   1st Qu.: 4.000
##  Median : 1.000   Median :  119.44   Median : 5.000
##  Mean   : 1.348   Mean   :  217.73   Mean   : 5.051
##  3rd Qu.: 1.000   3rd Qu.:  274.95   3rd Qu.: 6.000
##  Max.   :32.000   Max.   :24570.65   Max.   :16.000
##
##                              origin_up       pow_max         churn
```

```
##                                   :  87   Min.   :  1.00   Mode :logical
##  ewxeelcelemmiwuafmddpobolfuxioce:   1   1st Qu.: 12.50   FALSE:14236
##  kamkkxfxxuwbdslkwifmmcsiusiuosws:4489   Median : 13.86   TRUE :1525
##  ldkssxwpmemidmecebumciepifcamkci:3592   Mean   : 20.50
##  lxidpiddsbxsbosboudacockeimpuepw:7590   3rd Qu.: 19.80
##  usapbepcfoloekilkwsdiboslwaxobdp:   2   Max.   :500.00
##
```

It is wired that some data which are supposed to be positive has some negative values, for example, prices/consumption should probably be positive. Here we assume that somehow we added a negative symbol by mistake, so we scan the two datasets and modify them to positive (or change to zero, more information needed. Besides, we can combine this scan with the above one to save time).

```
scan_i = 1
for (scan_i in 1:nrow(hist_new)) {
  for (j in 3:8){
    if (hist_new[scan_i, j] < 0) hist_new[scan_i, j] = - hist_new[scan_i, j]
  }
}

### seems all numeric variables in training dataset are strictly positive
### more information needed
train_num <- which(unlist(lapply(train_new, is.numeric)))
scan_i = 1
for (scan_i in 1: nrow(train_new)) {
  for (j in train_num) {
    if (train_new[scan_i, j] < 0) train_new[scan_i, j] = - train_new[scan_i, j]
  }
}
```

The updated summary is as follows:

```
summary(hist_new)
```

```
##       id               price_date          price_p1_var     price_p2_var
##  Length:189132      Min.   :2014-12-01   Min.   :0.0000   Min.   :0.00000
##  Class :character   1st Qu.:2015-03-01   1st Qu.:0.1260   1st Qu.:0.00000
##  Mode  :character   Median :2015-06-01   Median :0.1460   Median :0.08547
##                     Mean   :2015-06-16   Mean   :0.1410   Mean   :0.05438
##                     3rd Qu.:2015-09-01   3rd Qu.:0.1516   3rd Qu.:0.10178
##                     Max.   :2015-12-01   Max.   :0.2807   Max.   :0.22979
##   price_p3_var      price_p1_fix    price_p2_fix    price_p3_fix
##  Min.   :0.00000   Min.   : 0.00   Min.   : 0.00   Min.   : 0.000
##  1st Qu.:0.00000   1st Qu.:40.73   1st Qu.: 0.00   1st Qu.: 0.000
##  Median :0.00000   Median :44.27   Median : 0.00   Median : 0.000
##  Mean   :0.03071   Mean   :43.33   Mean   :10.69   Mean   : 6.455
##  3rd Qu.:0.07256   3rd Qu.:44.44   3rd Qu.:24.34   3rd Qu.:16.226
##  Max.   :0.11410   Max.   :59.44   Max.   :36.49   Max.   :17.458
```

```
summary(train_new)
```

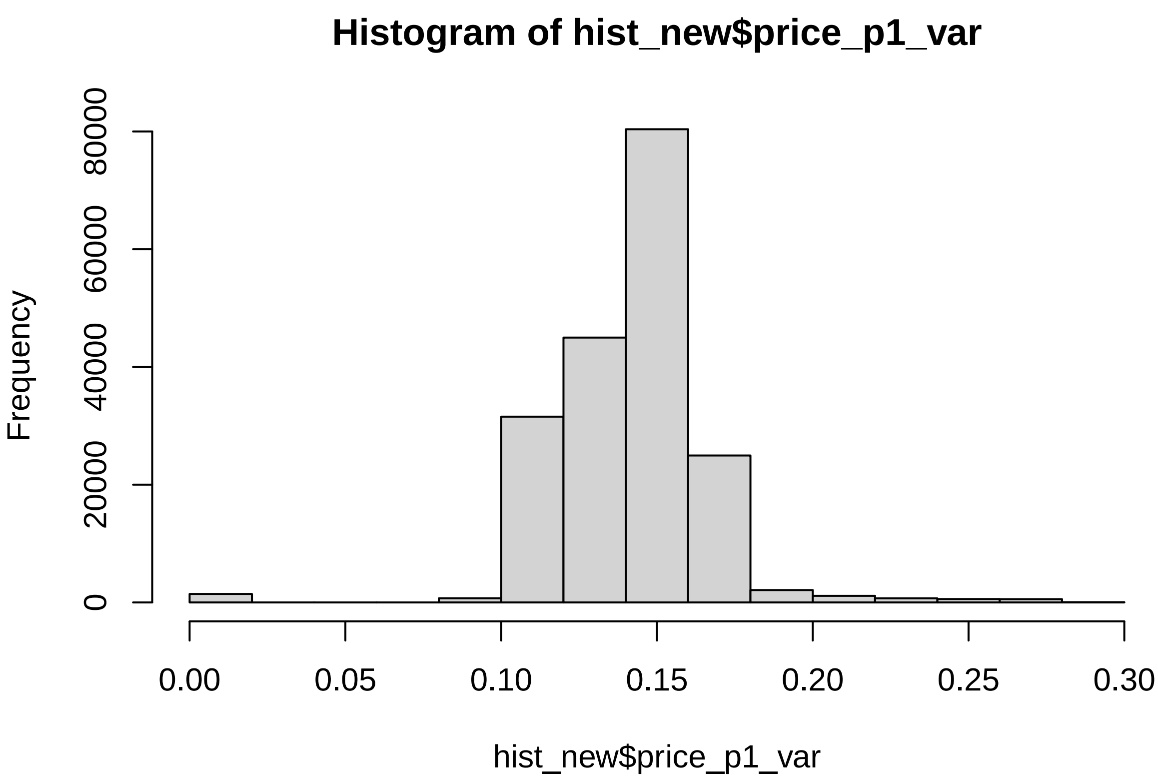
```
##       id                                       activity_new
```

```
##  Length:15761                                      :9360
##  Class :character   apdekpcbwosbxepsfxclislboipuxpop:1532
##  Mode  :character   kkklcdamwfafdcfwofuscwfwadblfmce: 420
##                     kwuslieomapmswolewpobpplkaooaaew: 226
##                     fmwdwsxillemwbbwelxsampiuwwpcdcb: 216
##                     ckfxocssowaeipxueikxcmaxdmcduxsa: 187
##                     (Other)                         :3820
##                            channel_sales     cons_12m         cons_gas_12m
##  foosdfpfkusacimwkcsosbicdxkicaua:7151   Min.   :       0   Min.   :      0
##                                  :4177   1st Qu.:    5896   1st Qu.:      0
##  lmkebamcaaclubfxadlmueccxoimlema:2052   Median :   15257   Median :      0
##  usilxuppasemubllopkaafesmlibmsdf:1418   Mean   :  191379   Mean   :  31376
##  ewpakwlliwisiwduibdlfmalxowmwpci: 949   3rd Qu.:   49590   3rd Qu.:      0
##  sddiedcslfslkckwlfkdpoeeailfpeds:  10   Max.   :16097108   Max.   :4154590
##  (Other)                         :   4
##  cons_last_month     date_activ           date_end
##  Min.   :      0   Min.   :2000-07-25   Min.   :2013-05-06
##  1st Qu.:      0   1st Qu.:2010-01-11   1st Qu.:2016-04-27
##  Median :    910   Median :2011-02-21   Median :2016-08-01
##  Mean   :  19381   Mean   :2011-01-08   Mean   :2016-07-28
##  3rd Qu.:   4142   3rd Qu.:2012-04-17   3rd Qu.:2016-11-01
##  Max.   :4538720   Max.   :2014-09-01   Max.   :2017-06-13
##
##  date_modif_prod       date_renewal       forecast_cons_12m
##  Min.   :2000-07-25   Min.   :2013-06-26   Min.   :     0.0
##  1st Qu.:2010-08-05   1st Qu.:2015-04-17   1st Qu.:   514.2
##  Median :2013-04-25   Median :2015-07-27   Median :  1179.8
##  Mean   :2012-12-11   Mean   :2015-07-21   Mean   :  2364.4
##  3rd Qu.:2015-05-24   3rd Qu.:2015-10-30   3rd Qu.:  2685.8
##  Max.   :2016-01-29   Max.   :2016-01-28   Max.   :103801.9
##
##  forecast_cons_year forecast_discount_energy forecast_meter_rent_12m
##  Min.   :      0    Min.   : 0.0000          Min.   :   0.00
##  1st Qu.:      0    1st Qu.: 0.0000          1st Qu.:  16.23
##  Median :    383    Median : 0.0000          Median :  19.44
##  Mean   :   1918    Mean   : 0.9792          Mean   :  70.38
##  3rd Qu.:   1999    3rd Qu.: 0.0000          3rd Qu.: 131.52
##  Max.   :175375     Max.   :50.0000          Max.   :2411.69
##
##  forecast_price_energy_p1 forecast_price_energy_p2 forecast_price_pow_p1
##  Min.   :0.0000           Min.   :0.00000          Min.   : 0.00
##  1st Qu.:0.1152           1st Qu.:0.00000          1st Qu.:40.61
##  Median :0.1429           Median :0.08616          Median :44.31
##  Mean   :0.1359           Mean   :0.05291          Mean   :43.53
##  3rd Qu.:0.1463           3rd Qu.:0.09884          3rd Qu.:44.31
##  Max.   :0.2740           Max.   :0.19598          Max.   :59.44
##
##   has_gas            imp_cons         margin_gross_pow_ele margin_net_pow_ele
##  Mode :logical    Min.   :    0.00   Min.   :  0.00       Min.   :  0.00
##  FALSE:12864      1st Qu.:    0.00   1st Qu.: 12.36       1st Qu.: 12.36
##  TRUE :2897       Median :   44.94   Median : 21.09       Median : 21.09
##                   Mean   :  197.28   Mean   : 23.58       Mean   : 24.15
##                   3rd Qu.:  218.25   3rd Qu.: 29.64       3rd Qu.: 29.76
##                   Max.   :15042.79   Max.   :525.54       Max.   :615.66
```

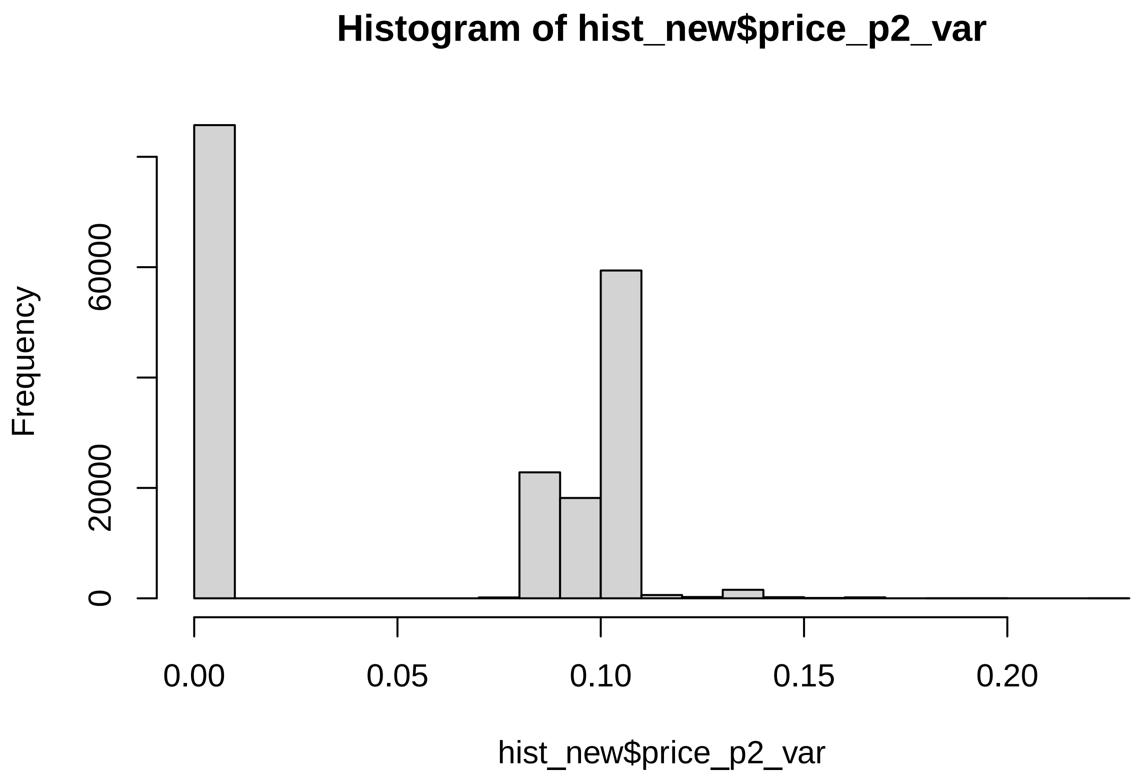
```
##
##    nb_prod_act       net_margin       num_years_antig
##   Min.   : 1.000   Min.   :    0.00   Min.   : 1.000
##   1st Qu.: 1.000   1st Qu.:   52.83   1st Qu.: 4.000
##   Median : 1.000   Median :  120.75   Median : 5.000
##   Mean   : 1.348   Mean   :  221.69   Mean   : 5.051
##   3rd Qu.: 1.000   3rd Qu.:  276.27   3rd Qu.: 6.000
##   Max.   :32.000   Max.   :24570.65   Max.   :16.000
##
##                               origin_up       pow_max          churn
##                                    :  87   Min.   :  1.00   Mode :logical
##   ewxeelcelemmiwuafmddpobolfuxioce:   1   1st Qu.: 12.50   FALSE:14236
##   kamkkxfxxuwbdslkwifmmcsiusiuosws:4489   Median : 13.86   TRUE :1525
##   ldkssxwpmemidmecebumciepifcamkci:3592   Mean   : 20.50
##   lxidpiddsbxsbosboudacockeimpuepw:7590   3rd Qu.: 19.80
##   usapbepcfoloekilkwsdiboslwaxobdp:   2   Max.   :500.00
##
```

Now we are going to visualize the data to have a general idea about how they distribute. For example, we would like to know how the distribution of historical prices look like, the figures for the prices of energy/power for the 1st/2rd/3rd periods are listed below:

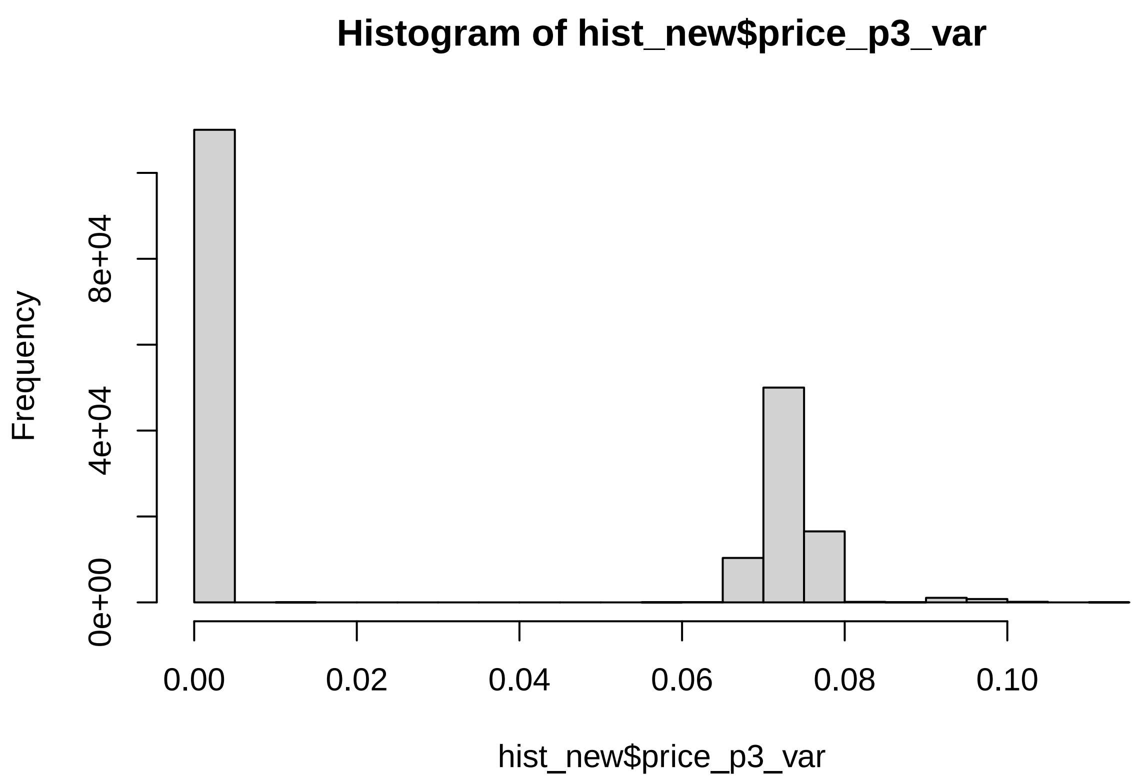
```
hist(hist_new$price_p1_var)
```



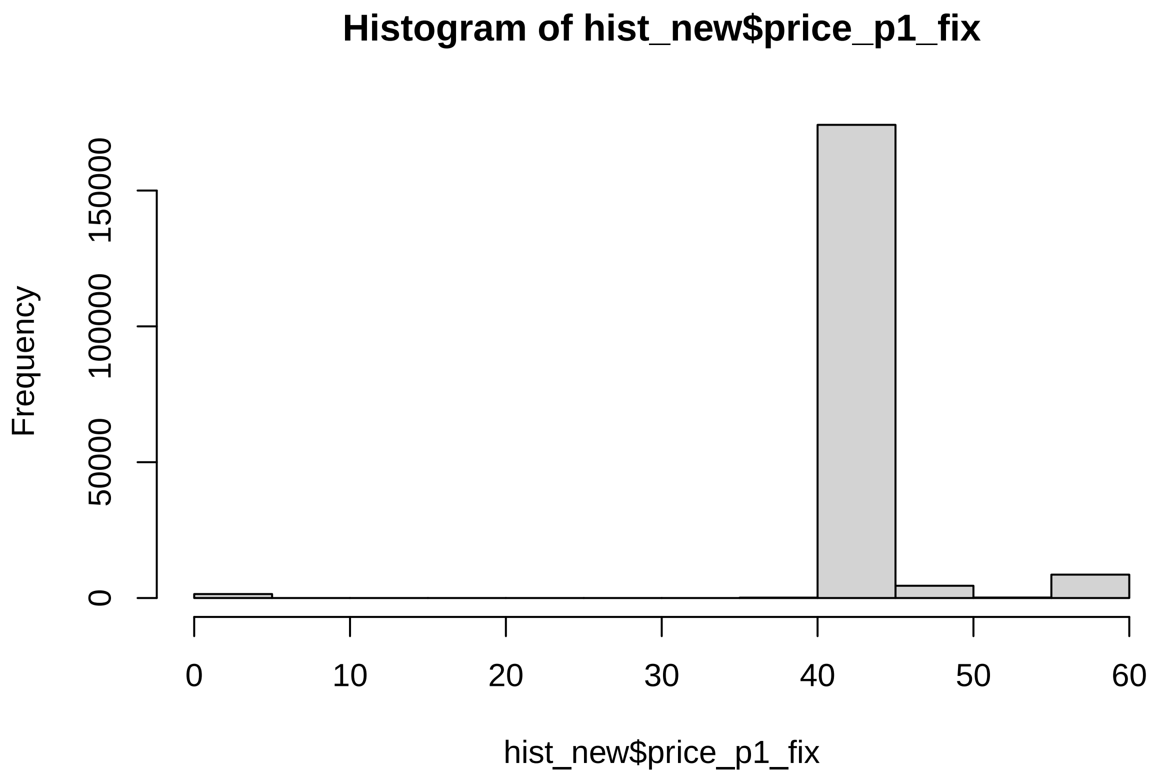

```
hist(hist_new$price_p2_var)
```



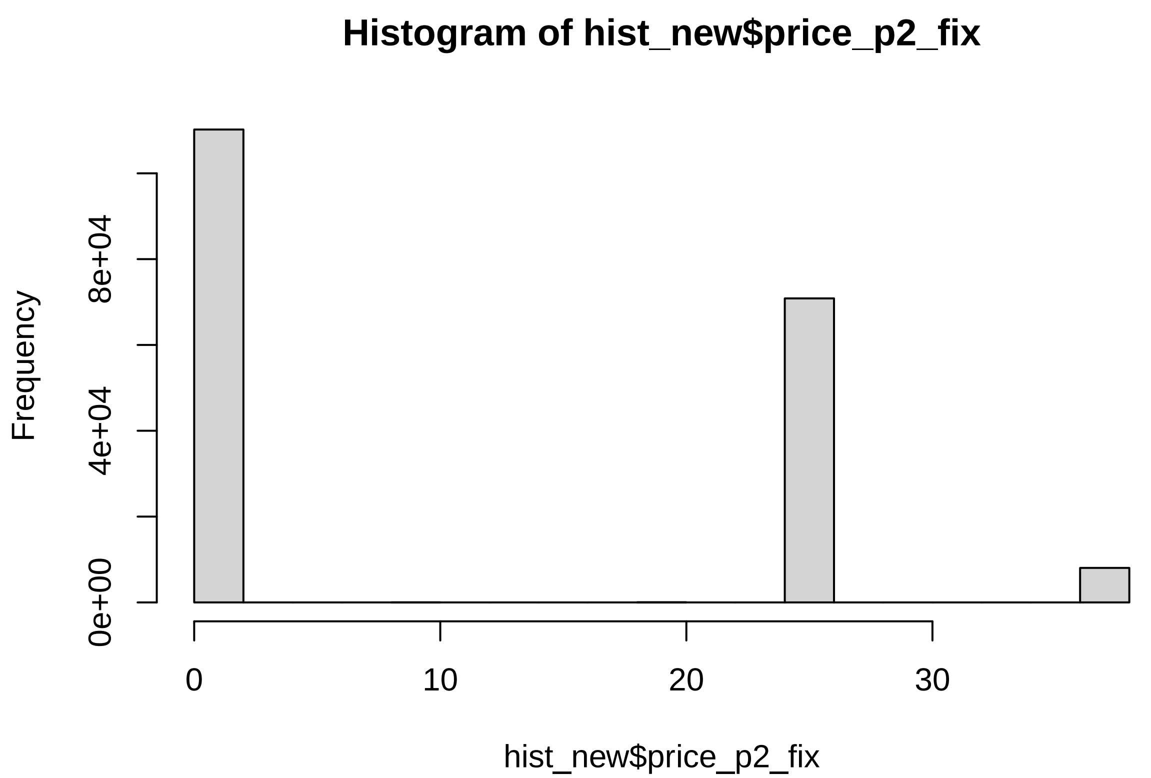

```
hist(hist_new$price_p3_var)
```



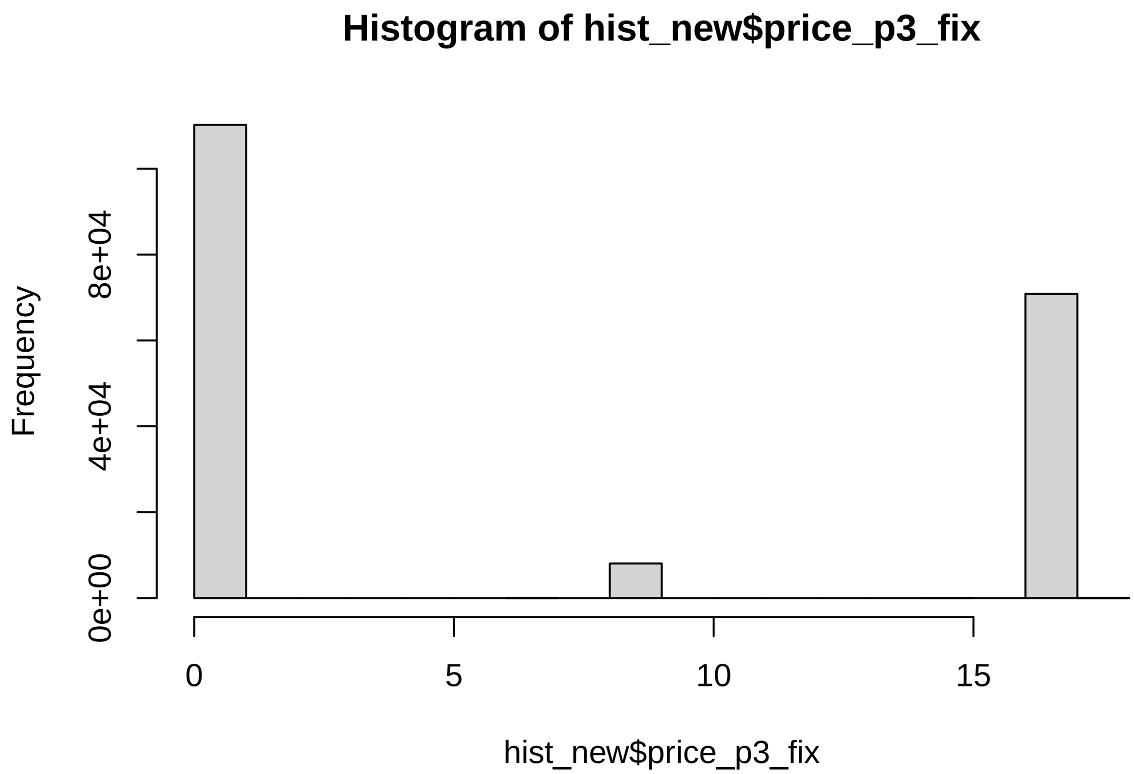
```
hist(hist_new$price_p1_fix)
```




```
hist(hist_new$price_p2_fix)
```



```
hist(hist_new$price_p3_fix)
```



For example, we see that thedistribution of prices of energy for the 1st period is unimodal with most prices concentrate at around 0.15, and very few price is close to 0 (might due to errors, as I don't believe there is such cheap price compared to others). Also, in training dataset, we can see the distribution of category of the company's activity, code of the sales channel and so on (issues showing the axis labels can be fixed by shorten the encrypted code, rename the variables or restyle the labelling area).

```
summary(train_new$activity_new)
```

```
##                                  apdekpcbwosbxepsfxclislboipuxpop
##                             9360                             1532
## kkklcdamwfafdcfwofuscwfwadblfmce kwuslieomapmswolewpobpplkaooaaew
##                              420                              226
## fmwdwsxillemwbbwelxsampiuwwpcdcb ckfxocssowaeipxueikxcmaxdmcduxsa
##                              216                              187
## cwofmuicebbcmiaaxufmfimpowpacobu wxemiwkumpibllwklfbcooafckufkdlm
##                              120                              117
## cluecxlameloamldmasudocsbmaoamdw sfisfxfcocfpcmckuekokxuseixdaoeu
##                              115                               81
## sffadmsbuamddwapeumdfibkmpkdicmc sxublbwoeuckkocekklxkllcdxxaisop
##                               75                               72
## dupxuibdflmskeieweeofcaluuuiioix ipdldckuswupeifllfbwccfpeafludfi
##                               61                               61
## saxlifeumaobawxpemwuopbwwldlucff daobdssbkieoukwxbopxiiospudkopwl
##                               61                               57
## ibkiiwcxiccxpoedpweiuxwbxbuewbxm cfdsselwimsklimddecfifseabdkxfcs
##                               55                               50
```

```
## bwpaswkpcilmlklklcapcwwumwaodaoo ilkfsaapsxpkcpswbllddfmpamwelpxi
##                               48                               47
## bxopwkbwdewxssbmkwcummkaakbwafxf mpicaaibskkfmxoblmwwwuuwpkecacil
##                               46                               44
## balskueexlmuccwdffilikwxasupasxf sumdxiaiudmaioicexmiwuudlblkissm
##                               43                               43
## ppcxfxbffsxaakxamcdpexdoxulfwwae lkeudbeowbapkpfodoxacpwdpaeuwxcx
##                               39                               38
## ddkpdekmbfdffwdmabkiiilolsxswccl axsupumdipebmlbiwolspmkdouoiddbc
##                               37                               34
## kmxccaddbdpaaolkbidlobeefsbbcxca ckadsdebplpkplelfspfoiucmxkeppus
##                               34                               33
## fkmblacmaapkaoauabpwpuweokkeiali dfcsaaowsemmabpepocaeaaecfwppxxk
##                               33                               31
## fcbfabofwcdaosksieduepeeusawfdsi xcswxcciuaxpacidfbixxmwkdmmskxkc
##                               29                               29
## cssldxpacdmuuaulamxdekcokibauube kkpddsilciodwwwffucmkflilcpfaumo
##                               28                               28
## oclfuccbxapuklpeowbabcpawcbwxesk pmedwkpuckbppeoecxiccwxluwkxdkpe
##                               28                               28
## cccpsslxcemdlomsaffxsecccbxpdkax duiwascsdupcmdfkspbukuuaklsawmmc
##                               26                               26
## fibkpxbliefxfmeielcidsckcxkpofaa ifppdlcfssupdcsdcclkoubulccouwml
##                               26                               26
## wdkbuxwfkbefwplcoudfalpfafdfpfax clbkplmouokdpxiwxebwculxxsdiuwap
##                               26                               25
## almlfkoedpwfdmmsebsdwueskducuiok dwamuluiuaiowuxmesuuilkbobidcmfo
##                               24                               24
## edxmolisbfbwlpmccduowkxpkiiooess pbpfffswspwswuxudcdibsmdkpokflmi
##                               24                               24
## uiouuawillpcssldoeemcddcpfseebsw sscfoipxikopfskekuobeuxkxmwsuucb
##                               24                               23
## alkuukubieaxcobeeowowmokpbilomax acefxcckbdxakciukwuwepweawbkwmii
##                               22                               21
## cxdlpsmkulssdwsoskdmisdmdbcuebww pffpiboilxxdeluedfxssmaklbdplfmi
##                               21                               21
## piaosodsowlfpxipbiudiiwuikoeiisd spcildxusfwkiacbxokefewaoalakiee
##                               21                               21
## xbsbaipfluioualwapemiublmepsbuoo libuewofdiwukcoeempcibcwcwepldap
##                               21                               20
## bdbcaommfeelfuofobfauflkiolollwk fcdfsumaxdslpwpxekaxasfuffeakxca
##                               19                               19
## afeccskfmobewicibxofslkxecsuekfi ffmciapbdkcwwiwpuakakmiexskcmxfc
##                               18                               18
## kmkacdccelocksmlpallpcwpiicoewsw mksmfeexfwuuwsbpamfmxbikklcdwkbb
##                               18                               18
## pxxlamsdbssumpslpkduwskxodummews wdkxkxomwacxkwubwisleblebfekluib
##                               18                               18
## bapcuxcousodpaabofsesslupodaapcx idpmdduoieixxoklkufmspokimxcxeid
##                               17                               17
## plflscsfepabwxekcdlecbilsbxakwsd owuppmeskiukobiwdkfxoufexesaewil
##                               17                               16
## ppbmluelablufxsafiemmpxufupiwaik akokxbmlwukcmwlimosloemdplieuuwm
##                               16                               15
```

```
## apmpdisoaulbesoawkkekkcpokeaeucl epmwweimsesebmlpseufxpckcxmmuxol
##                               15                               15
## wlxfbefauebfbauopppswxppaafdkoap axekkipoplpalkpikkkfdumlapcufmlb
##                               15                               14
## ibfoefbbekcwubufxslcoewkfswxolua mkauefdplcsocbdeeopxiiuoumpawpds
##                               14                               14
## xwwsfoileuxkwbcxupadudcfoecmmdda acpmlkfcadicfcpslmoxcdakikieeeso
##                               14                               13
## ciixbauekwabolfbbbsswfupoiioowsd cwleuplwopmllxkbabaoeopmxxmfaiod
##                               13                               13
## ebadppbpcufaidikpolbbxxfuelueofp fcoesawwkbuwfswmpimwkiplsumkoiei
##                               13                               13
## imfclodmbmabakpawdmwfssefabcemoe swxildmwpxuuwuuoesmobpewaakakssi
##                               13                               13
## xlpufbwemoedxpsssmkbipwwicsadebw dmklwapxmxxalfwupxepeiuoooduaueb
##                               13                               12
## ixuciffexbsibwibpcwdmfwcoixkfscw okxcpskmecbumwcifxfxdofxocupwwom
##                               12                               12
## uapfospaxexfspkkskumakxcdlwuiuul cpsbiipoacmouecemlddaxxdllacksaw
##                               12                               11
## iabskbxembdweacmalxplabbxupsaadc iimppdwbwecsmxcpliaesdiasxccpwie
##                               11                               11
## mbiecfsdmkwubbksoapxsficcmioesue ppwwsldwidxcfoieopwwsxaixpkbaswl
##                               11                               11
## uaxxxwkppmwfciofupisxsdeauikeppw wixkdsawloxffiwmwswkcudoewxmawou
##                               11                               11
## xkfpfmcwobuumawmkxleudppfwiwwbmb                          (Other)
##                               11                             1124
```

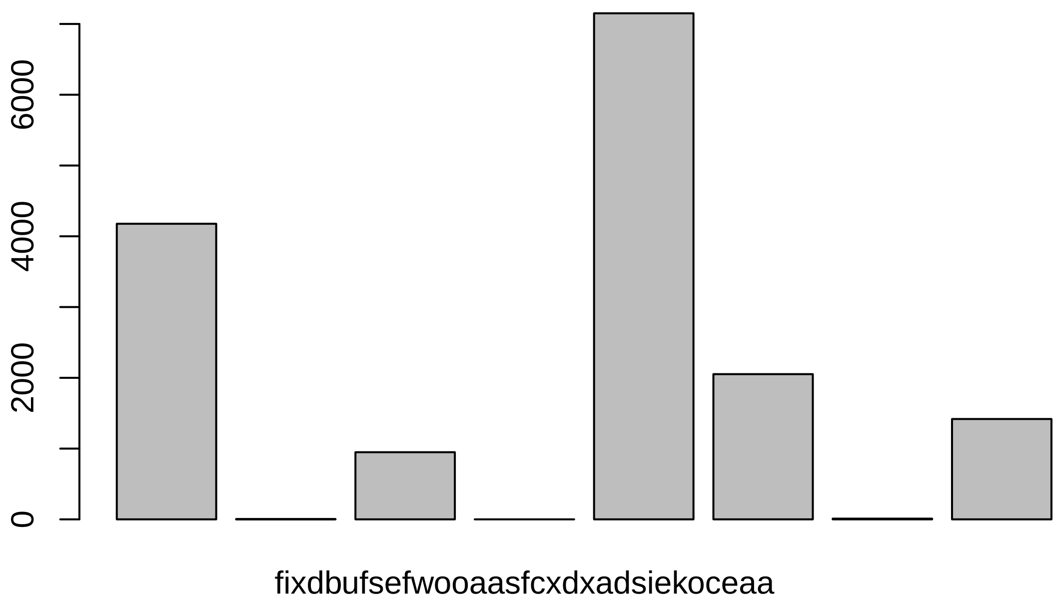
```
barplot(summary(train_new$activity_new))
```


```
summary(train_new$channel_sales)
```

```
##                                  epumfxlbckeskwekxbiuasklxalciiuu
```

```
##                               4177                                  4
## ewpakwlliwisiwduibdlfmalxowmwpci fixdbufsefwooaasfcxdxadsiekoceaa
##                                949                                  0
## foosdfpfkusacimwkcsosbicdxkicaua lmkebamcaaclubfxadlmueccxoimlema
##                               7151                               2052
## sddiedcslfslkckwlfkdpoeeailfpeds usilxuppasemubllopkaafesmlibmsdf
##                                 10                               1418
```

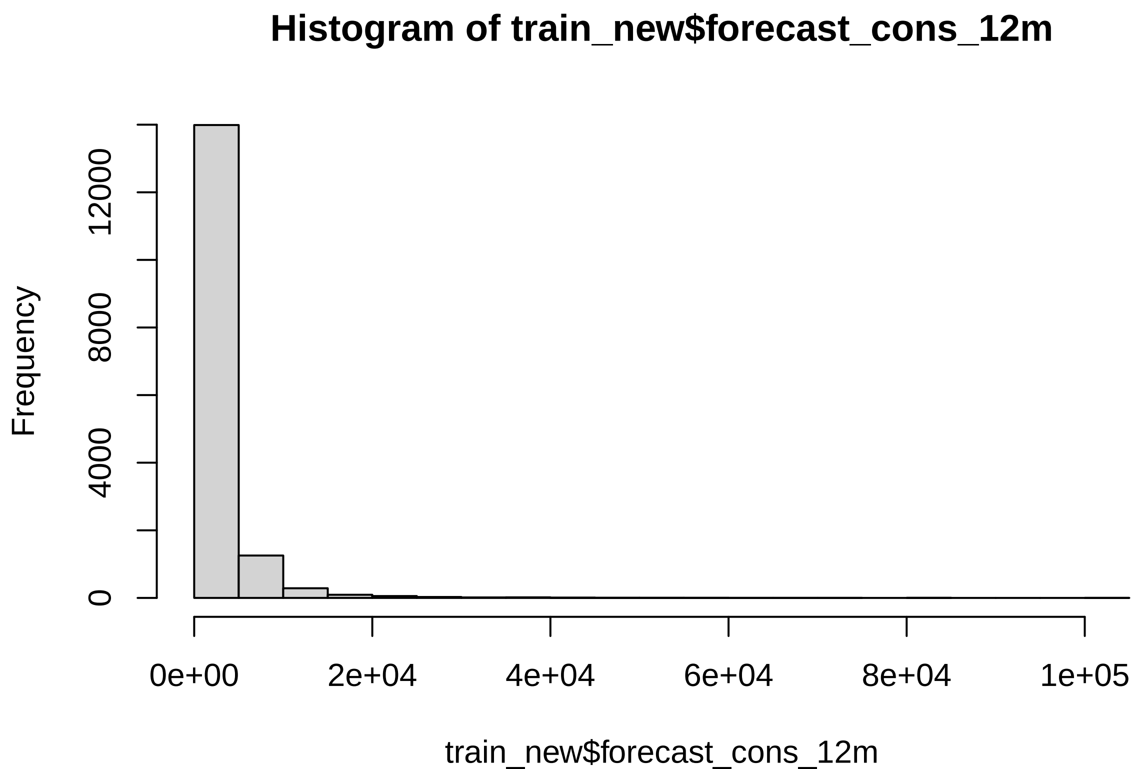
```
barplot(summary(train_new$channel_sales))
```




```
summary(train_new$forecast_cons_12m)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##      0.0    514.2   1179.8   2364.4   2685.8 103801.9
```

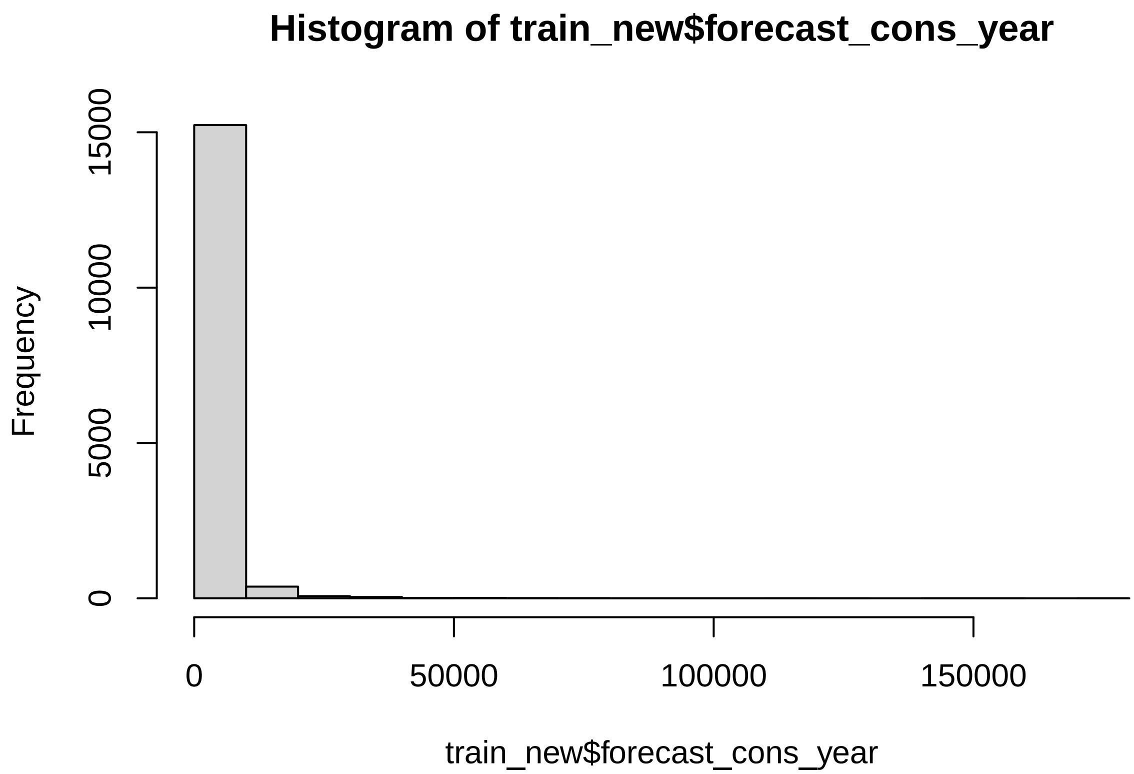
```
hist(train_new$forecast_cons_12m)
```




```
summary(train_new$forecast_cons_year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0     383    1918    1999  175375
```

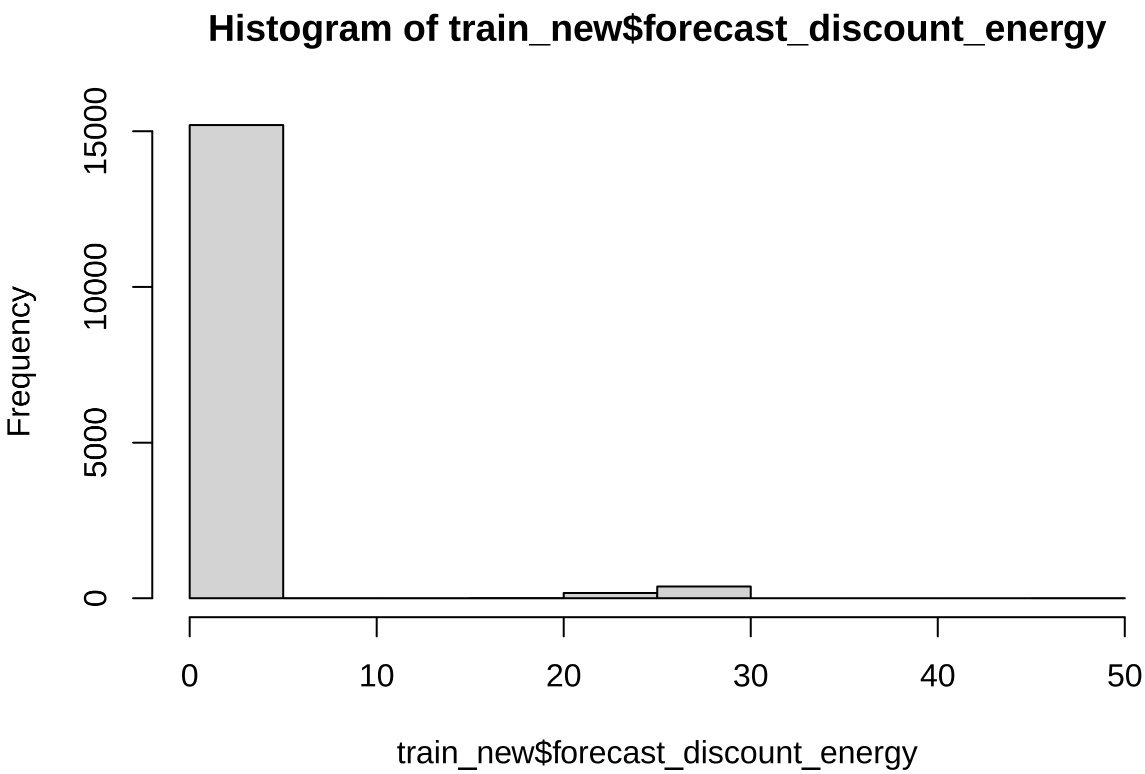
```
hist(train_new$forecast_cons_year)
```




```
summary(train_new$forecast_discount_energy)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.9792  0.0000 50.0000
```

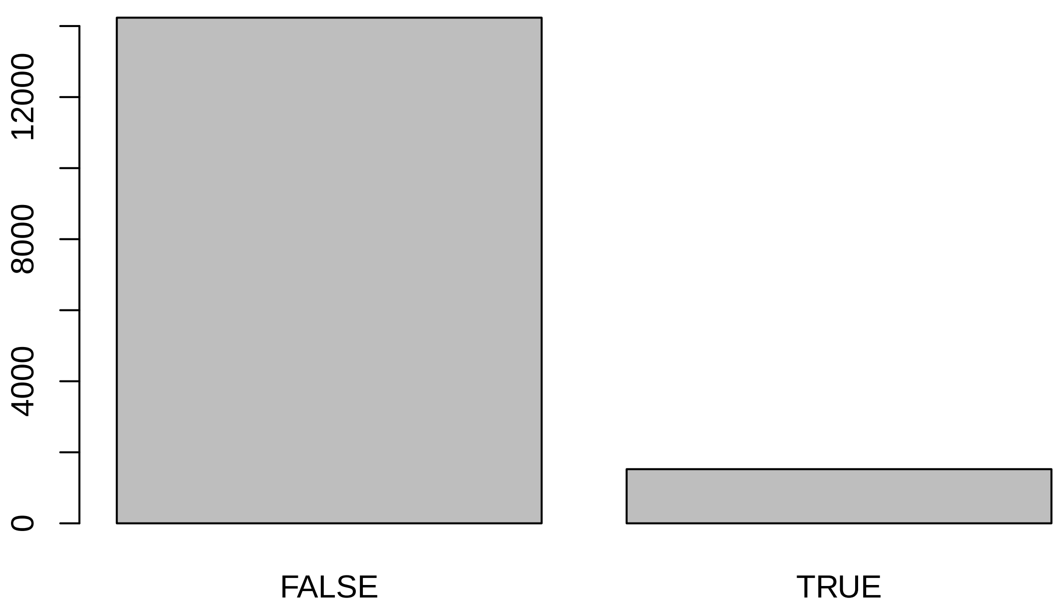
```
hist(train_new$forecast_discount_energy)
```




```
summary(train_new$churn)
```

```
##    Mode   FALSE    TRUE
## logical   14236    1525
```

```
barplot(table(train_new$churn))
```



For example, we see that 1525 out of 15761 customers chose to churn. To make more beautiful plots we can use advanced plotting packages such as `ggplot2` or `plotly`, or other visualisation software like Tableau.