DM_EDA+Associaton Rules

Group 5

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
#load the packages
library(readr)
library(readxl)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tid
yverse 1.2.1 ──
```

```
## ✔ ggplot2 3.2.1     ✔ purrr   0.3.3
## ✔ tibble  2.1.3     ✔ dplyr   0.8.3
## ✔ tidyr   1.0.0     ✔ stringr 1.4.0
## ✔ ggplot2 3.2.1     ✔ forcats 0.4.0
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse
_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(caret)
```

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

```
##
## Attaching package: 'caret'
```

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

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

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

```
library(leaps)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```r
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(formattable)
```

```
##
## Attaching package: 'formattable'
```

```
## The following object is masked from 'package:MASS':
##
##     area
```

```r
library(outliers)
library(ggplot2)
library(cowplot)
```

```
##
## ********************************************************
```

```
## Note: As of version 1.0.0, cowplot does not change the
```

```
##   default ggplot2 theme anymore. To recover the previous
```

```
##   behavior, execute:
##   theme_set(theme_cowplot())
```

```
## ********************************************************
```

```r
library(arules)
```

```
## Warning: package 'arules' was built under R version 3.6.2
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
##
## Attaching package: 'arules'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```r
library(arulesViz)
```

```
## Loading required package: grid
```

```
## Registered S3 method overwritten by 'seriation':
##   method         from
##   reorder.hclust gclus
```

```
#upload the target dataset
churn_data <- read_csv("~/Desktop/MBRChurnModel_FirstYear_MSK (1).cs
v")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   RENEW = col_character(),
##   M2EXCFLG = col_character(),
##   F2HOMRGN = col_character(),
##   HOMEFCTYCHANGE = col_character(),
##   RECENTMOVING = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
#check the missing values
sapply(churn_data, function(x) sum(is.na(x)))
```

```
##          RENEW       A2ACCIPK       A2ACCTYP       M2EXCFLG
B2BUSTYP
##              0              0              0              0
0
##       F2HOMRGN       F2HOMFCY            AGE         TENURE
ZIPCODE
##              0              0              0              0
0
##       MBRCOUNT       DISTANCE  EARLYFAREWELL HOMEFCTYCHANGE   REC
ENTMOVING
##              0              0              0              0
0
##        SHOP1YR         SHOP6M         SHOP3M       ECOMSHOP
GASSHOP
##              0              0              0              0
0
##    MEDICALSHOP    GROCERYSHOP
##              0              0
```

```
#drop the irrelavant column (customer No.)
churn_data$A2ACCIPK <- NULL

head(churn_data)
```

```
## # A tibble: 6 x 21
##   RENEW A2ACCTYP M2EXCFLG B2BUSTYP F2HOMRGN F2HOMFCY   AGE TENURE
ZIPCODE
##   <chr>    <dbl> <chr>       <dbl> <chr>       <dbl> <dbl>  <dbl>
<dbl>
## 1 N            1 N               0 NE           1078    42      1
20715
## 2 N            1 N               0 BO            847    61      1
77346
## 3 N            1 N               0 BO            847    52      1
91024
## 4 N            1 N               0 SE            185    32      1
32789
## 5 N            1 E               0 BA            472    46      1
93960
## 6 N            1 E               0 BD            823    36      1
94544
## # … with 12 more variables: MBRCOUNT <dbl>, DISTANCE <dbl>,
## #   EARLYFAREWELL <dbl>, HOMEFCTYCHANGE <chr>, RECENTMOVING <chr>
,
## #   SHOP1YR <dbl>, SHOP6M <dbl>, SHOP3M <dbl>, ECOMSHOP <dbl>,
## #   GASSHOP <dbl>, MEDICALSHOP <dbl>, GROCERYSHOP <dbl>
```

```
str(churn_data)
```

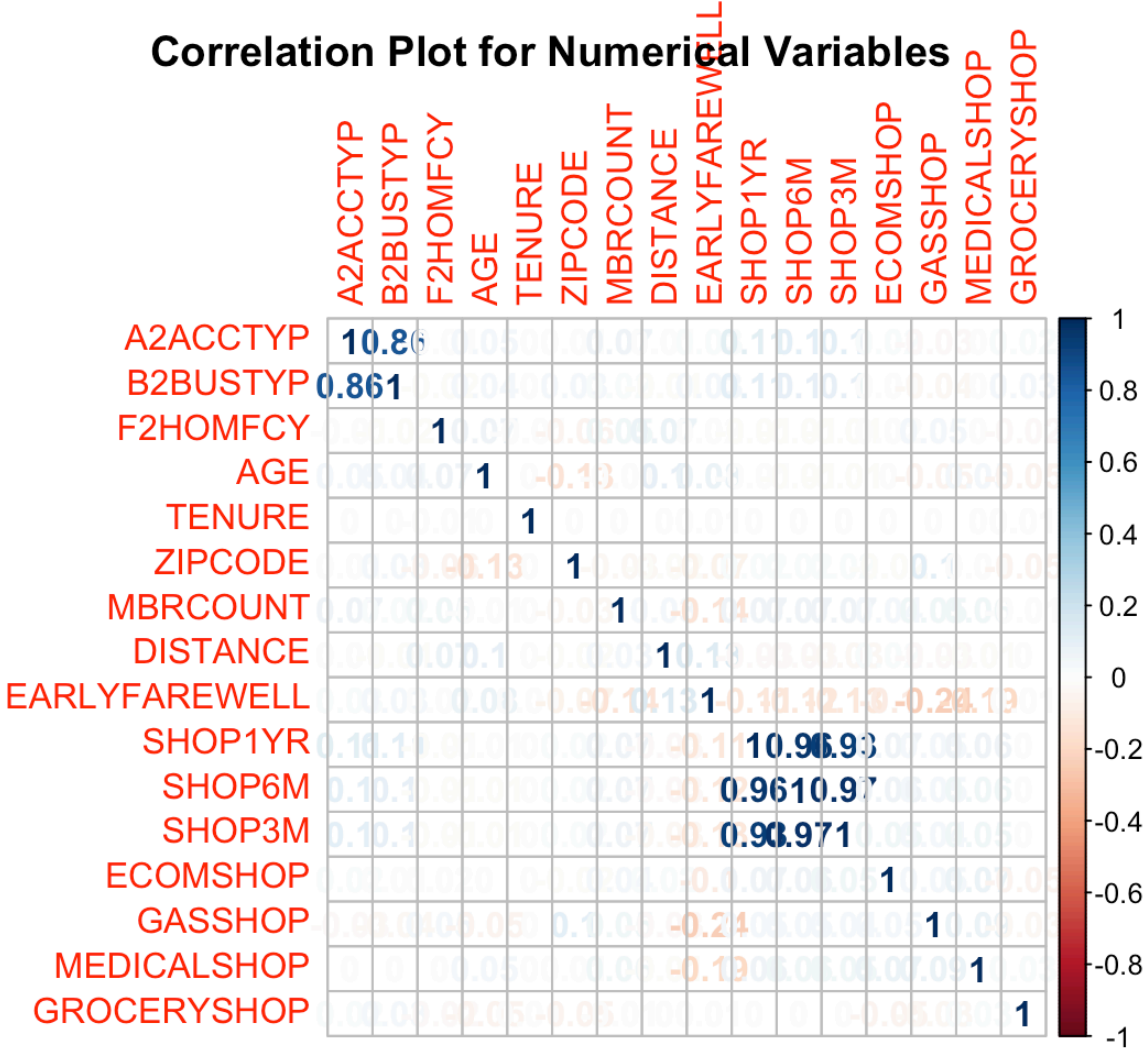
```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 120450 o
bs. of  21 variables:
##  $ RENEW         : chr  "N" "N" "N" "N" ...
##  $ A2ACCTYP      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ M2EXCFLG      : chr  "N" "N" "N" "N" ...
##  $ B2BUSTYP      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ F2HOMRGN      : chr  "NE" "BO" "BO" "SE" ...
##  $ F2HOMFCY      : num  1078 847 847 185 472 ...
##  $ AGE           : num  42 61 52 32 46 36 34 45 52 32 ...
```

```
##  $ TENURE        : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ ZIPCODE       : num  20715 77346 91024 32789 93960 ...
##  $ MBRCOUNT      : num  2 2 2 2 2 1 2 2 2 2 ...
##  $ DISTANCE      : num  7.53 6.05 7.89 3.43 26.29 ...
##  $ EARLYFAREWELL : num  75 320 350 137 41 53 38 363 53 64 ...
##  $ HOMEFCTYCHANGE: chr  "Y" "N" "N" "N" ...
##  $ RECENTMOVING  : chr  "N" "N" "N" "N" ...
##  $ SHOP1YR       : num  1385 3500 114 997 12579 ...
##  $ SHOP6M        : num  827.7 0 0 23.2 73 ...
##  $ SHOP3M        : num  253 0 0 0 73 ...
##  $ ECOMSHOP      : num  0 1 0 0 0 0 0 0 0 0 ...
##  $ GASSHOP       : num  0.0293 0 0 0.0251 0 ...
##  $ MEDICALSHOP   : num  0.0173 0 0.30377 0 0.00818 ...
##  $ GROCERYSHOP   : num  0.523 0 0.234 0.405 0.936 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   RENEW = col_character(),
##   ..   A2ACCIPK = col_double(),
##   ..   A2ACCTYP = col_double(),
##   ..   M2EXCFLG = col_character(),
##   ..   B2BUSTYP = col_double(),
##   ..   F2HOMRGN = col_character(),
##   ..   F2HOMFCY = col_double(),
##   ..   AGE = col_double(),
##   ..   TENURE = col_double(),
##   ..   ZIPCODE = col_double(),
##   ..   MBRCOUNT = col_double(),
##   ..   DISTANCE = col_double(),
##   ..   EARLYFAREWELL = col_double(),
##   ..   HOMEFCTYCHANGE = col_character(),
##   ..   RECENTMOVING = col_character(),
##   ..   SHOP1YR = col_double(),
##   ..   SHOP6M = col_double(),
##   ..   SHOP3M = col_double(),
##   ..   ECOMSHOP = col_double(),
##   ..   GASSHOP = col_double(),
##   ..   MEDICALSHOP = col_double(),
##   ..   GROCERYSHOP = col_double()
##   .. )
```

1. Binary classificatin outcome: Renew(Y or N) ——chr (initial 21 features)

2. A2ACCIPK: membership number
3. A2ACCTYP: account type: gold star/regular
4. M2EXCFLG: exclusive membership/ non-exclusive —chr 5.B2BUSTYP: if members B2B or not: 0=No; Y=yes 6.F2HOMRGN: region—chr 7.F2HOMFCY:warehouse number
5. AGE 9.TENURE: Number of months the customer has stayed 10.ZIPCODE: customer zipcode 11.MERCOUNT: number of cards hold 12.DISTANCE: miles to the warehouse 13.EARLYFAREWELL: number of days not shop 14.HOMEFCTYCHANGE: does customer change the home warehouse they are used to go?—-chr (yes or no) 15.RECENTMOVING: recent move —chr (yes or no) 16.SHOP1YR:shopping times in 1 year 17.SHOP6M: shopping times in 6 months 18.SHOP3M: shopping times in 3 months 19.ECOMSHOP: e-comme shopping % (shopping kinds) 20.GASSHOP:gas shopping% 21.MEDICALSHOP: medical shopping% 21.GROCERYSHOP:grocery shopping%

```
#Correlation between numeric variables
numeric_var <- sapply(churn_data, is.numeric)
matrix <- cor(churn_data[,numeric_var])

corrplot(matrix,main="\n\nCorrelation Plot for Numerical Variables",
method="number")
```



From the correlation plot, we can see that: B2B and A2A are correlated;(0.86) Shop1yr and shop6m are correlated;(0.96) Shop1yr and shop3m are correlated ;(0.93) shop6m and shop3m are correlated;(0.97)

```
#get the numerical variables
numeric_var <- sapply(churn_data, is.numeric)

#get the mean, max, min for numerical varianles columns from the dat
aframe
colMeans(churn_data[numeric_var])
```

```
##      A2ACCTYP      B2BUSTYP      F2HOMFCY           AGE        TE
NURE
##  1.053375e+00  3.138016e+02  6.708431e+02  4.321900e+01  9.997592
e-01
##       ZIPCODE      MBRCOUNT      DISTANCE EARLYFAREWELL       SHO
P1YR
##  6.063731e+04  1.796920e+00  1.040175e+01  7.518697e+01  2.504813
e+03
##        SHOP6M        SHOP3M      ECOMSHOP       GASSHOP   MEDICAL
SHOP
##  1.218073e+03  5.824226e+02  5.571661e-01  1.624577e+00  2.038110
e+00
##   GROCERYSHOP
##  3.496129e+00
```

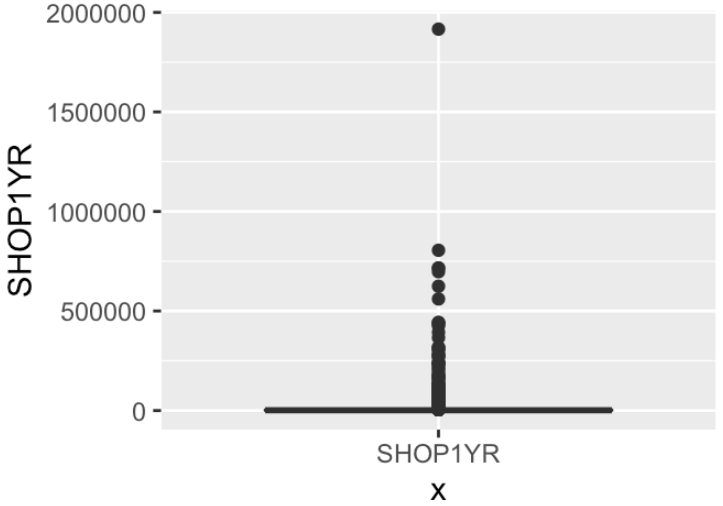
```
sapply(churn_data[numeric_var],max)
```

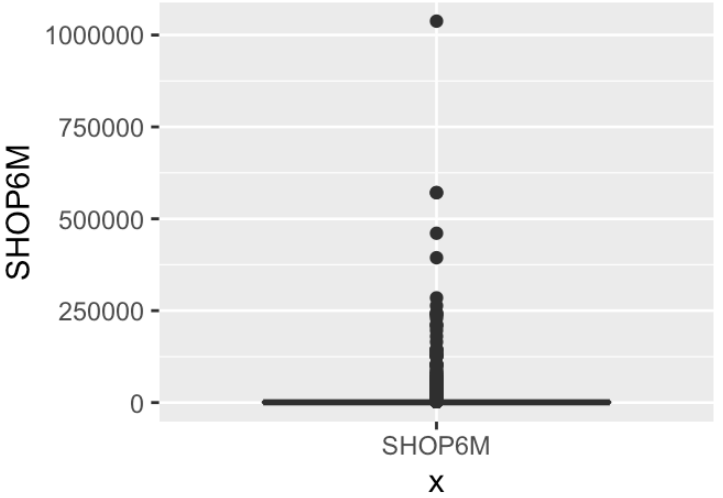
```
##      A2ACCTYP      B2BUSTYP      F2HOMFCY           AGE        TE
NURE
##        2.0000     9999.0000     1342.0000      108.0000        1.
0000
##       ZIPCODE      MBRCOUNT      DISTANCE EARLYFAREWELL       SHO
P1YR
##    99925.0000       19.0000      463.3391      409.0000  1915896.
0000
##        SHOP6M        SHOP3M      ECOMSHOP       GASSHOP   MEDICAL
SHOP
##  1037777.0000   532868.0000        9.0000        9.0000        9.
0000
##   GROCERYSHOP
##        9.0000
```

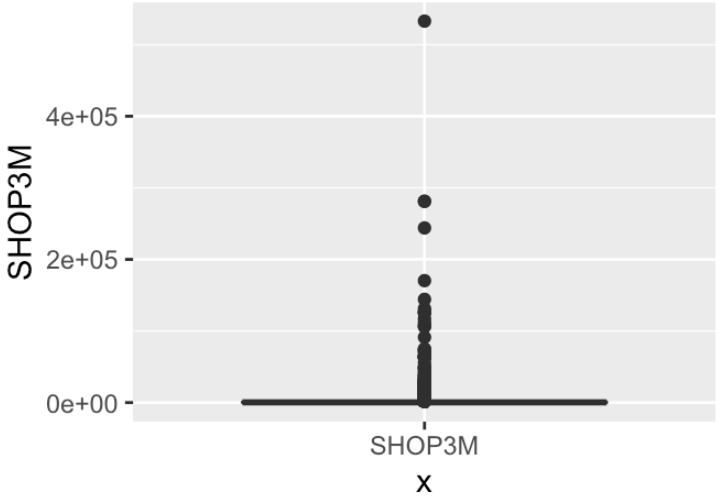
```
sapply(churn_data[numeric_var],min)
```

```
##      A2ACCTYP      B2BUSTYP      F2HOMFCY           AGE        TE
NURE
##     1.0000000     0.0000000     1.0000000    19.0000000     0.000
0000
##       ZIPCODE      MBRCOUNT      DISTANCE EARLYFAREWELL       SHO
P1YR
##   601.0000000     1.0000000     0.1168919    10.0000000     0.000
0000
##        SHOP6M        SHOP3M      ECOMSHOP       GASSHOP   MEDICAL
SHOP
##     0.0000000     0.0000000     0.0000000     0.0000000     0.000
0000
##   GROCERYSHOP
##     0.0000000
```

```
#deal with the outliers
#univariate outliers: (SHOPIYR;SHOP6M;SHOP3M)
p1 <- ggplot(churn_data, aes(x = "SHOP1YR", y = SHOP1YR)) +
    geom_boxplot()

p2<-ggplot(churn_data, aes(x = "SHOP6M", y = SHOP6M)) +
    geom_boxplot()

p3 <- ggplot(churn_data, aes(x = "SHOP3M", y = SHOP3M)) +
    geom_boxplot()
grid.arrange(p1,p2,p3,ncol=2)
```

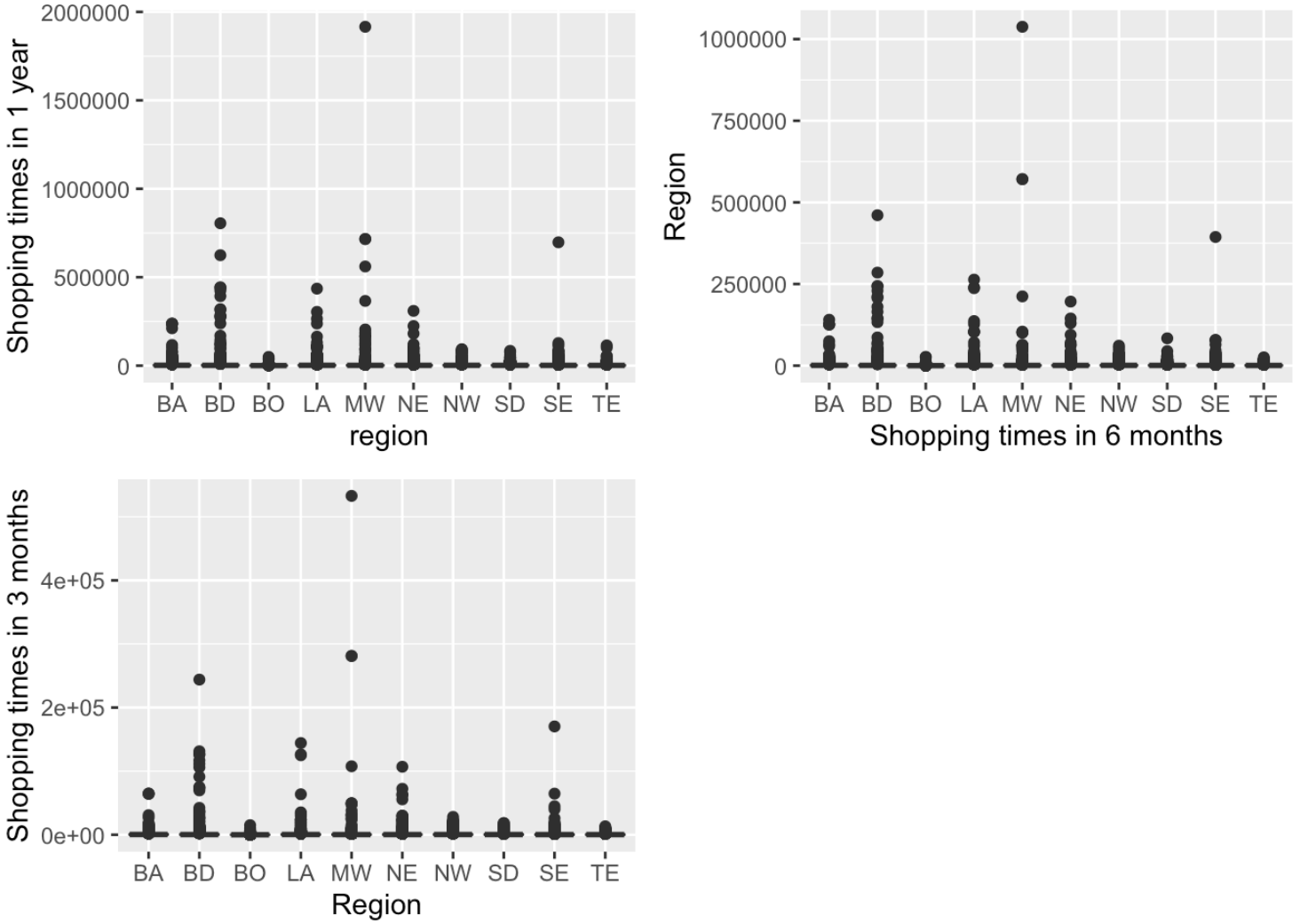






In general, an outlier is usually defined as an observation more than 3 standard deviations from the mean

```
#multivariate outliers: (SHOPIYR;SHOP6M;SHOP3M) VS (Region)
p1 <- ggplot(churn_data, aes(x =F2HOMRGN,y =SHOP1YR)) +
    geom_boxplot()+
  xlab("region")+
  ylab("Shopping times in 1 year")

p2 <- ggplot(churn_data, aes( x=F2HOMRGN,y=SHOP6M)) +
    geom_boxplot()+
    xlab("Shopping times in 6 months")+
     ylab("Region")

p3 <- ggplot(churn_data, aes(x=F2HOMRGN,y=SHOP3M)) +
    geom_boxplot()+
    xlab("Region")+
    ylab("Shopping times in 3 months")

grid.arrange(p1,p2,p3,ncol=2)
```



One way to identify outliers is to determine which points have a z-score that's far from 0. We can use the scores() function in the outliers package

```
#identify which roes contain outliers (SHOP1YR)
library(outliers)
# get the z-scores for
outlier_scores_1YR <- scores(churn_data$SHOP1YR)

#use threshold =3
#it is "TRUE" if outlier_scores is greater than 3
# it is false if outlier_scores is less than negative 3
is_outlier1YR <- outlier_scores_1YR > 3 | outlier_scores_1YR < -3

# add a column with info whether the refund_value is an outlier
churn_data$is_outlier <- is_outlier1YR

# create a dataframe with only outliers
churn_outliers_1YR <- churn_data[outlier_scores_1YR > 3| outlier_sco
res_1YR < -3, ]
str(churn_outliers_1YR)
```

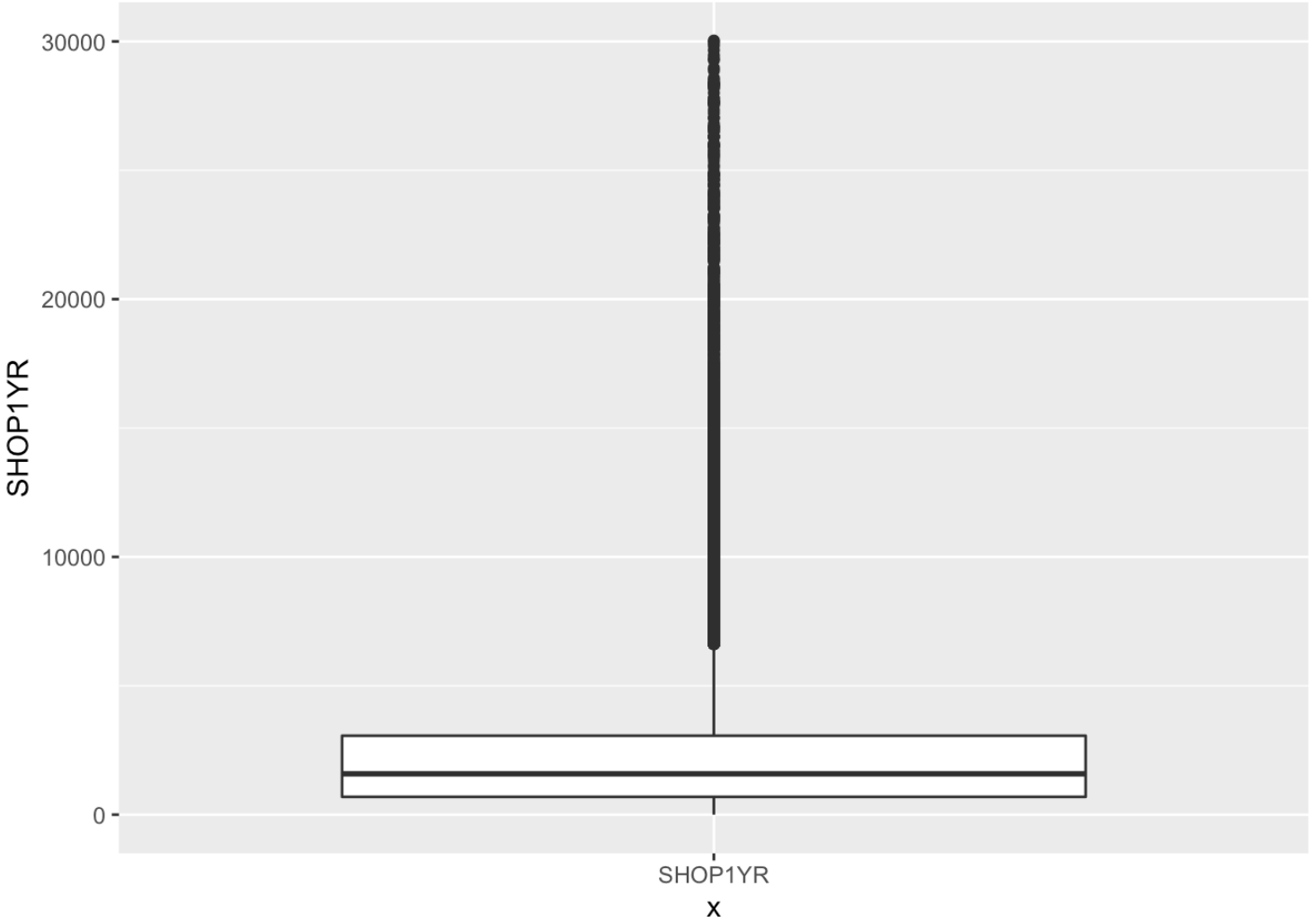
```
## Classes 'tbl_df', 'tbl' and 'data.frame':    254 obs. of  22 vari
ables:
##  $ RENEW         : chr  "Y" "Y" "N" "Y" ...
##  $ A2ACCTYP      : num  1 2 1 1 1 2 2 2 2 1 ...
##  $ M2EXCFLG      : chr  "E" "E" "E" "E" ...
##  $ B2BUSTYP      : num  0 5993 0 0 0 ...
##  $ F2HOMRGN      : chr  "MW" "BD" "NW" "LA" ...
##  $ F2HOMFCY      : num  1040 767 10 741 473 847 214 230 6 128 ...
##  $ AGE           : num  38 58 27 42 35 48 47 54 39 40 ...
##  $ TENURE        : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ ZIPCODE       : num  60175 98002 99654 93420 91789 ...
##  $ MBRCOUNT      : num  2 2 2 2 2 2 2 5 2 2 ...
##  $ DISTANCE      : num  3.7 4.88 29.52 13.58 4.11 ...
##  $ EARLYFAREWELL : num  25 15 15 101 12 12 14 11 11 16 ...
##  $ HOMEFCTYCHANGE: chr  "N" "N" "N" "Y" ...
##  $ RECENTMOVING  : chr  "N" "N" "N" "N" ...
##  $ SHOP1YR       : num  31225 392954 41426 58166 61630 ...
##  $ SHOP6M        : num  5725 229623 12453 752 45187 ...
##  $ SHOP3M        : num  2128 105936 2762 164 23077 ...
##  $ ECOMSHOP      : num  0.036 0 0 0 0.279 ...
##  $ GASSHOP       : num  0 0 0.00272 0 0.0188 0 3 2 9 8 ...
##  $ MEDICALSHOP   : num  0.379436 0.000351 0.00776 0.000945 0.3041
92 ...
##  $ GROCERYSHOP   : num  0.0771 0.00751 0.43977 0.0104 0.14652 ...
##  $ is_outlier    : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
```

```
#Remove rows with outliers from churn dataset
churn_clean1<- churn_data[churn_data$is_outlier== F, ]
str(churn_clean1)
```

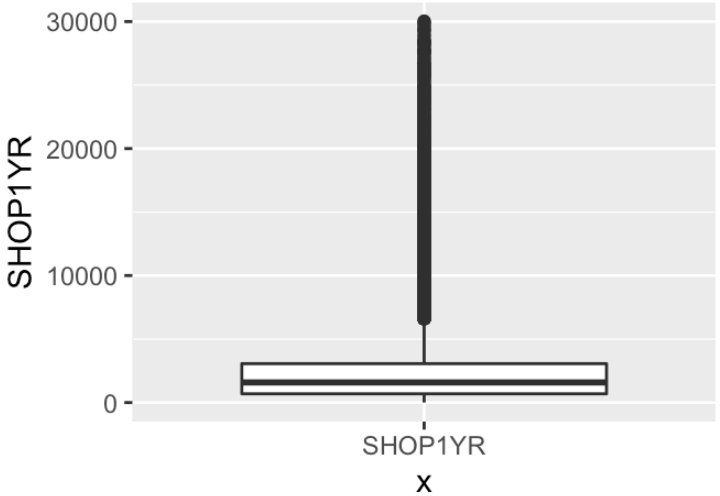
```
## Classes 'tbl_df', 'tbl' and 'data.frame':    120196 obs. of  22 v
ariables:
##  $ RENEW         : chr  "N" "N" "N" "N" ...
##  $ A2ACCTYP      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ M2EXCFLG      : chr  "N" "N" "N" "N" ...
##  $ B2BUSTYP      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ F2HOMRGN      : chr  "NE" "BO" "BO" "SE" ...
##  $ F2HOMFCY      : num  1078 847 847 185 472 ...
##  $ AGE           : num  42 61 52 32 46 36 34 45 52 32 ...
##  $ TENURE        : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ ZIPCODE       : num  20715 77346 91024 32789 93960 ...
##  $ MBRCOUNT      : num  2 2 2 2 2 1 2 2 2 2 ...
##  $ DISTANCE      : num  7.53 6.05 7.89 3.43 26.29 ...
##  $ EARLYFAREWELL : num  75 320 350 137 41 53 38 363 53 64 ...
##  $ HOMEFCTYCHANGE: chr  "Y" "N" "N" "N" ...
##  $ RECENTMOVING  : chr  "N" "N" "N" "N" ...
##  $ SHOP1YR       : num  1385 3500 114 997 12579 ...
##  $ SHOP6M        : num  827.7 0 0 23.2 73 ...
##  $ SHOP3M        : num  253 0 0 0 73 ...
##  $ ECOMSHOP      : num  0 1 0 0 0 0 0 0 0 0 ...
##  $ GASSHOP       : num  0.0293 0 0 0.0251 0 ...
##  $ MEDICALSHOP   : num  0.0173 0 0.30377 0 0.00818 ...
##  $ GROCERYSHOP   : num  0.523 0 0.234 0.405 0.936 ...
##  $ is_outlier    : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

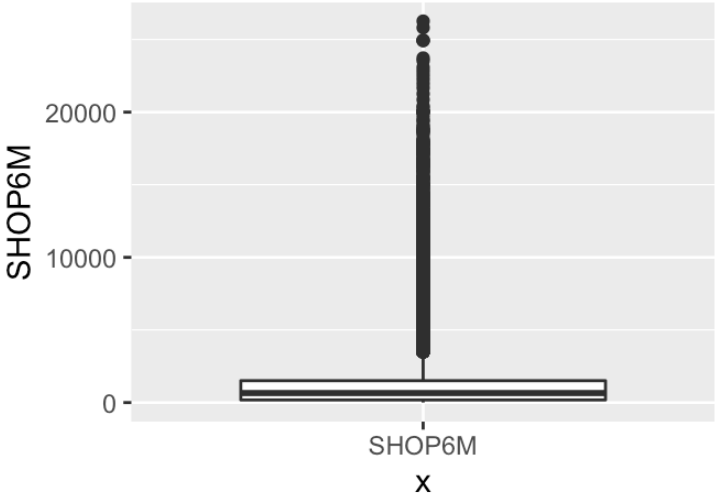
we removes outliers when: we don’t have a lot of time to figure out why you have outliers we have a large amount of data without outliers we have outliers due to measurement or data entry errors

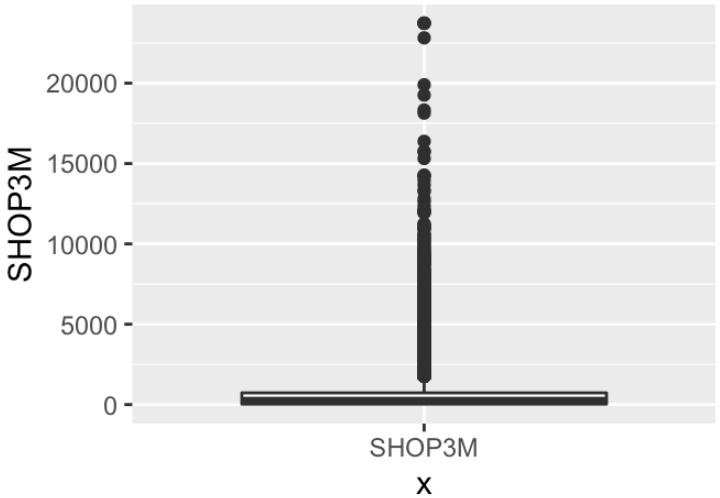
```
#check the clean churn dataset
p1 <- ggplot(churn_clean1, aes(x = "SHOP1YR", y = SHOP1YR)) +
    geom_boxplot()
p1
```




```
p2<-ggplot(churn_clean1, aes(x = "SHOP6M", y = SHOP6M)) +
    geom_boxplot()

p3 <- ggplot(churn_clean1, aes(x = "SHOP3M", y = SHOP3M)) +
    geom_boxplot()
grid.arrange(p1,p2,p3,ncol=2)
```







column of shop3m still has outliers….

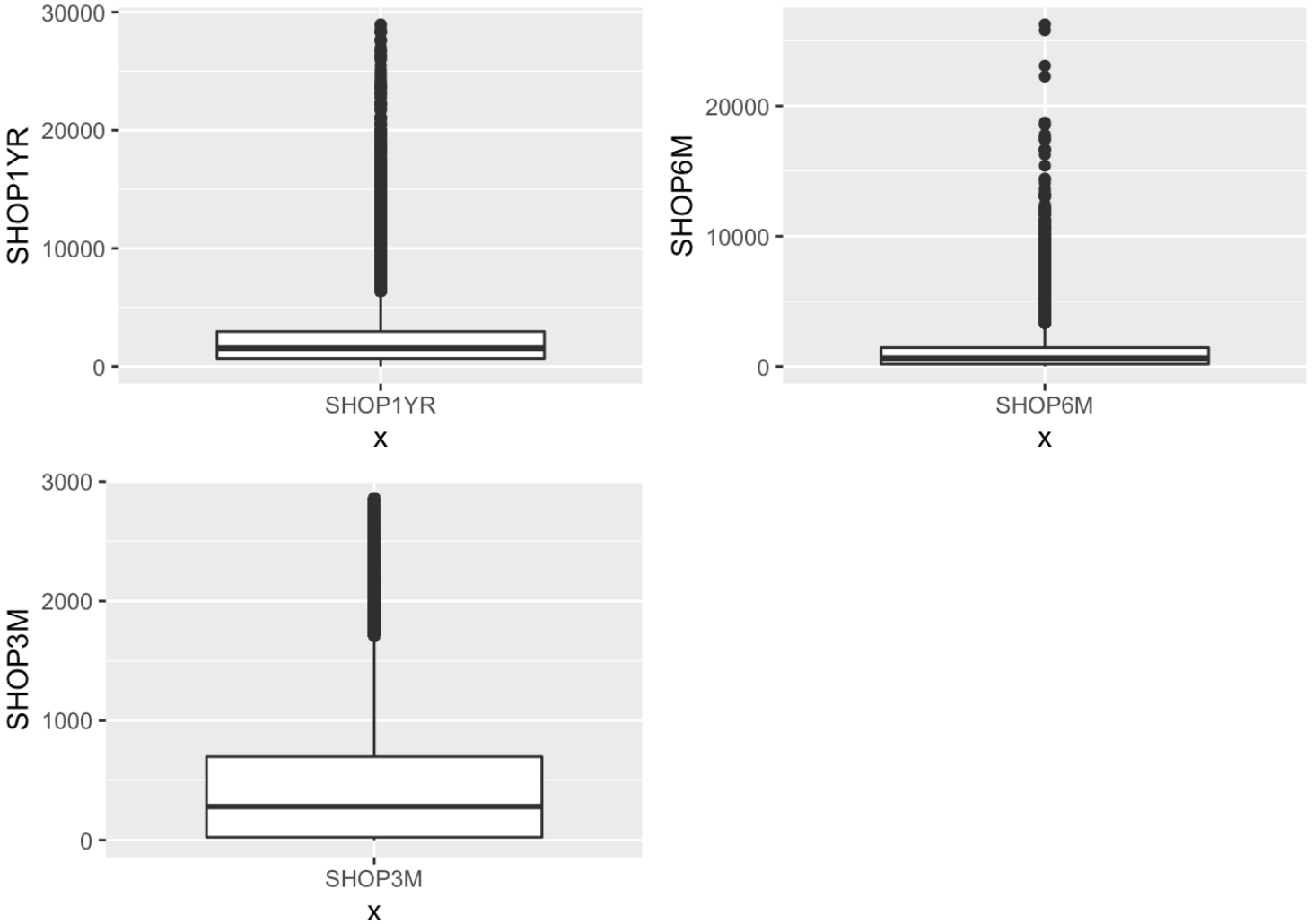
```
#identify which roes contain outliers (SHOP3M)
library(outliers)
# get the z-scores for
outlier_scores_3m <- scores(churn_clean1$SHOP3M)

#use threshold =3
#it is "TRUE" if outlier_scores is greater than 3
# it is false if outlier_scores is less than negative 3
is_outlier3m <- outlier_scores_3m > 3 | outlier_scores_3m < -3

# add a column with info whether the refund_value is an outlier
churn_clean1$is_outlieraa <- is_outlier3m

# create a dataframe with only outliers
churn_outliers_3m <- churn_clean1[outlier_scores_3m > 3| outlier_sco
res_3m < -3, ]
str(churn_outliers_3m)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1884 obs. of  23 var
iables:
##  $ RENEW         : chr  "Y" "Y" "Y" "Y" ...
##  $ A2ACCTYP      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ M2EXCFLG      : chr  "E" "E" "E" "E" ...
##  $ B2BUSTYP      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ F2HOMRGN      : chr  "SE" "NW" "BA" "SE" ...
##  $ F2HOMFCY      : num  93 733 38 1229 1206 ...
##  $ AGE           : num  65 48 58 46 58 54 42 49 50 54 ...
##  $ TENURE        : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ ZIPCODE       : num  33412 84043 95215 33133 27523 ...
##  $ MBRCOUNT      : num  2 2 2 1 2 2 1 2 2 2 ...
##  $ DISTANCE      : num  8.53 2.95 9.59 6.29 6.34 ...
##  $ EARLYFAREWELL : num  12 16 13 30 13 12 15 12 12 12 ...
##  $ HOMEFCTYCHANGE: chr  "N" "Y" "N" "N" ...
##  $ RECENTMOVING  : chr  "N" "N" "N" "N" ...
##  $ SHOP1YR       : num  18794 12574 10267 7900 11917 ...
##  $ SHOP6M        : num  4385 9406 5045 7452 6828 ...
##  $ SHOP3M        : num  2964 3029 3297 4644 3351 ...
##  $ ECOMSHOP      : num  0.0431 0.091 0 0.0137 0.0498 0.00521 0 0.
0549 0.00915 0 ...
##  $ GASSHOP       : num  0 0.00386 0 0 0.23255 ...
##  $ MEDICALSHOP   : num  0.0255 0.00852 0 0.0357 0.13301 ...
##  $ GROCERYSHOP   : num  0.0535 0.2923 0.106 0.4889 0.2946 ...
##  $ is_outlier    : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ is_outlieraa  : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
```

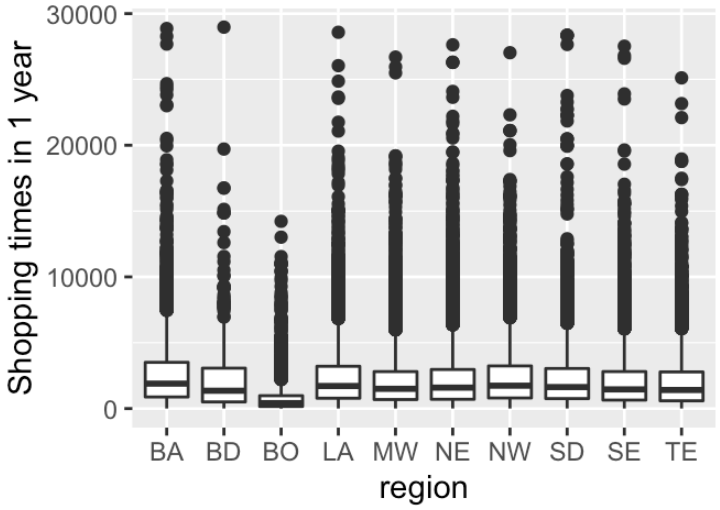
```
#Remove rows with outliers from churn dataset
churn_clean2<- churn_clean1[churn_clean1$is_outlieraa== F, ]
str(churn_clean2)
```

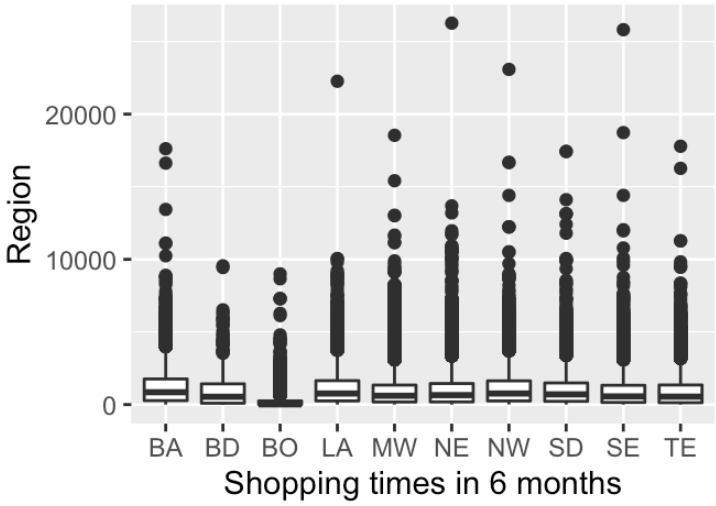
```
## Classes 'tbl_df', 'tbl' and 'data.frame':    118312 obs. of  23 v
ariables:
##  $ RENEW         : chr  "N" "N" "N" "N" ...
##  $ A2ACCTYP      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ M2EXCFLG      : chr  "N" "N" "N" "N" ...
##  $ B2BUSTYP      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ F2HOMRGN      : chr  "NE" "BO" "BO" "SE" ...
##  $ F2HOMFCY      : num  1078 847 847 185 472 ...
##  $ AGE           : num  42 61 52 32 46 36 34 45 52 32 ...
##  $ TENURE        : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ ZIPCODE       : num  20715 77346 91024 32789 93960 ...
##  $ MBRCOUNT      : num  2 2 2 2 2 1 2 2 2 2 ...
##  $ DISTANCE      : num  7.53 6.05 7.89 3.43 26.29 ...
##  $ EARLYFAREWELL : num  75 320 350 137 41 53 38 363 53 64 ...
##  $ HOMEFCTYCHANGE: chr  "Y" "N" "N" "N" ...
##  $ RECENTMOVING  : chr  "N" "N" "N" "N" ...
##  $ SHOP1YR       : num  1385 3500 114 997 12579 ...
##  $ SHOP6M        : num  827.7 0 0 23.2 73 ...
##  $ SHOP3M        : num  253 0 0 0 73 ...
##  $ ECOMSHOP      : num  0 1 0 0 0 0 0 0 0 0 ...
##  $ GASSHOP       : num  0.0293 0 0 0.0251 0 ...
##  $ MEDICALSHOP   : num  0.0173 0 0.30377 0 0.00818 ...
##  $ GROCERYSHOP   : num  0.523 0 0.234 0.405 0.936 ...
##  $ is_outlier    : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ is_outlieraa  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

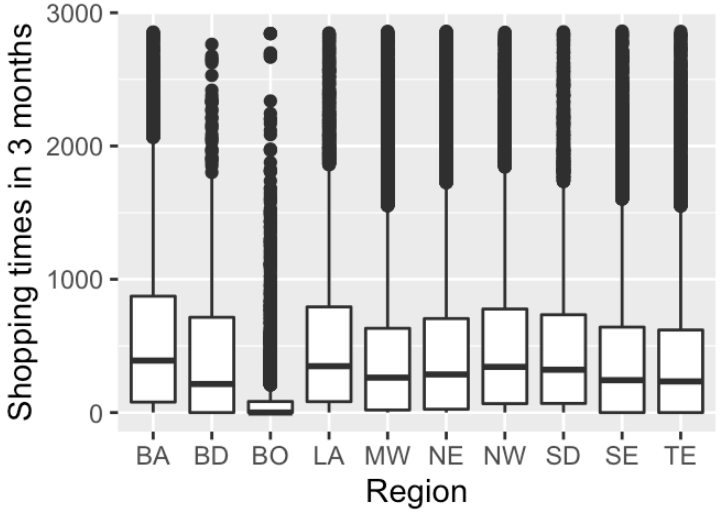
```
#check the clean churn dataset
p1 <- ggplot(churn_clean2, aes(x = "SHOP1YR", y = SHOP1YR)) +
    geom_boxplot()

p2<-ggplot(churn_clean2, aes(x = "SHOP6M", y = SHOP6M)) +
    geom_boxplot()

p3 <- ggplot(churn_clean2, aes(x = "SHOP3M", y = SHOP3M)) +
    geom_boxplot()
grid.arrange(p1,p2,p3,ncol=2)
```




```
#check the clean churn dataset
p1 <- ggplot(churn_clean2, aes(x =F2HOMRGN,y =SHOP1YR)) +
    geom_boxplot()+
  xlab("region")+
  ylab("Shopping times in 1 year")

p2 <- ggplot(churn_clean2, aes( x=F2HOMRGN,y=SHOP6M)) +
    geom_boxplot()+
    xlab("Shopping times in 6 months")+
     ylab("Region")

p3 <- ggplot(churn_clean2, aes(x=F2HOMRGN,y=SHOP3M)) +
    geom_boxplot()+
    xlab("Region")+
    ylab("Shopping times in 3 months")

grid.arrange(p1,p2,p3,ncol=2)
```








```
#use the churn_clean12 (no outliers in SHOP1YR and SHOP3m)

New1 <- churn_clean2[,!names(churn_clean2) %in% c("is_outlier","is_o
utlieraa")]
New11 <- mutate(New1, "TOTALECOM"=SHOP1YR*ECOMSHOP, "TOTALGAS"=SHOP1
YR*GASSHOP , "TOTALMEDICAL"=SHOP1YR*MEDICALSHOP, "TOTALGROCER"=SHOP1
YR*GROCERYSHOP)
head(New11)
```

```
## # A tibble: 6 x 25
##   RENEW A2ACCTYP M2EXCFLG B2BUSTYP F2HOMRGN F2HOMFCY   AGE TENURE
ZIPCODE
##   <chr>    <dbl> <chr>       <dbl> <chr>       <dbl> <dbl>  <dbl>
<dbl>
## 1 N            1 N               0 NE           1078    42      1
20715
## 2 N            1 N               0 BO            847    61      1
77346
## 3 N            1 N               0 BO            847    52      1
91024
## 4 N            1 N               0 SE            185    32      1
32789
## 5 N            1 E               0 BA            472    46      1
93960
## 6 N            1 E               0 BD            823    36      1
94544
## # … with 16 more variables: MBRCOUNT <dbl>, DISTANCE <dbl>,
## #   EARLYFAREWELL <dbl>, HOMEFCTYCHANGE <chr>, RECENTMOVING <chr>
,
## #   SHOP1YR <dbl>, SHOP6M <dbl>, SHOP3M <dbl>, ECOMSHOP <dbl>,
## #   GASSHOP <dbl>, MEDICALSHOP <dbl>, GROCERYSHOP <dbl>, TOTALECO
M <dbl>,
## #   TOTALGAS <dbl>, TOTALMEDICAL <dbl>, TOTALGROCER <dbl>
```

```
dim(New11)
```

```
## [1] 118312     25
```

```
str(New11)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    118312 obs. of  25 v
ariables:
##  $ RENEW         : chr  "N" "N" "N" "N" ...
##  $ A2ACCTYP      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ M2EXCFLG      : chr  "N" "N" "N" "N" ...
##  $ B2BUSTYP      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ F2HOMRGN      : chr  "NE" "BO" "BO" "SE" ...
##  $ F2HOMFCY      : num  1078 847 847 185 472 ...
##  $ AGE           : num  42 61 52 32 46 36 34 45 52 32 ...
##  $ TENURE        : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ ZIPCODE       : num  20715 77346 91024 32789 93960 ...
##  $ MBRCOUNT      : num  2 2 2 2 2 1 2 2 2 2 ...
##  $ DISTANCE      : num  7.53 6.05 7.89 3.43 26.29 ...
##  $ EARLYFAREWELL : num  75 320 350 137 41 53 38 363 53 64 ...
##  $ HOMEFCTYCHANGE: chr  "Y" "N" "N" "N" ...
##  $ RECENTMOVING  : chr  "N" "N" "N" "N" ...
##  $ SHOP1YR       : num  1385 3500 114 997 12579 ...
##  $ SHOP6M        : num  827.7 0 0 23.2 73 ...
##  $ SHOP3M        : num  253 0 0 0 73 ...
##  $ ECOMSHOP      : num  0 1 0 0 0 0 0 0 0 0 ...
##  $ GASSHOP       : num  0.0293 0 0 0.0251 0 ...
##  $ MEDICALSHOP   : num  0.0173 0 0.30377 0 0.00818 ...
##  $ GROCERYSHOP   : num  0.523 0 0.234 0.405 0.936 ...
##  $ TOTALECOM     : num  0 3500 0 0 0 ...
##  $ TOTALGAS      : num  40.6 0 0 25 0 ...
##  $ TOTALMEDICAL  : num  24 0 34.5 0 102.9 ...
##  $ TOTALGROCER   : num  724.1 0 26.6 404.3 11772.5 ...
```

```
New2 <- New11[,!names(New11) %in% c("ECOMSHOP","GASSHOP","MEDICALSHO
P","GROCERYSHOP")]
head(New2)
```

```
## # A tibble: 6 x 21
##   RENEW A2ACCTYP M2EXCFLG B2BUSTYP F2HOMRGN F2HOMFCY   AGE TENURE
ZIPCODE
##   <chr>    <dbl> <chr>       <dbl> <chr>       <dbl> <dbl>  <dbl>
<dbl>
## 1 N            1 N               0 NE           1078    42      1
20715
## 2 N            1 N               0 BO            847    61      1
77346
## 3 N            1 N               0 BO            847    52      1
91024
## 4 N            1 N               0 SE            185    32      1
32789
## 5 N            1 E               0 BA            472    46      1
93960
## 6 N            1 E               0 BD            823    36      1
94544
## # … with 12 more variables: MBRCOUNT <dbl>, DISTANCE <dbl>,
## #   EARLYFAREWELL <dbl>, HOMEFCTYCHANGE <chr>, RECENTMOVING <chr>
,
## #   SHOP1YR <dbl>, SHOP6M <dbl>, SHOP3M <dbl>, TOTALECOM <dbl>,
## #   TOTALGAS <dbl>, TOTALMEDICAL <dbl>, TOTALGROCER <dbl>
```

```
dim(New2)
```

```
## [1] 118312     21
```

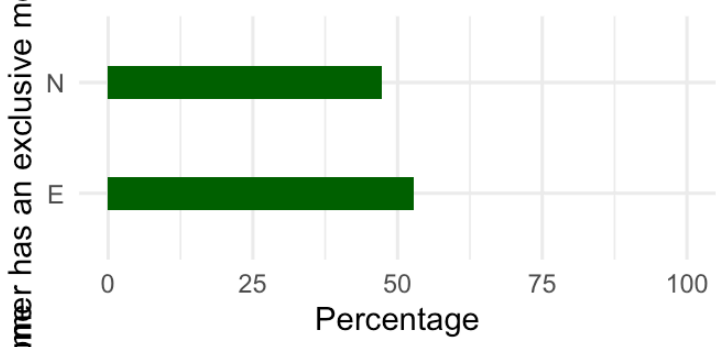
```
str(New2)
```

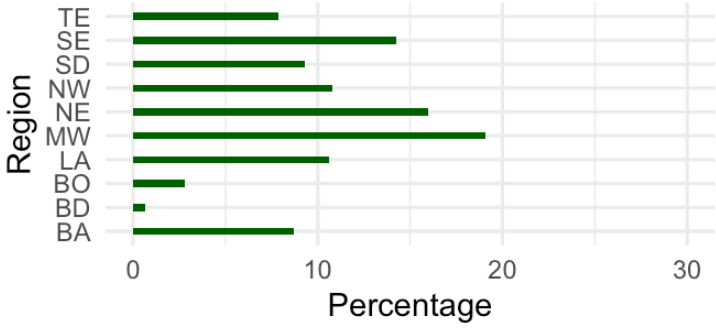
```
## Classes 'tbl_df', 'tbl' and 'data.frame':    118312 obs. of  21 v
ariables:
##  $ RENEW         : chr  "N" "N" "N" "N" ...
##  $ A2ACCTYP      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ M2EXCFLG      : chr  "N" "N" "N" "N" ...
##  $ B2BUSTYP      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ F2HOMRGN      : chr  "NE" "BO" "BO" "SE" ...
##  $ F2HOMFCY      : num  1078 847 847 185 472 ...
##  $ AGE           : num  42 61 52 32 46 36 34 45 52 32 ...
##  $ TENURE        : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ ZIPCODE       : num  20715 77346 91024 32789 93960 ...
##  $ MBRCOUNT      : num  2 2 2 2 2 1 2 2 2 2 ...
##  $ DISTANCE      : num  7.53 6.05 7.89 3.43 26.29 ...
##  $ EARLYFAREWELL : num  75 320 350 137 41 53 38 363 53 64 ...
##  $ HOMEFCTYCHANGE: chr  "Y" "N" "N" "N" ...
##  $ RECENTMOVING  : chr  "N" "N" "N" "N" ...
##  $ SHOP1YR       : num  1385 3500 114 997 12579 ...
##  $ SHOP6M        : num  827.7 0 0 23.2 73 ...
##  $ SHOP3M        : num  253 0 0 0 73 ...
##  $ TOTALECOM     : num  0 3500 0 0 0 ...
##  $ TOTALGAS      : num  40.6 0 0 25 0 ...
##  $ TOTALMEDICAL  : num  24 0 34.5 0 102.9 ...
##  $ TOTALGROCER   : num  724.1 0 26.6 404.3 11772.5 ...
```

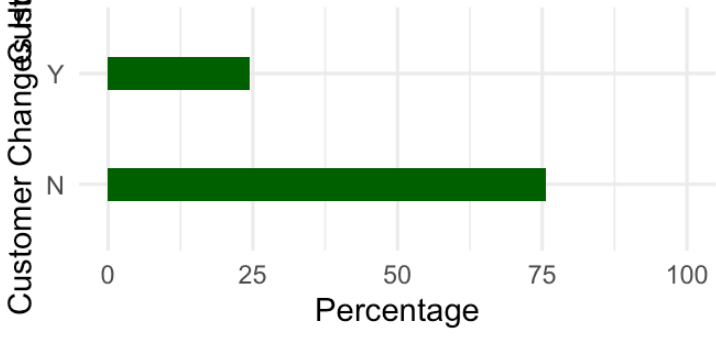
Exploratory Data Analysis

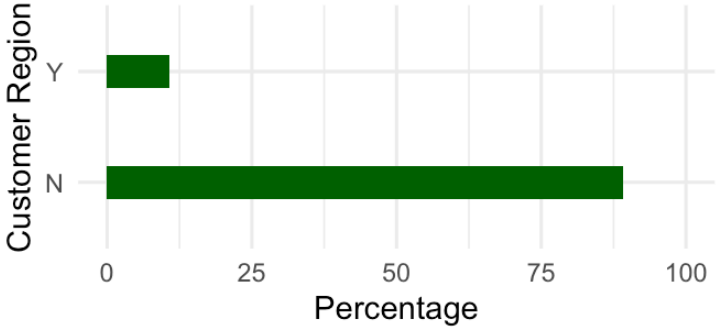
```
#Step 2
#data visualization for categorical variables
library(ggplot2)
library(cowplot)

#RENEW(Y or N) -----chr
#M2EXCFLG:exclusive membership/ non-exclusive (Y or N) ---chr
#F2HOMRGN: region---chr
#HOMEFCTYCHANGE: does customer change the home warehouse they are us
ed to go?----chr (Y or N)
#RECENTMOVING: recent move ---chr (Y or N)

p1 <- ggplot(data=New2, aes(x=M2EXCFLG))+
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.3,fill
="darkgreen") +
```

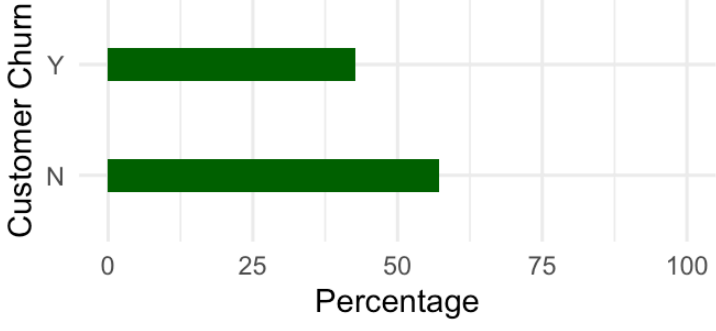
```
  ylab("Percentage") + ylim(0,100)+
  xlab("Customer has an exclusive membership")+
  coord_flip() + theme_minimal()

p2 <- ggplot(data=New2, aes(x=F2HOMRGN))+
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.3,fill
="darkgreen") +
  ylab("Percentage") + ylim(0,30)+
  xlab("Region")+
  coord_flip() + theme_minimal()

p3 <- ggplot(data=New2, aes(x=HOMEFCTYCHANGE))+
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.3,fill
="darkgreen") +
  ylab("Percentage") + ylim(0,100)+
  xlab("Customer Changes Home")+
  coord_flip() + theme_minimal()

p4 <- ggplot(data=New2, aes(x=RECENTMOVING))+
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.3,fill
="darkgreen") +
  ylab("Percentage") + ylim(0,100)+
  xlab("Customer Region")+
  coord_flip() + theme_minimal()

p5 <- ggplot(data=New2, aes(x=RENEW))+
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.3,fill
="darkgreen") +
  ylab("Percentage") + ylim(0,100)+
  xlab("Customer Churn")+
  coord_flip() + theme_minimal()

#get the bar plots of categorical variables
grid.arrange(p1, p2, p3, p4, p5)
```


Customer has an exclusive m
N
E
0
25
50
75
100
Percentage


Region
TE
SE
SD
NW
NE
MW
LA
BO
BD
BA
0
10
20
30
Percentage


Customer Changes
Y
N
0
25
50
75
100
Percentage


Customer Region
Y
N
0
25
50
75
100
Percentage


Customer Churn
Y
N
0
25
50
75
100
Percentage

```
#Step 2
##data visualization for numerical variables

shop1year <- ggplot(data=New2, aes(SHOP1YR)) +
  geom_histogram(fill="darkred") +
  geom_vline(aes(xintercept = mean(SHOP1YR)), linetype = "dashed")+
  xlim(0,15000)


shop6m <- ggplot(data=New2, aes(SHOP6M)) +
  geom_histogram(fill="darkred") +
  geom_vline(aes(xintercept = mean(SHOP6M)), linetype = "dashed")+
  xlim(0,10000)

shop3m <- ggplot(data=New2, aes(SHOP3M)) +
  geom_histogram(fill="darkred") +
  geom_vline(aes(xintercept = mean(SHOP6M)), linetype = "dashed")+
  xlim(0,5000)


grid.arrange(shop1year,shop6m,shop3m,ncol=2)
```

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

```
## Warning: Removed 205 rows containing non-finite values (stat_bin)
.
```

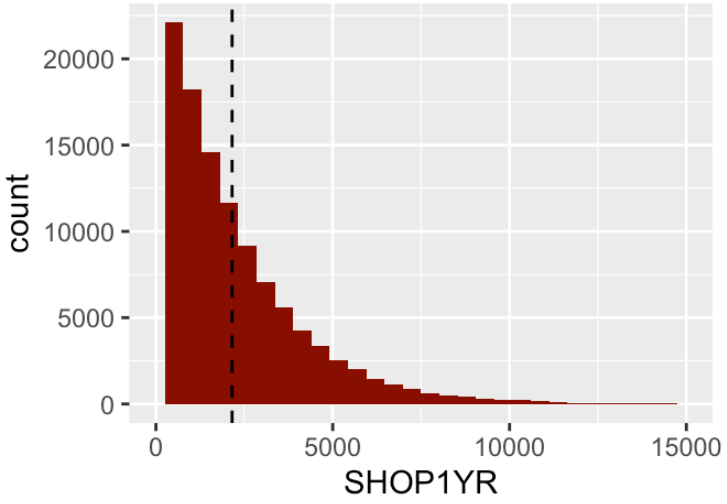
```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

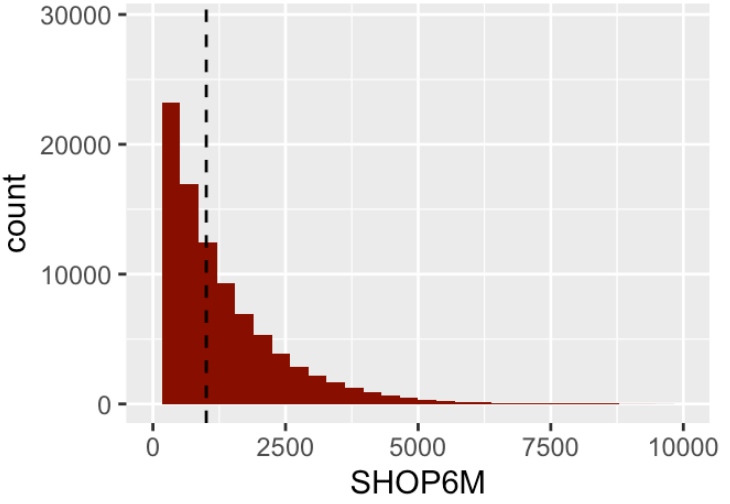
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`
.
```

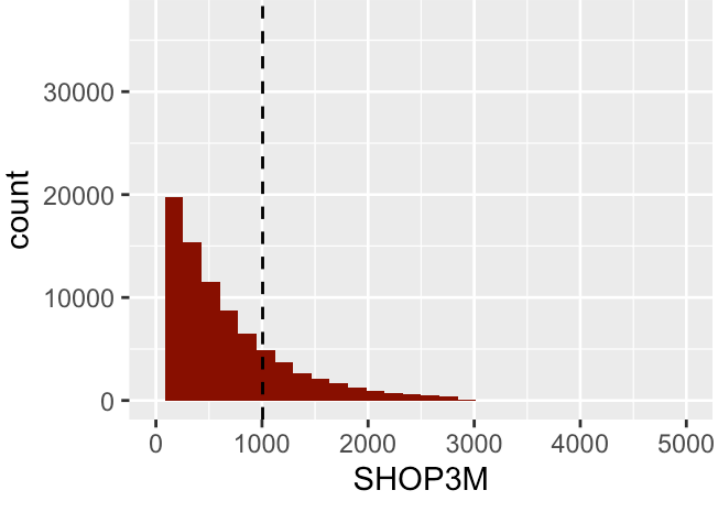
```
## Warning: Removed 66 rows containing non-finite values (stat_bin).

## Warning: Removed 2 rows containing missing values (geom_bar).
```

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

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```








```
mean(New2$SHOP1YR)
```

```
## [1] 2153.639
```

```
mean(New2$SHOP6M)
```

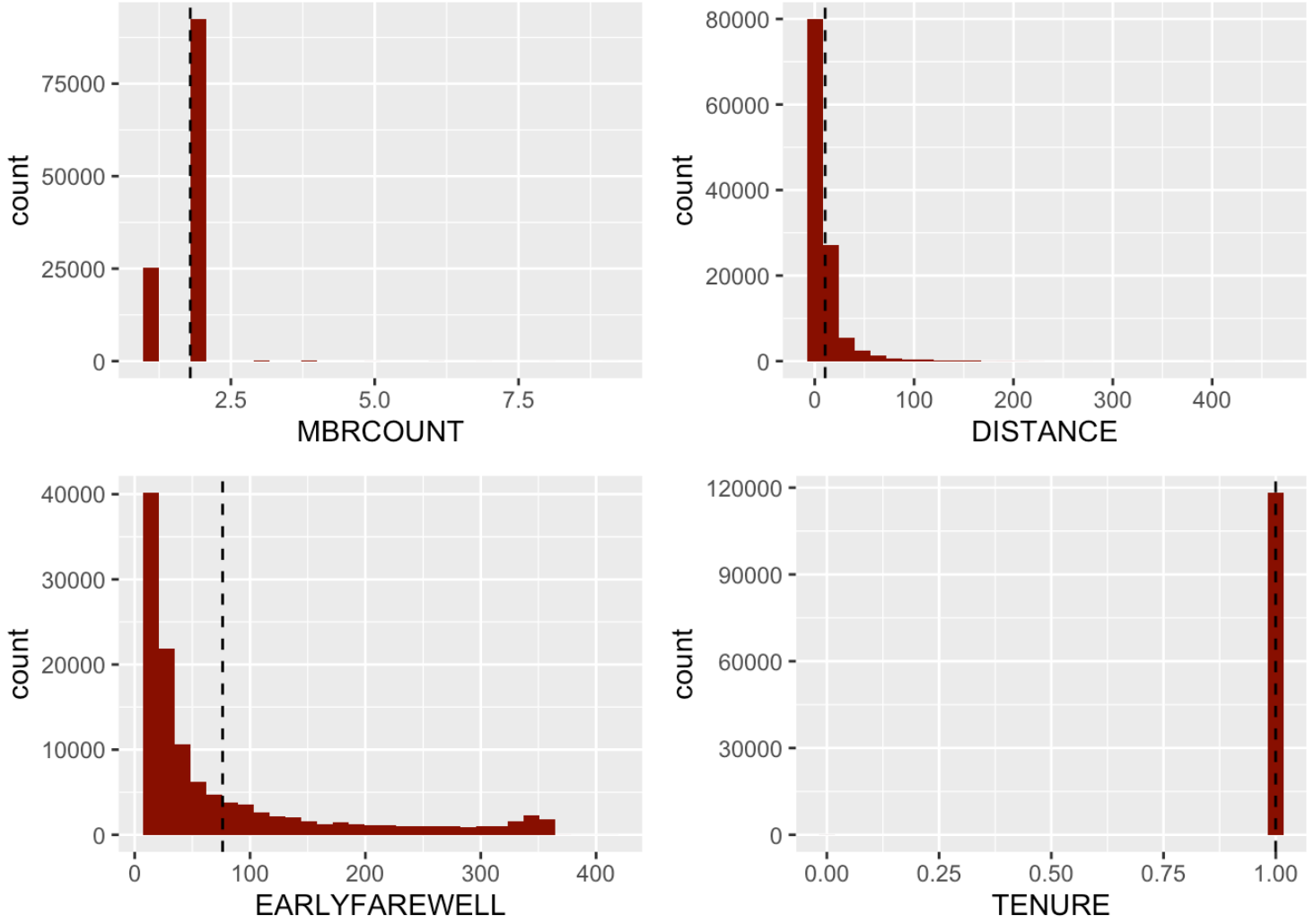
```
## [1] 1006.062
```

```
mean(New2$SHOP3M)
```

```
## [1] 467.3689
```

```
#MBRCOUNT: number of cards hold
#DISTANCE: miles to the warehouse
#EARLYFAREWELL: number of days not shop
#TENURE: Number of months the customer has stayed


p1 <- ggplot(data=New2, aes(MBRCOUNT)) +
  geom_histogram(fill="darkred") +
  geom_vline(aes(xintercept = mean(MBRCOUNT)), linetype = "dashed")

p2 <- ggplot(data=New2, aes(DISTANCE)) +
  geom_histogram(fill="darkred") +
  geom_vline(aes(xintercept = mean(DISTANCE)), linetype = "dashed")

p3 <- ggplot(data=New2, aes(EARLYFAREWELL)) +
  geom_histogram(fill="darkred") +
  geom_vline(aes(xintercept = mean(EARLYFAREWELL)), linetype = "dash
ed")


p4 <- ggplot(data=New2, aes(TENURE)) +
  geom_histogram(fill="darkred") +
  geom_vline(aes(xintercept = mean(TENURE)), linetype = "dashed")

grid.arrange(p1,p2,p3,p4,ncol=2)
```

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



Association Rules

```
library(tidyverse)
churn_data <- read_csv("~/Desktop/MBRChurnModel_FirstYear_MSK (1).cs
v")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   RENEW = col_character(),
##   M2EXCFLG = col_character(),
##   F2HOMRGN = col_character(),
##   HOMEFCTYCHANGE = col_character(),
##   RECENTMOVING = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
library(data.table)
setDT(churn_data)[AGE <1, agegroup := "0-1"]
churn_data[AGE >0 & AGE <5, agegroup := "1-4"]
churn_data[AGE >4 & AGE <10, agegroup := "5-9"]
churn_data[AGE >9 & AGE <15, agegroup := "10-14"]
churn_data[AGE >14 & AGE <20, agegroup := "15-19"]
churn_data[AGE >19 & AGE <25, agegroup := "20-24"]
churn_data[AGE >24 & AGE <30, agegroup := "25-29"]
churn_data[AGE >29 & AGE <35, agegroup := "30-34"]
churn_data[AGE >34 & AGE <40, agegroup := "35-39"]
churn_data[AGE >39 & AGE <45, agegroup := "40-44"]
churn_data[AGE >44 & AGE <50, agegroup := "45-49"]
churn_data[AGE >49 & AGE <55, agegroup := "50-54"]
churn_data[AGE >54 & AGE <60, agegroup := "55-59"]
churn_data[AGE >59 & AGE <65, agegroup := "60-64"]
churn_data[AGE >64 & AGE <70, agegroup := "65-69"]
churn_data[AGE >69 & AGE <75, agegroup := "70-74"]
churn_data[AGE >74 & AGE <80, agegroup := "75-79"]
churn_data[AGE >79 & AGE <85, agegroup := "80-84"]
churn_data[AGE >84, agegroup := "85+"]
```

```
churn_data$EARLYFAREWELL<- as.integer(churn_data$EARLYFAREWELL)
setDT(churn_data)
churn_data[EARLYFAREWELL >=0 & EARLYFAREWELL <60, earlygroup := "0 -
60"]
churn_data[EARLYFAREWELL >=60  & EARLYFAREWELL <120, earlygroup := "
61 - 120"]
churn_data[EARLYFAREWELL >=120 & EARLYFAREWELL <180, earlygroup := "
121 - 180"]
churn_data[EARLYFAREWELL >=180 & EARLYFAREWELL <240, earlygroup := "
181 - 240"]
churn_data[EARLYFAREWELL >=240 & EARLYFAREWELL <300, earlygroup := "
241 - 300"]
churn_data[EARLYFAREWELL >=300 & EARLYFAREWELL <360, earlygroup := "
301 - 360"]
churn_data[EARLYFAREWELL >=360 & EARLYFAREWELL <420, earlygroup := "
361 - 420"]
```

```
setDT(churn_data)[DISTANCE < 10, DISTANCEGroup := "Less then 10"]
churn_data[DISTANCE >=10 & DISTANCE < 20, DISTANCEGroup := "10-20"]
churn_data[DISTANCE >=20 & DISTANCE < 30, DISTANCEGroup := "20-30"]
churn_data[DISTANCE >=30 & DISTANCE < 40, DISTANCEGroup := "30-40"]
churn_data[DISTANCE >=40 & DISTANCE < 50, DISTANCEGroup := "40-50"]
churn_data[DISTANCE >=50 & DISTANCE < 60, DISTANCEGroup := "50-60"]
churn_data[DISTANCE >=60 & DISTANCE < 70, DISTANCEGroup := "60-70"]
churn_data[DISTANCE >=70 & DISTANCE < 80, DISTANCEGroup := "70-80"]
churn_data[DISTANCE >=80 & DISTANCE < 90, DISTANCEGroup := "80-90"]
churn_data[DISTANCE >=90 & DISTANCE < 100, DISTANCEGroup := "90-100"
]
churn_data[DISTANCE >=100, DISTANCEGroup := "100+"]
```

```
setDT(churn_data)[SHOP1YR <1000, shop1YrGROUP := "0-1000"]
churn_data[SHOP1YR >=1000 & SHOP1YR <5000, shop1YrGROUP := "1001-500
0"]
churn_data[SHOP1YR >=5000 & SHOP1YR <10000, shop1YrGROUP := "5001-10
000"]
churn_data[SHOP1YR >=10000 & SHOP1YR <50000, shop1YrGROUP := "10001-
50000"]
churn_data[SHOP1YR >=50000 & SHOP1YR <100000, shop1YrGROUP := "50001
-100000"]
churn_data[SHOP1YR >=100000 & SHOP1YR <200000, shop1YrGROUP := "1000
01-200000"]
churn_data[SHOP1YR >=200000 & SHOP1YR <300000, shop1YrGROUP := "2000
01-300000"]
churn_data[SHOP1YR >=300000 & SHOP1YR <400000, shop1YrGROUP := "3000
1-400000"]
churn_data[SHOP1YR >=400000 & SHOP1YR <500000, shop1YrGROUP := "4000
1-500000"]
churn_data[SHOP1YR >=600000 & SHOP1YR <700000, shop1YrGROUP := "6000
1-700000"]
churn_data[SHOP1YR >=700000 & SHOP1YR <800000, shop1YrGROUP := "7000
01-800000"]
churn_data[SHOP1YR >=800000 & SHOP1YR <900000, shop1YrGROUP := "8000
01-900000"]
churn_data[SHOP1YR >=900000 & SHOP1YR <1000000, shop1YrGROUP := "900
001-1000000"]
churn_data[SHOP1YR >=1000000 & SHOP1YR<1100000, shop1YrGROUP := "100
0001-1100000"]
churn_data[SHOP1YR >=1100000 & SHOP1YR<1200000, shop1YrGROUP := "110
```

```
0001-1200000"]
churn_data[SHOP1YR >=1200000 & SHOP1YR<1300000, shop1YrGROUP := "120
0001-1300000"]
churn_data[SHOP1YR >=1300000 & SHOP1YR<1400000, shop1YrGROUP := "130
0001-1400000"]
churn_data[SHOP1YR >=1400000 & SHOP1YR<1500000, shop1YrGROUP := "140
0001-1500000"]
churn_data[SHOP1YR >=1500000 & SHOP1YR<1600000, shop1YrGROUP := "150
0001-1600000"]
churn_data[SHOP1YR >=1600000 & SHOP1YR<1700000, shop1YrGROUP := "160
0001-1700000"]
churn_data[SHOP1YR >=1700000 & SHOP1YR<1800000, shop1YrGROUP := "170
0001-1800000"]
churn_data[SHOP1YR >=1800000 & SHOP1YR<1900000, shop1YrGROUP := "180
0001-1900000"]
churn_data[SHOP1YR >=1900000, shop1YrGROUP := "1900001+"]
```

```
churn_data1 <- churn_data[,-c(2,8,12,13, 17:22)]
```

```
sapply(churn_data1, function(x) sum(is.na(x)))
```

```
##          RENEW       A2ACCTYP       M2EXCFLG       B2BUSTYP
F2HOMRGN
##              0              0              0              0
0
##       F2HOMFCY         TENURE        ZIPCODE       MBRCOUNT HOMEF
CTYCHANGE
##              0              0              0              0
0
##   RECENTMOVING        SHOP1YR       agegroup     earlygroup  DIST
ANCEGroup
##              0              0              0              0
0
##   shop1YrGROUP
##              1
```

```
churn_data1<- na.omit(churn_data1)
```

```
churn_data1[,1:16] <- lapply(churn_data1[,1:16], factor)
```

s

```
str(churn_data)
```

```
## Classes 'data.table' and 'data.frame':   120450 obs. of  26 varia
bles:
##  $ RENEW         : chr  "N" "N" "N" "N" ...
##  $ A2ACCIPK      : num  280928 280100 279886 279912 279896 ...
##  $ A2ACCTYP      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ M2EXCFLG      : chr  "N" "N" "N" "N" ...
##  $ B2BUSTYP      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ F2HOMRGN      : chr  "NE" "BO" "BO" "SE" ...
##  $ F2HOMFCY      : num  1078 847 847 185 472 ...
##  $ AGE           : num  42 61 52 32 46 36 34 45 52 32 ...
##  $ TENURE        : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ ZIPCODE       : num  20715 77346 91024 32789 93960 ...
##  $ MBRCOUNT      : num  2 2 2 2 2 1 2 2 2 2 ...
##  $ DISTANCE      : num  7.53 6.05 7.89 3.43 26.29 ...
##  $ EARLYFAREWELL : int  75 320 350 137 41 53 38 363 53 64 ...
##  $ HOMEFCTYCHANGE: chr  "Y" "N" "N" "N" ...
##  $ RECENTMOVING  : chr  "N" "N" "N" "N" ...
##  $ SHOP1YR       : num  1385 3500 114 997 12579 ...
##  $ SHOP6M        : num  827.7 0 0 23.2 73 ...
##  $ SHOP3M        : num  253 0 0 0 73 ...
##  $ ECOMSHOP      : num  0 1 0 0 0 0 0 0 0 0 ...
##  $ GASSHOP       : num  0.0293 0 0 0.0251 0 ...
##  $ MEDICALSHOP   : num  0.0173 0 0.30377 0 0.00818 ...
##  $ GROCERYSHOP   : num  0.523 0 0.234 0.405 0.936 ...
##  $ agegroup      : chr  "40-44" "60-64" "50-54" "30-34" ...
##  $ earlygroup    : chr  "61 - 120" "301 - 360" "301 - 360" "121 -
180" ...
##  $ DISTANCEGroup : chr  "Less then 10" "Less then 10" "Less then
10" "Less then 10" ...
##  $ shop1YrGROUP  : chr  "1001-5000" "1001-5000" "0-1000" "0-1000"
...
##  - attr(*, "spec")=
##   .. cols(
##   ..   RENEW = col_character(),
```

```
##   ..   A2ACCIPK = col_double(),
##   ..   A2ACCTYP = col_double(),
##   ..   M2EXCFLG = col_character(),
##   ..   B2BUSTYP = col_double(),
##   ..   F2HOMRGN = col_character(),
##   ..   F2HOMFCY = col_double(),
##   ..   AGE = col_double(),
##   ..   TENURE = col_double(),
##   ..   ZIPCODE = col_double(),
##   ..   MBRCOUNT = col_double(),
##   ..   DISTANCE = col_double(),
##   ..   EARLYFAREWELL = col_double(),
##   ..   HOMEFCTYCHANGE = col_character(),
##   ..   RECENTMOVING = col_character(),
##   ..   SHOP1YR = col_double(),
##   ..   SHOP6M = col_double(),
##   ..   SHOP3M = col_double(),
##   ..   ECOMSHOP = col_double(),
##   ..   GASSHOP = col_double(),
##   ..   MEDICALSHOP = col_double(),
##   ..   GROCERYSHOP = col_double()
##   .. )
##  - attr(*, ".internal.selfref")=<externalptr>
```

Frequent Itemset Generation: Find all frequent item-sets with support >= pre-determined min_support count

```
library(arules)
library(arulesViz)
NotR_rules <- apriori(data=churn_data1, parameter=list (supp=0.048,c
onf = 0.9), appearance = list (rhs='RENEW=N'))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime suppor
t minlen
##         0.9    0.1    1 none FALSE            TRUE       5   0.04
8      1
##  maxlen target   ext
##      10  rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 5781
##
## set item appearances ...[1 item(s)] done [0.00s].
## set transactions ...[32136 item(s), 120449 transaction(s)] done [
0.58s].
## sorting and recoding items ... [40 item(s)] done [0.02s].
## creating transaction tree ... done [0.09s].
## checking subsets of size 1 2 3 4 5 6 7 8 9 10
```

```
## Warning in apriori(data = churn_data1, parameter = list(supp = 0.
048, conf
## = 0.9), : Mining stopped (maxlen reached). Only patterns up to a
length of
## 10 returned!
```

```
##  done [0.46s].
## writing ... [16 rule(s)] done [0.00s].
## creating S4 object  ... done [0.05s].
```

Confidence(A=>B) = P(B|A) = P(A and B)/ P(A)

Lift(A=>B = Support/(Supp(A)Supp(B))

```
inspect(head(sort(NotR_rules, by = "count"), 15))
```

```
##      lhs                          rhs          support confidence
lift count
## [1]  {earlygroup=301 - 360} => {RENEW=N} 0.05309301  0.9167144 1.
623737  6395
## [2]  {TENURE=1,
##       earlygroup=301 - 360} => {RENEW=N} 0.05309301  0.9167144 1.
623737  6395
## [3]  {HOMEFCTYCHANGE=N,
##       earlygroup=301 - 360} => {RENEW=N} 0.05188918  0.9187123 1.
627275  6250
## [4]  {TENURE=1,
##       HOMEFCTYCHANGE=N,
##       earlygroup=301 - 360} => {RENEW=N} 0.05188918  0.9187123 1.
627275  6250
## [5]  {B2BUSTYP=0,
##       earlygroup=301 - 360} => {RENEW=N} 0.04961436  0.9162833 1.
622973  5976
## [6]  {B2BUSTYP=0,
##       TENURE=1,
##       earlygroup=301 - 360} => {RENEW=N} 0.04961436  0.9162833 1.
622973  5976
## [7]  {A2ACCTYP=1,
##       earlygroup=301 - 360} => {RENEW=N} 0.04907471  0.9171451 1.
624499  5911
## [8]  {A2ACCTYP=1,
##       B2BUSTYP=0,
##       earlygroup=301 - 360} => {RENEW=N} 0.04907471  0.9171451 1.
624499  5911
## [9]  {A2ACCTYP=1,
##       TENURE=1,
##       earlygroup=301 - 360} => {RENEW=N} 0.04907471  0.9171451 1.
624499  5911
## [10] {A2ACCTYP=1,
##       B2BUSTYP=0,
##       TENURE=1,
##       earlygroup=301 - 360} => {RENEW=N} 0.04907471  0.9171451 1.
624499  5911
## [11] {earlygroup=301 - 360,
##       shop1YrGROUP=0-1000}  => {RENEW=N} 0.04871771  0.9214824 1.
632182  5868
## [12] {TENURE=1,
```

```
##       earlygroup=301 - 360,
##       shop1YrGROUP=0-1000}  => {RENEW=N} 0.04871771  0.9214824 1.
632182  5868
## [13] {RECENTMOVING=N,
##       earlygroup=301 - 360} => {RENEW=N} 0.04853506  0.9137230 1.
618438  5846
## [14] {TENURE=1,
##       RECENTMOVING=N,
##       earlygroup=301 - 360} => {RENEW=N} 0.04853506  0.9137230 1.
618438  5846
## [15] {B2BUSTYP=0,
##       HOMEFCTYCHANGE=N,
##       earlygroup=301 - 360} => {RENEW=N} 0.04851846  0.9185791 1.
627039  5844
```

```
library(arules)
library(arulesViz)
RD_rules <- apriori(data=churn_data1, parameter=list (supp=0.002,con
f = 0.827), appearance = list (rhs='shop1YrGROUP=1001-5000'))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime suppor
t minlen
##       0.827    0.1    1 none FALSE            TRUE       5   0.00
2      1
##  maxlen target   ext
##      10  rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 240
##
## set item appearances ...[1 item(s)] done [0.00s].
## set transactions ...[32136 item(s), 120449 transaction(s)] done [
0.46s].
## sorting and recoding items ... [265 item(s)] done [0.02s].
## creating transaction tree ... done [0.08s].
## checking subsets of size 1 2 3 4 5 6 7 8 9 10
```

```
## Warning in apriori(data = churn_data1, parameter = list(supp = 0.
002, conf
## = 0.827), : Mining stopped (maxlen reached). Only patterns up to
a length
## of 10 returned!
```

```
##  done [4.55s].
## writing ... [19 rule(s)] done [0.02s].
## creating S4 object  ... done [0.10s].
```

```
inspect(head(sort(RD_rules, by = "confidence"), 20))
```

```
##      lhs                              rhs
support confidence     lift count
## [1]  {RENEW=Y,
##       A2ACCTYP=1,
```

```
##       M2EXCFLG=N,
##       F2HOMRGN=MW,
##       MBRCOUNT=2,
##       RECENTMOVING=N,
##       agegroup=25-29,
##       earlygroup=0 - 60}          => {shop1YrGROUP=1001-5000} 0.0
02092172  0.8400000 1.530877   252
## [2]  {RENEW=Y,
##       A2ACCTYP=1,
##       M2EXCFLG=N,
##       B2BUSTYP=0,
##       F2HOMRGN=MW,
##       MBRCOUNT=2,
##       RECENTMOVING=N,
##       agegroup=25-29,
##       earlygroup=0 - 60}          => {shop1YrGROUP=1001-5000} 0.0
02092172  0.8400000 1.530877   252
## [3]  {RENEW=Y,
##       A2ACCTYP=1,
##       M2EXCFLG=N,
##       F2HOMRGN=MW,
##       TENURE=1,
##       MBRCOUNT=2,
##       RECENTMOVING=N,
##       agegroup=25-29,
##       earlygroup=0 - 60}          => {shop1YrGROUP=1001-5000} 0.0
02083870  0.8394649 1.529901   251
## [4]  {RENEW=Y,
##       M2EXCFLG=N,
##       B2BUSTYP=0,
##       F2HOMRGN=MW,
##       MBRCOUNT=2,
##       RECENTMOVING=N,
##       agegroup=25-29,
##       earlygroup=0 - 60}          => {shop1YrGROUP=1001-5000} 0.0
02092172  0.8372093 1.525791   252
## [5]  {RENEW=Y,
##       M2EXCFLG=N,
##       B2BUSTYP=0,
##       F2HOMRGN=MW,
##       TENURE=1,
##       MBRCOUNT=2,
```

```
##       RECENTMOVING=N,
##       agegroup=25-29,
##       earlygroup=0 - 60}          => {shop1YrGROUP=1001-5000} 0.0
02083870  0.8366667 1.524802   251
## [6]  {RENEW=Y,
##       A2ACCTYP=1,
##       M2EXCFLG=N,
##       F2HOMRGN=MW,
##       MBRCOUNT=2,
##       agegroup=25-29,
##       earlygroup=0 - 60}          => {shop1YrGROUP=1001-5000} 0.0
02440867  0.8352273 1.522178   294
## [7]  {RENEW=Y,
##       A2ACCTYP=1,
##       M2EXCFLG=N,
##       B2BUSTYP=0,
##       F2HOMRGN=MW,
##       MBRCOUNT=2,
##       agegroup=25-29,
##       earlygroup=0 - 60}          => {shop1YrGROUP=1001-5000} 0.0
02440867  0.8352273 1.522178   294
## [8]  {RENEW=Y,
##       M2EXCFLG=N,
##       F2HOMRGN=MW,
##       MBRCOUNT=2,
##       RECENTMOVING=N,
##       agegroup=25-29,
##       earlygroup=0 - 60}          => {shop1YrGROUP=1001-5000} 0.0
02100474  0.8349835 1.521734   253
## [9]  {RENEW=Y,
##       A2ACCTYP=1,
##       M2EXCFLG=N,
##       F2HOMRGN=MW,
##       TENURE=1,
##       MBRCOUNT=2,
##       agegroup=25-29,
##       earlygroup=0 - 60}          => {shop1YrGROUP=1001-5000} 0.0
02432565  0.8347578 1.521323   293
## [10] {RENEW=Y,
##       A2ACCTYP=1,
##       M2EXCFLG=N,
##       B2BUSTYP=0,
```

```
##       F2HOMRGN=MW,
##       TENURE=1,
##       MBRCOUNT=2,
##       agegroup=25-29,
##       earlygroup=0 - 60}          => {shop1YrGROUP=1001-5000} 0.0
02432565  0.8347578 1.521323   293
## [11] {RENEW=Y,
##       M2EXCFLG=N,
##       F2HOMRGN=MW,
##       TENURE=1,
##       MBRCOUNT=2,
##       RECENTMOVING=N,
##       agegroup=25-29,
##       earlygroup=0 - 60}          => {shop1YrGROUP=1001-5000} 0.0
02092172  0.8344371 1.520738   252
## [12] {RENEW=Y,
##       M2EXCFLG=N,
##       B2BUSTYP=0,
##       F2HOMRGN=MW,
##       MBRCOUNT=2,
##       agegroup=25-29,
##       earlygroup=0 - 60}          => {shop1YrGROUP=1001-5000} 0.0
02449169  0.8333333 1.518727   295
## [13] {RENEW=Y,
##       M2EXCFLG=N,
##       B2BUSTYP=0,
##       F2HOMRGN=MW,
##       TENURE=1,
##       MBRCOUNT=2,
##       agegroup=25-29,
##       earlygroup=0 - 60}          => {shop1YrGROUP=1001-5000} 0.0
02440867  0.8328612 1.517866   294
## [14] {RENEW=Y,
##       M2EXCFLG=N,
##       F2HOMRGN=MW,
##       MBRCOUNT=2,
##       agegroup=25-29,
##       earlygroup=0 - 60}          => {shop1YrGROUP=1001-5000} 0.0
02457472  0.8314607 1.515314   296
## [15] {RENEW=Y,
##       M2EXCFLG=N,
##       F2HOMRGN=MW,
```

```
##       TENURE=1,
##       MBRCOUNT=2,
##       agegroup=25-29,
##       earlygroup=0 - 60}          => {shop1YrGROUP=1001-5000} 0.0
02449169  0.8309859 1.514449   295
## [16] {RENEW=Y,
##       M2EXCFLG=N,
##       F2HOMRGN=MW,
##       MBRCOUNT=2,
##       agegroup=30-34,
##       earlygroup=0 - 60,
##       DISTANCEGroup=Less then 10} => {shop1YrGROUP=1001-5000} 0.0
02150288  0.8274760 1.508052   259
## [17] {RENEW=Y,
##       M2EXCFLG=N,
##       F2HOMRGN=MW,
##       TENURE=1,
##       MBRCOUNT=2,
##       agegroup=30-34,
##       earlygroup=0 - 60,
##       DISTANCEGroup=Less then 10} => {shop1YrGROUP=1001-5000} 0.0
02150288  0.8274760 1.508052   259
## [18] {RENEW=Y,
##       M2EXCFLG=N,
##       B2BUSTYP=0,
##       F2HOMRGN=MW,
##       MBRCOUNT=2,
##       agegroup=30-34,
##       earlygroup=0 - 60,
##       DISTANCEGroup=Less then 10} => {shop1YrGROUP=1001-5000} 0.0
02108776  0.8273616 1.507843   254
## [19] {RENEW=Y,
##       M2EXCFLG=N,
##       B2BUSTYP=0,
##       F2HOMRGN=MW,
##       TENURE=1,
##       MBRCOUNT=2,
##       agegroup=30-34,
##       earlygroup=0 - 60,
##       DISTANCEGroup=Less then 10} => {shop1YrGROUP=1001-5000} 0.0
02108776  0.8273616 1.507843   254
```