5200Appendix1_CodeandOutput

Group of NBA Shot Result Analysis

October 22, 2017

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(car)
```

```
##
## Attaching package: 'car'
```

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

```
library(RcmdrMisc)
```

```
## Loading required package: sandwich
```

```
library(stringr)

#####I. Data import

##Dataset1
shot=read.csv("shot_logs.csv")
##supplementary data
players=read.csv("players_stats.csv")

#####II. Dataset1 diagnosis & cleaning

##1. Wrong data types / Too many levels
#Diagnose wrong data type:
#GAME_ID,GAME_CLOCK,CLOSEST_DEFENDER,CLOSEST_DEFENDER_PLAYER_ID,PLAYER_NAME,PLAYER_IDshould be charact
er.
#PERIOD,PTS_TYPE should be factor.
str(shot)
```

```
## 'data.frame':    128069 obs. of  21 variables:
##  $ GAME_ID                  : int  21400899 21400899 21400899 21400899 21400899 21400899 21400899
 21400899 21400899 21400890 ...
##  $ MATCHUP                  : Factor w/ 1808 levels "DEC 01, 2014 - DEN @ UTA",..: 1291 1291 1291
 1291 1291 1291 1291 1291 1291 1277 ...
##  $ LOCATION                 : Factor w/ 2 levels "A","H": 1 1 1 1 1 1 1 1 1 2 ...
##  $ W                        : Factor w/ 2 levels "L","W": 2 2 2 2 2 2 2 2 2 2 ...
##  $ FINAL_MARGIN             : int  24 24 24 24 24 24 24 24 24 1 ...
##  $ SHOT_NUMBER              : int  1 2 3 4 5 6 7 8 9 1 ...
##  $ PERIOD                   : int  1 1 1 2 2 2 4 4 4 2 ...
##  $ GAME_CLOCK               : Factor w/ 719 levels "0:00","0:01",..: 70 15 1 228 155 615 136 600 4
34 213 ...
##  $ SHOT_CLOCK               : num  10.8 3.4 NA 10.3 10.9 9.1 14.5 3.4 12.4 17.4 ...
##  $ DRIBBLES                 : int  2 0 3 2 2 2 11 3 0 0 ...
##  $ TOUCH_TIME               : num  1.9 0.8 2.7 1.9 2.7 4.4 9 2.5 0.8 1.1 ...
##  $ SHOT_DIST                : num  7.7 28.2 10.1 17.2 3.7 18.4 20.7 3.5 24.6 22.4 ...
##  $ PTS_TYPE                 : int  2 3 2 2 2 2 2 2 3 3 ...
##  $ SHOT_RESULT              : Factor w/ 2 levels "made","missed": 1 2 2 2 2 2 2 1 2 2 ...
##  $ CLOSEST_DEFENDER         : Factor w/ 473 levels "Acy, Quincy",..: 15 51 51 62 471 456 219 351 3
14 132 ...
##  $ CLOSEST_DEFENDER_PLAYER_ID: int  101187 202711 202711 203900 201152 101114 101127 203486 202721
 201961 ...
##  $ CLOSE_DEF_DIST           : num  1.3 6.1 0.9 3.4 1.1 2.6 6.1 2.1 7.3 19.8 ...
##  $ FGM                      : int  1 0 0 0 0 0 0 1 0 0 ...
##  $ PTS                      : int  2 0 0 0 0 0 0 2 0 0 ...
##  $ player_name              : Factor w/ 281 levels "aaron brooks",..: 36 36 36 36 36 36 36 36 36 3
6 ...
##  $ player_id                : int  203148 203148 203148 203148 203148 203148 203148 203148 203148
 203148 ...
```

```
#Process wrong data type
shot$GAME_CLOCK<-as.character(shot$GAME_CLOCK)
shot$GAME_CLOCK<-as.numeric(gsub(pattern = ":", replacement = ".", shot$GAME_CLOCK))
shot[, c(7,13,18)] <- lapply(shot[, c(7,13,18), drop=FALSE],
                             as.factor)
shot[, c(1,2,8,15,16,20,21)] <- lapply(shot[, c(1,2,8,15,16,20,21), drop=FALSE],
                                       as.character)

#check
str(shot)
```

```
## 'data.frame':    128069 obs. of  21 variables:
##  $ GAME_ID                  : chr  "21400899" "21400899" "21400899" "21400899" ...
##  $ MATCHUP                  : chr  "MAR 04, 2015 - CHA @ BKN" "MAR 04, 2015 - CHA @ BKN" "MAR 04,
 2015 - CHA @ BKN" "MAR 04, 2015 - CHA @ BKN" ...
##  $ LOCATION                 : Factor w/ 2 levels "A","H": 1 1 1 1 1 1 1 1 1 2 ...
##  $ W                        : Factor w/ 2 levels "L","W": 2 2 2 2 2 2 2 2 2 2 ...
##  $ FINAL_MARGIN             : int  24 24 24 24 24 24 24 24 24 1 ...
##  $ SHOT_NUMBER              : int  1 2 3 4 5 6 7 8 9 1 ...
##  $ PERIOD                   : Factor w/ 7 levels "1","2","3","4",..: 1 1 1 2 2 2 4 4 4 2 ...
##  $ GAME_CLOCK               : chr  "1:09" "0:14" "0:0" "11:47" ...
##  $ SHOT_CLOCK               : num  10.8 3.4 NA 10.3 10.9 9.1 14.5 3.4 12.4 17.4 ...
##  $ DRIBBLES                 : int  2 0 3 2 2 2 11 3 0 0 ...
##  $ TOUCH_TIME               : num  1.9 0.8 2.7 1.9 2.7 4.4 9 2.5 0.8 1.1 ...
##  $ SHOT_DIST                : num  7.7 28.2 10.1 17.2 3.7 18.4 20.7 3.5 24.6 22.4 ...
##  $ PTS_TYPE                 : Factor w/ 2 levels "2","3": 1 2 1 1 1 1 1 1 2 2 ...
##  $ SHOT_RESULT              : Factor w/ 2 levels "made","missed": 1 2 2 2 2 2 2 1 2 2 ...
##  $ CLOSEST_DEFENDER         : chr  "Anderson, Alan" "Bogdanovic, Bojan" "Bogdanovic, Bojan" "Brow
n, Markel" ...
##  $ CLOSEST_DEFENDER_PLAYER_ID: chr  "101187" "202711" "202711" "203900" ...
##  $ CLOSE_DEF_DIST           : num  1.3 6.1 0.9 3.4 1.1 2.6 6.1 2.1 7.3 19.8 ...
##  $ FGM                      : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 2 1 1 ...
##  $ PTS                      : int  2 0 0 0 0 0 0 2 0 0 ...
##  $ player_name              : chr  "brian roberts" "brian roberts" "brian roberts" "brian roberts"
...
##  $ player_id                : chr  "203148" "203148" "203148" "203148" ...
```

```
##2. Missing Value
#Diagnose missing value
sapply(shot, function(x)(sum(is.na(x))))
```

```
##                    GAME_ID                     MATCHUP
##                          0                           0
##                   LOCATION                           W
##                          0                           0
##               FINAL_MARGIN                 SHOT_NUMBER
##                          0                           0
##                     PERIOD                  GAME_CLOCK
##                          0                           0
##                 SHOT_CLOCK                    DRIBBLES
##                       5567                           0
##                 TOUCH_TIME                   SHOT_DIST
##                          0                           0
##                   PTS_TYPE                 SHOT_RESULT
##                          0                           0
##           CLOSEST_DEFENDER CLOSEST_DEFENDER_PLAYER_ID
##                          0                           0
##              CLOSE_DEF_DIST                         FGM
##                          0                           0
##                         PTS                 player_name
##                          0                           0
##                   player_id
##                          0
```

```
sum(is.na(shot$SHOT_CLOCK))
```

```
## [1] 5567
```

```
#Process missing value
shotnona<-na.omit(shot)
#Check
sum(is.na(shotnona$SHOT_CLOCK))
```

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

```
##3. Invalid Data:
#Diagnose invalid Data: TOUCH_TIME<0
summary(shotnona)
```

```
##    GAME_ID            MATCHUP          LOCATION  W
##  Length:122502      Length:122502      A:61315   L:60353
##  Class :character   Class :character   H:61187   W:62149
##  Mode  :character   Mode  :character
##
##
##
##
##   FINAL_MARGIN       SHOT_NUMBER     PERIOD      GAME_CLOCK
##  Min.    :-53.0000   Min.    : 1.000   1:32383   Length:122502
##  1st Qu.: -8.0000   1st Qu.: 3.000   2:30148   Class :character
##  Median :  1.0000   Median : 5.000   3:31032   Mode  :character
##  Mean   :  0.2524   Mean   : 6.476   4:27941
##  3rd Qu.:  9.0000   3rd Qu.: 9.000   5:  809
##  Max.    : 53.0000   Max.    :37.000   6:  149
##                                       7:   40
##    SHOT_CLOCK        DRIBBLES         TOUCH_TIME          SHOT_DIST
##  Min.    : 0.00   Min.    : 0.000   Min.    :-100.500   Min.    : 0.00
##  1st Qu.: 8.20   1st Qu.: 0.000   1st Qu.:   0.900   1st Qu.: 4.70
##  Median :12.30   Median : 1.000   Median :   1.600   Median :13.40
##  Mean   :12.45   Mean   : 1.989   Mean   :   2.748   Mean   :13.44
##  3rd Qu.:16.68   3rd Qu.: 2.000   3rd Qu.:   3.700   3rd Qu.:22.40
##  Max.    :24.00   Max.    :32.000   Max.    :  24.900   Max.    :43.50
##
##  PTS_TYPE  SHOT_RESULT    CLOSEST_DEFENDER   CLOSEST_DEFENDER_PLAYER_ID
##  2:90852   made  :55880   Length:122502      Length:122502
##  3:31650   missed:66622   Class :character   Class :character
##                           Mode  :character   Mode  :character
##
##
##
##
##  CLOSE_DEF_DIST   FGM            PTS         player_name
##  Min.    : 0.000   0:66622   Min.    :0.000   Length:122502
##  1st Qu.: 2.300   1:55880   1st Qu.:0.000   Class :character
##  Median : 3.700             Median :0.000   Mode  :character
##  Mean   : 4.122             Mean   :1.005
##  3rd Qu.: 5.300             3rd Qu.:2.000
##  Max.    :53.200             Max.    :3.000
##
##   player_id
##  Length:122502
##  Class :character
##  Mode  :character
##
##
##
##
```

```
range(shotnona$TOUCH_TIME)
```

```
## [1] -100.5   24.9
```

```
#Process invalid data
shotnona<-dplyr::filter(shotnona, TOUCH_TIME <= 24 &  TOUCH_TIME >= 0)
#Check
range(shotnona$TOUCH_TIME)
```

```
## [1]  0.0 23.9
```

```
##4. Mismatching Data
#Diagnose mismatching data: "PTS_TYPE" & "SHOT_DIST“
count(shotnona[shotnona$PTS_TYPE == 3 & shotnona$SHOT_DIST <= 23.75,])
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  9136
```

```
count(shotnona[shotnona$PTS_TYPE == 2 & shotnona$SHOT_DIST > 23.75,])
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   508
```

```
#Process mismatching data
shotnona[shotnona$SHOT_DIST >= 23.75,]$PTS_TYPE = 3
shotnona[shotnona$SHOT_DIST < 23.75,]$PTS_TYPE = 2
#Check
count(shotnona[shotnona$PTS_TYPE == 3 & shotnona$SHOT_DIST <= 23.75,])
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1     0
```

```
count(shotnona[shotnona$PTS_TYPE == 2 & shotnona$SHOT_DIST > 23.75,])
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1     0
```

```
##5.Typos
#Diagnose typos in "Players name": We used names in supplement dataset to check the typos.
#Note: We have already checked the validity of names in supplement dataset using external resources: h
ttps://www.basketball-reference.com/players/.
NAME<-data.frame(unique(toupper(shotnona$player_name)))
ALLNAME<-unique(toupper(players$Name))
all(NAME$unique.toupper.shotnona.player_name..%in%ALLNAME)
```

```
## [1] FALSE
```

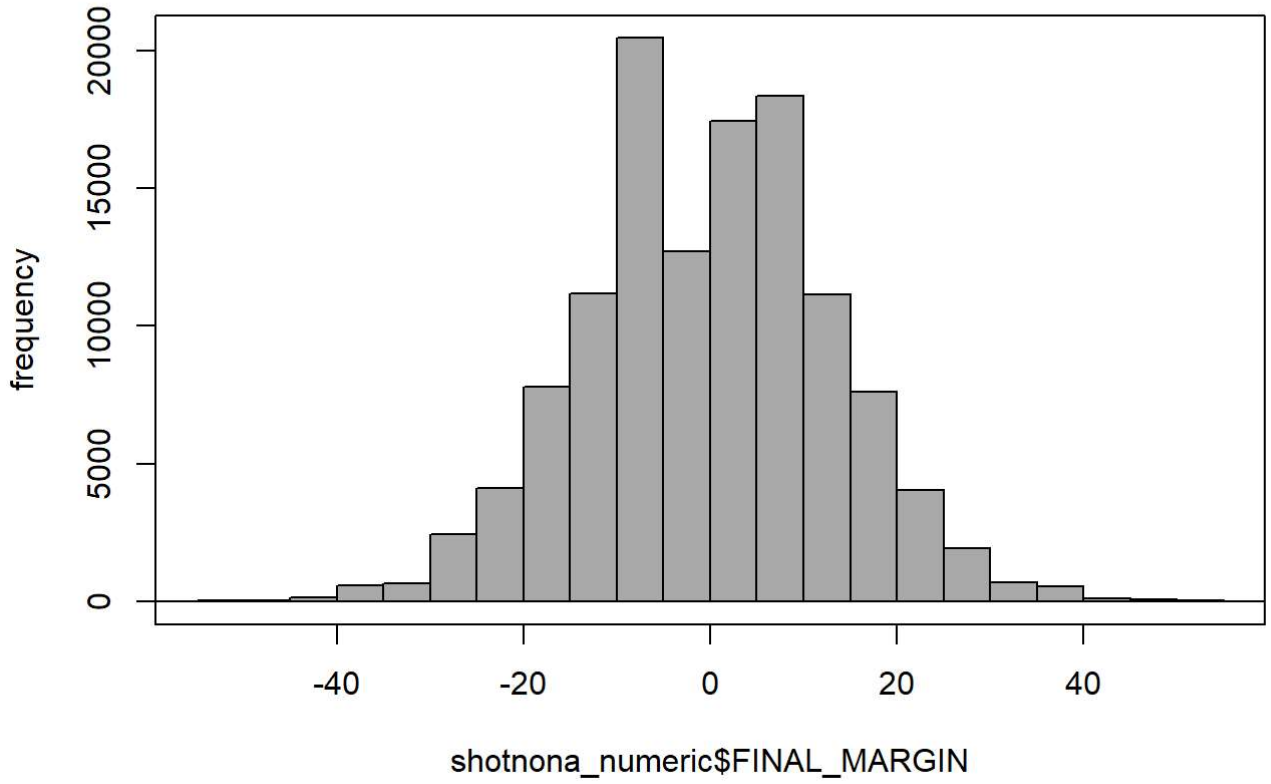
```
NAME$Judge<-ifelse(NAME$unique.toupper.shotnona.player_name.. %in% ALLNAME, 1, 2)
print(Typos<-NAME[NAME$Judge==2,])
```

```
##     unique.toupper.shotnona.player_name.. Judge
## 17                              JON INGLES     2
## 50                             OTTO PORTER     2
## 58                            NENE HILARIO     2
## 68                          JIMMER DREDETTE     2
## 78                               MNTA ELLIS     2
## 80                          JOSE JUAN BAREA     2
## 84                          AL FAROUQ AMINU     2
## 87                           DIRK NOWTIZSKI     2
## 94                              KYLE OQUINN     2
## 110                               CJ WATSON     2
## 143                         DANILO GALLINAI     2
## 155                             NERLES NOEL     2
## 166                                 OJ MAYO     2
## 178                               BENO URDIH     2
## 185                        AMARE STOUDEMIRE     2
## 189                        TIME HARDAWAY JR     2
## 192                             DJ AUGUSTIN     2
## 211                             DWAYNE WADE     2
## 213                             JAMES ENNIS     2
## 235                             STEVE ADAMS     2
## 259                             ALAN CRABBE     2
```

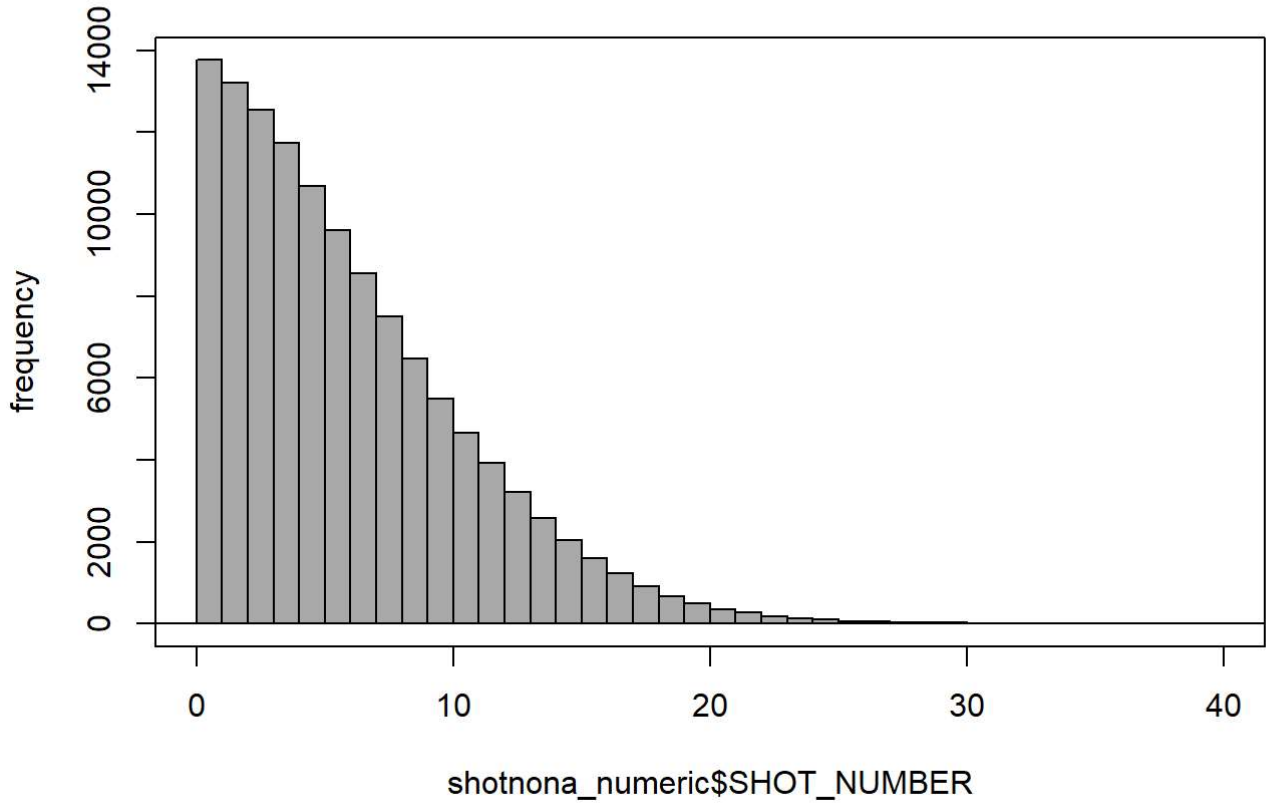
```
#Process typos: Some typos happened as spelling mistakes, while some typos happened because of the for
mat.
shotnona$player_name<-toupper(shotnona$player_name)
shotnona[shotnona$player_name == "JON INGLES",]$player_name  = "JOE INGLES"
shotnona[shotnona$player_name == "JIMMER DREDETTE",]$player_name  = "JIMMER FREDETTE"
shotnona[shotnona$player_name == "MNTA ELLIS",]$player_name  = "MONTA ELLIS"
shotnona[shotnona$player_name == "DIRK NOWTIZSKI",]$player_name  = "DIRK NOWITZKI"
shotnona[shotnona$player_name == "KYLE OQUINN",]$player_name  = "KYLE O'QUINN"
shotnona[shotnona$player_name == "DANILO GALLINAI",]$player_name  = "DANILO GALLINARI"
shotnona[shotnona$player_name == "NERLES NOEL",]$player_name  = "NERLENS NOEL"
shotnona[shotnona$player_name == "BENO URDIH",]$player_name  = "BENO UDRIH"
shotnona[shotnona$player_name == "OTTO PORTER",]$player_name  = "OTTO PORTER JR."
shotnona[shotnona$player_name == "NENE HILARIO",]$player_name  = "NENE"
shotnona[shotnona$player_name == "JOSE JUAN BAREA",]$player_name  = "J.J. BAREA"
shotnona[shotnona$player_name == "AL FAROUQ AMINU",]$player_name  = "AL-FAROUQ AMINU"
shotnona[shotnona$player_name == "CJ WATSON",]$player_name  = "C.J. WATSON"
shotnona[shotnona$player_name == "OJ MAYO",]$player_name  = "O.J. MAYO"
shotnona[shotnona$player_name == "AMARE STOUDEMIRE",]$player_name  = "AMAR'E STOUDEMIRE"
shotnona[shotnona$player_name == "TIME HARDAWAY JR",]$player_name  = "TIM HARDAWAY JR."
shotnona[shotnona$player_name == "DJ AUGUSTIN",]$player_name  = "D.J. AUGUSTIN"
shotnona[shotnona$player_name == "DWAYNE WADE",]$player_name  = "DWYANE WADE"
shotnona[shotnona$player_name == "JAMES ENNIS",]$player_name  = "JAMES ENNIS III"
shotnona[shotnona$player_name == "STEVE ADAMS",]$player_name  = "STEVEN ADAMS"
shotnona[shotnona$player_name == "ALAN CRABBE",]$player_name  = "ALLEN CRABBE"
#Check
all(unique(shotnona$player_name) %in% ALLNAME)
```

```
## [1] TRUE
```

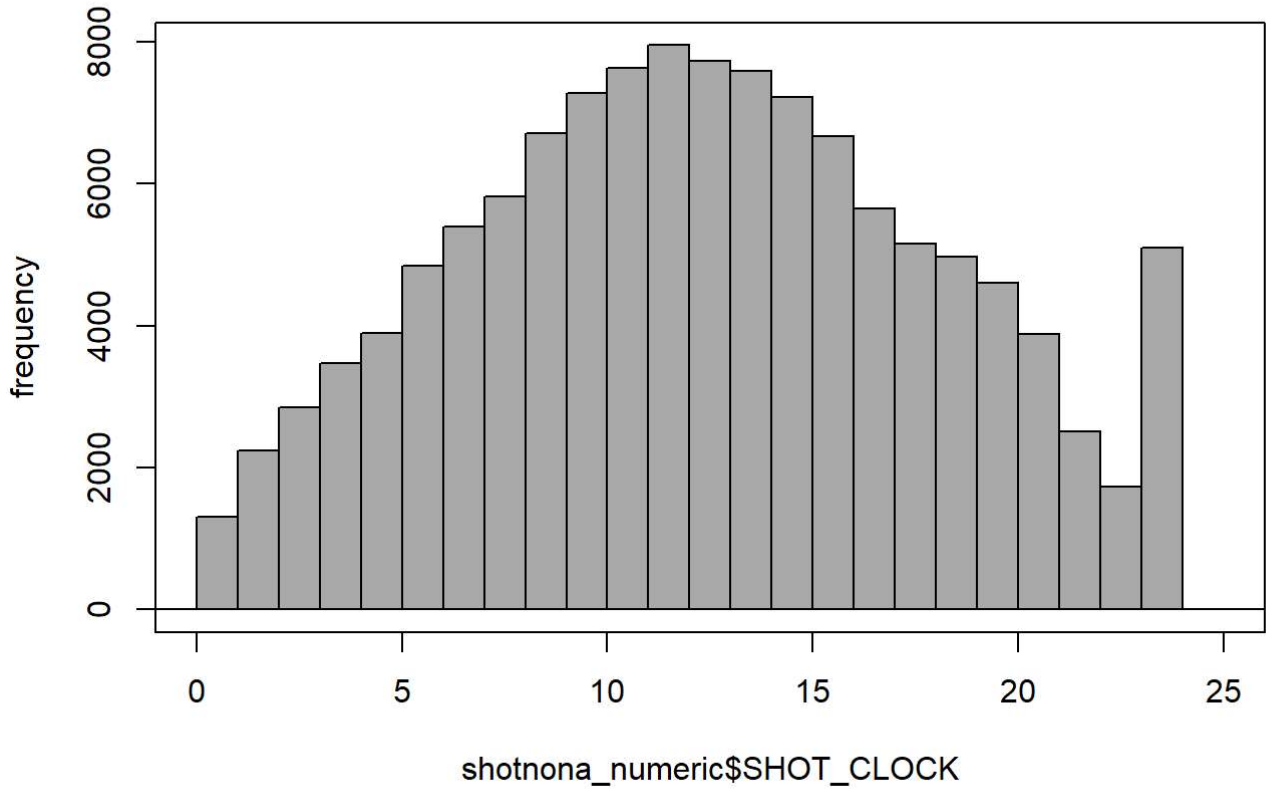
```
##6.strange distributions. We plot the distribution of numeric variables to see the distributions.
shotnona_numeric<-shotnona[,c(5,6,9,10,11,12,17)]
Hist(shotnona_numeric$FINAL_MARGIN, scale="frequency", breaks="Sturges", col="darkgray")
```



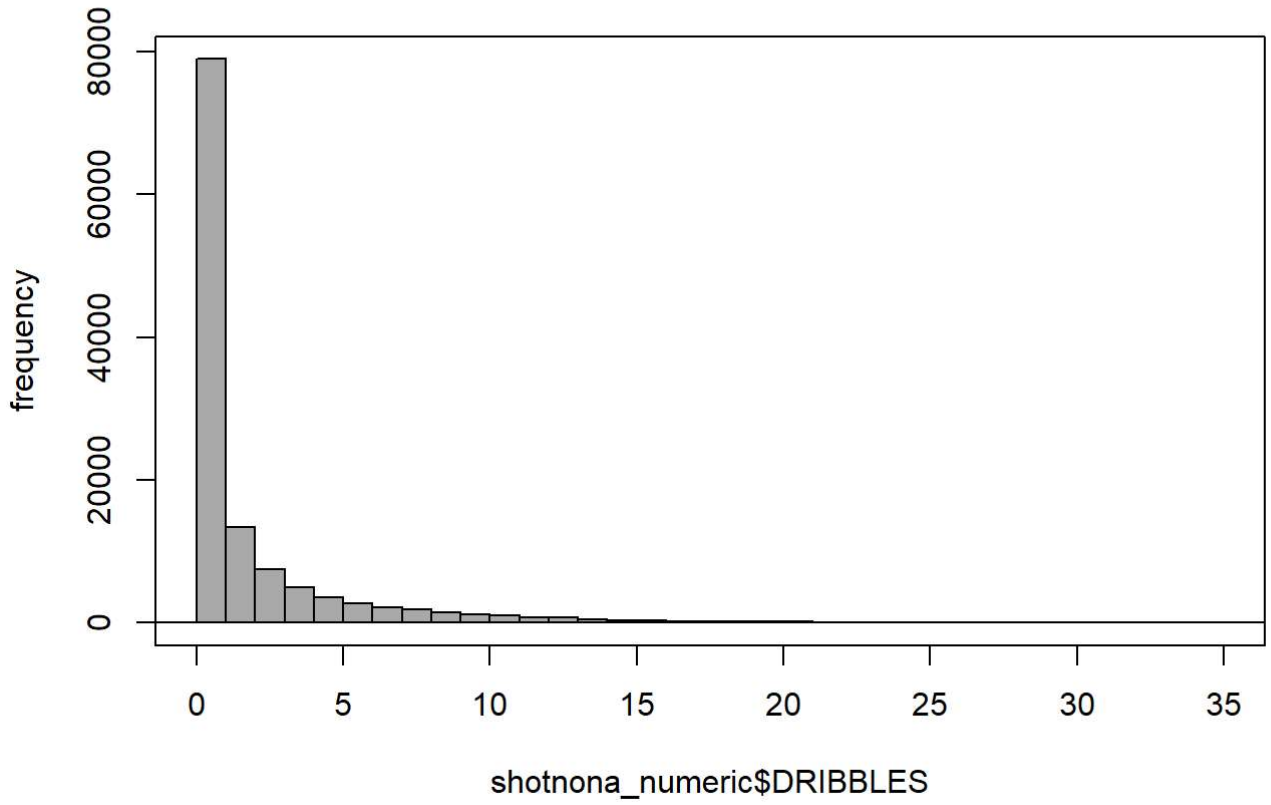

```
Hist(shotnona_numeric$SHOT_NUMBER, scale="frequency", breaks=seq(0,40,1), col="darkgray")
```



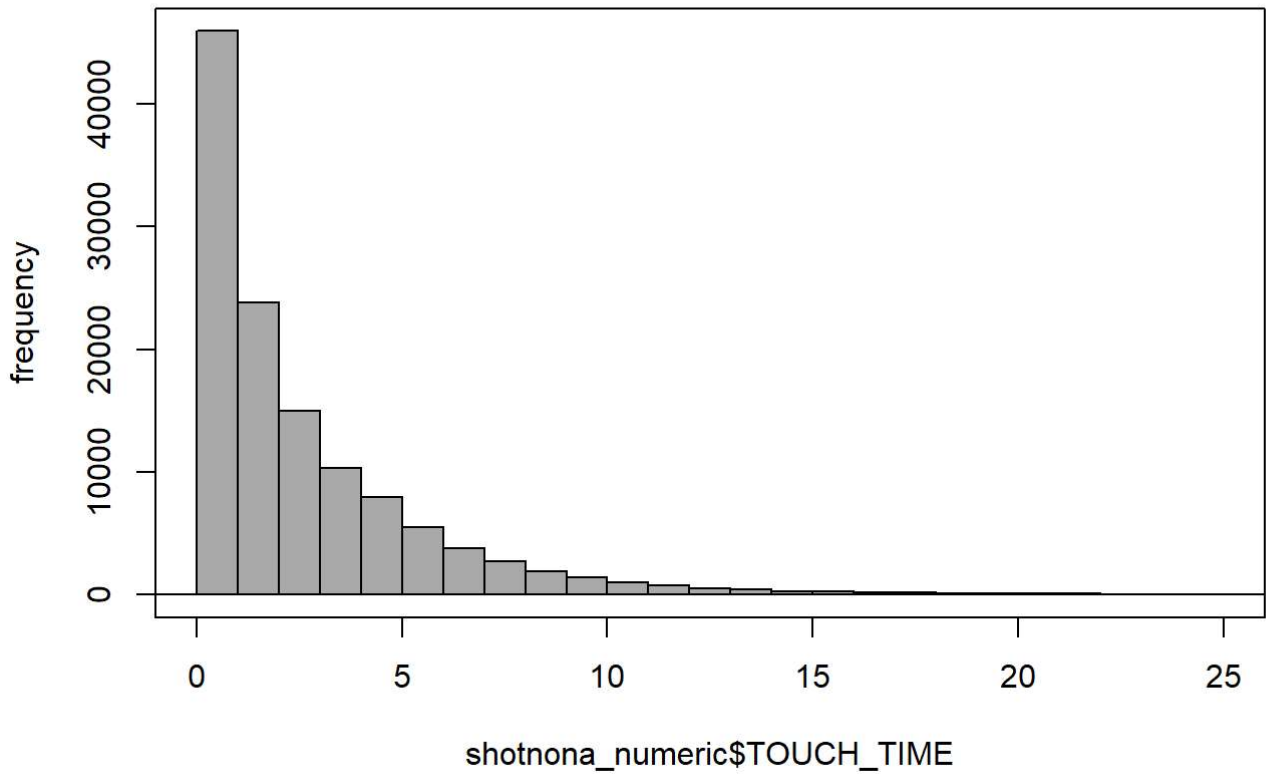

```
Hist(shotnona_numeric$SHOT_CLOCK, scale="frequency", breaks=seq(0,25,1), col="darkgray")
```



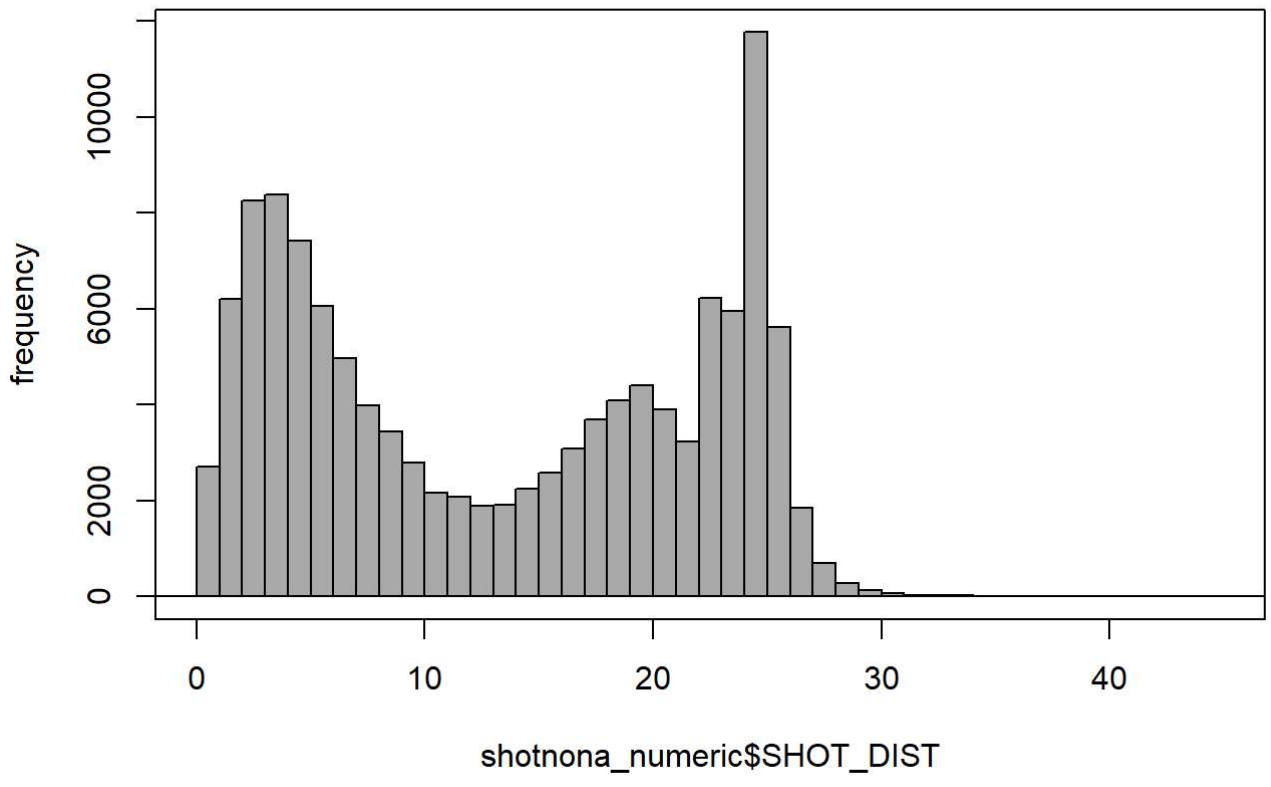
```
Hist(shotnona_numeric$DRIBBLES, scale="frequency", breaks=seq(0,35,1), col="darkgray")
```



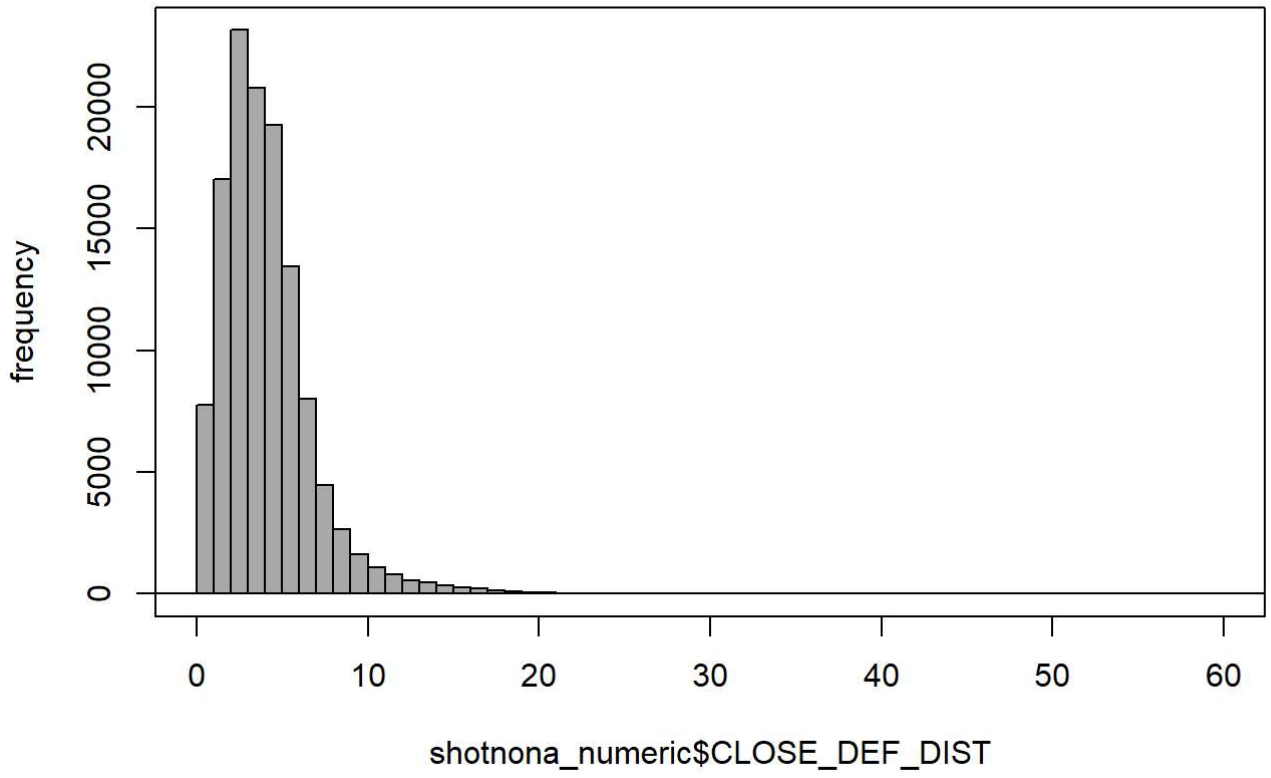

```
Hist(shotnona_numeric$TOUCH_TIME, scale="frequency", breaks=seq(0,25,1), col="darkgray")
```



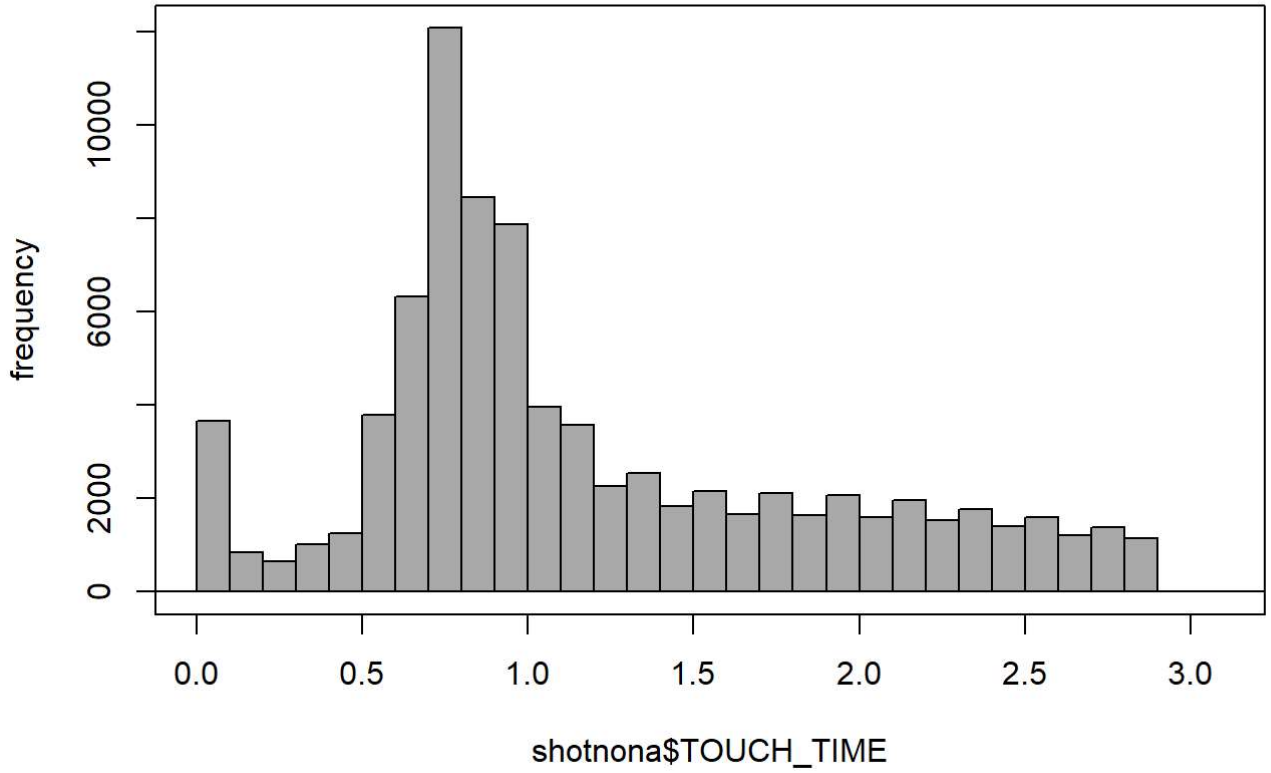

```
Hist(shotnona_numeric$SHOT_DIST, scale="frequency", breaks=seq(0,45,1), col="darkgray")
```



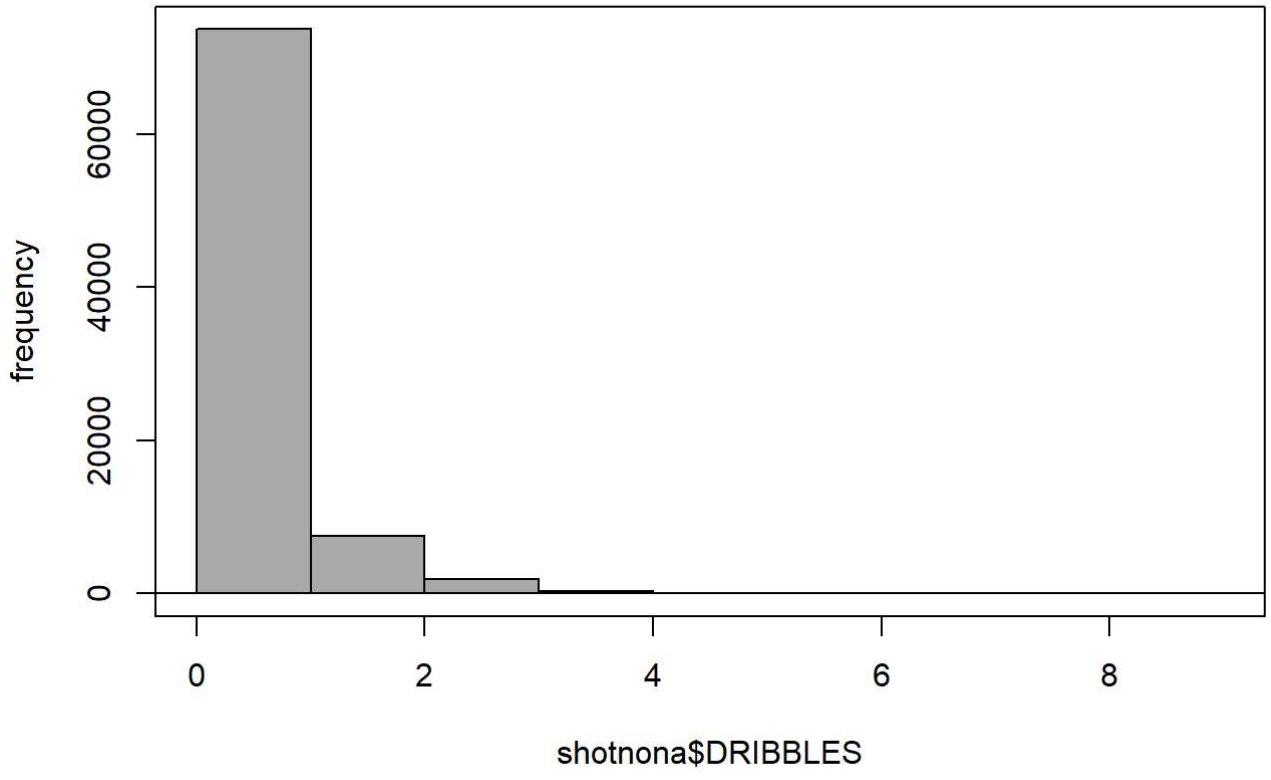
```
Hist(shotnona_numeric$CLOSE_DEF_DIST, scale="frequency", breaks=seq(0,60,1), col="darkgray")
```




```
#We found that the distributionS of "DRIBBLES","TOUCH_TIME" are highly skew.
#Process strange distribution:
#As we are more interested in short "TOUCH_TIME"---we could analyze performance in some short key mome
nts.
#"DRIBBLES" and "TOUCH_TIME" are somewhat corrlated: longer you hold the ball, more dribbles you could
 make.
#We decide to filter the data that on short TOUCH_TIME.
shotnona<-dplyr::filter(shotnona, TOUCH_TIME < 3)
#check
Hist(shotnona$TOUCH_TIME, scale="frequency", breaks=seq(0,3.1,0.1), col="darkgray")
```




```
Hist(shotnona$DRIBBLES, scale="frequency", breaks=seq(0,9,1), col="darkgray")
```

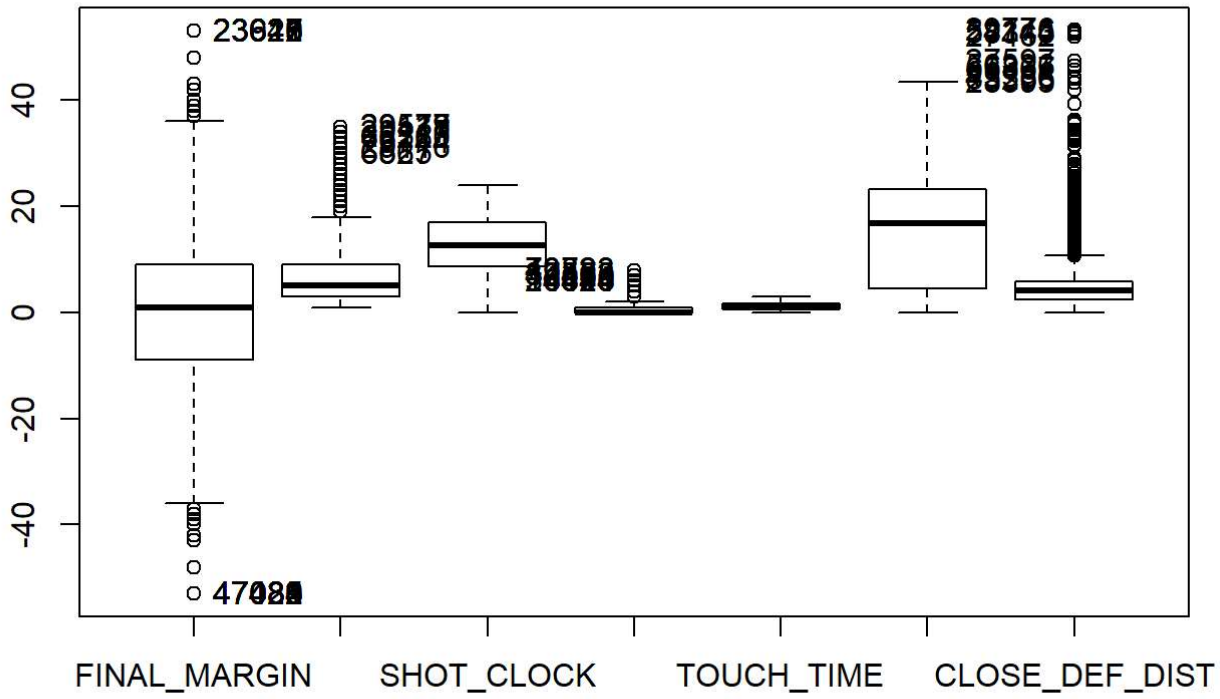


```
#explanation: The distribution of "DRIBBLES" looks still skew, but we think it's acceptable for our pr
oject.

##7.Outliers
#Diagnosis outliers of numeric variables.
shotnona_numeric<-shotnona[,c(5,6,9,10,11,12,17)]
Boxplot(shotnona_numeric,id.method="y")
```

```
##  [1] "47029" "47030" "47031" "47032" "47033" "47482" "47483" "47484"
##  [9] "47485" "47486" "23615" "23616" "23617" "23618" "23619" "23620"
## [17] "23621" "23622" "23623" "23941" "23539" "29118" "29577" "29117"
## [25] "6626"  "68510" "73744" "29116" "6567"  "6625"  "40823" "73783"
## [33] "34352" "39565" "47475" "6494"  "14170" "16848" "20820" "34611"
## [41] "29776" "54144" "58343" "27462" "27597" "56336" "68867" "45386"
## [49] "83236" "28395"
```

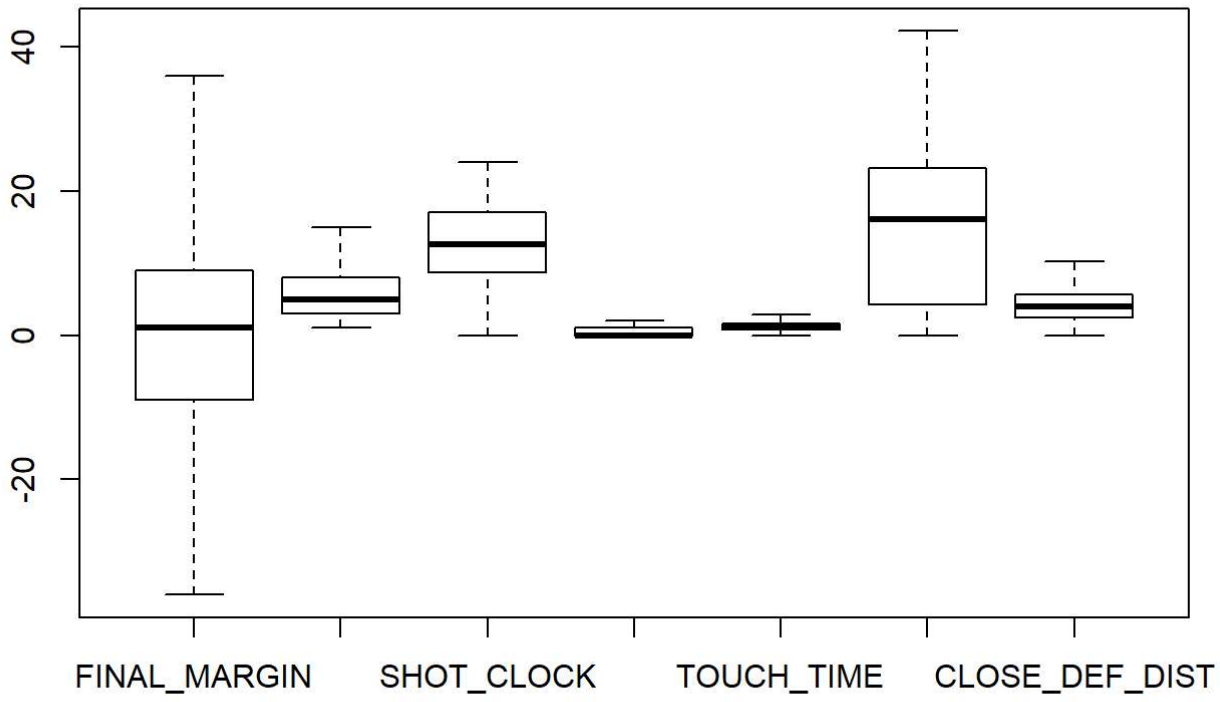
```
list<-Boxplot(shotnona_numeric,id.method="y")
```




```
outliers<-shotnona_numeric[list,]
head(outliers)
```

```
##       FINAL_MARGIN SHOT_NUMBER SHOT_CLOCK DRIBBLES TOUCH_TIME SHOT_DIST
## 47029          -53           1        8.3        0         0.8       0.4
## 47030          -53           3       22.8        0         1.1       6.2
## 47031          -53           4        9.6        0         1.5      15.1
## 47032          -53           5       23.6        0         0.6       2.6
## 47033          -53           6       24.0        0         0.0       2.9
## 47482          -53           3       17.8        0         0.9      24.5
##       CLOSE_DEF_DIST
## 47029            2.0
## 47030            3.1
## 47031            5.1
## 47032            2.6
## 47033            0.0
## 47482            5.6
```

```
#Processing outliers
#define a new function
remove_outliers <- function(x, na.rm = TRUE) {
  qnt <- quantile(x, probs=c(0.25, 0.75), na.rm = na.rm)
  H <- 1.5 * IQR(x, na.rm = na.rm)
  y <- x
  y[x < (qnt[1] - H)] <- NA
  y[x > (qnt[2] + H)] <- NA
  y
}
#iterate the function to for three times to remove all outliers.
shotnona[,c(5,6,9,10,11,12,17)]<-data.frame(sapply(shotnona[,c(5,6,9,10,11,12,17)],remove_outliers))
shotnona<-na.omit(shotnona)
shotnona[,c(5,6,9,10,11,12,17)]<-data.frame(sapply(shotnona[,c(5,6,9,10,11,12,17)],remove_outliers))
shotnona<-na.omit(shotnona)
shotnona[,c(5,6,9,10,11,12,17)]<-data.frame(sapply(shotnona[,c(5,6,9,10,11,12,17)],remove_outliers))
shotnona<-na.omit(shotnona)
#check
Boxplot(shotnona[,c(5,6,9,10,11,12,17)],id.method="y",id.n=10)
```




```
##8.Variables unrelated to our questions
#Diagnose
summary(shotnona)
```

```
##    GAME_ID            MATCHUP          LOCATION  W
##  Length:73152       Length:73152       A:36450   L:36049
##  Class :character   Class :character   H:36702   W:37103
##  Mode  :character   Mode  :character
##
##
##
##
##   FINAL_MARGIN       SHOT_NUMBER     PERIOD      GAME_CLOCK
##  Min.    :-36.0000   Min.    : 1.000   1:20342   Length:73152
##  1st Qu.: -9.0000   1st Qu.: 3.000   2:18816   Class :character
##  Median :  1.0000   Median : 5.000   3:18758   Mode  :character
##  Mean   :  0.2102   Mean   : 5.622   4:14872
##  3rd Qu.:  9.0000   3rd Qu.: 8.000   5:  301
##  Max.    : 36.0000   Max.    :15.000   6:   50
##                                       7:   13
##    SHOT_CLOCK        DRIBBLES         TOUCH_TIME       SHOT_DIST
##  Min.    : 0.00   Min.    :0.0000   Min.    :0.000   Min.    : 0.00
##  1st Qu.: 8.70   1st Qu.:0.0000   1st Qu.:0.800   1st Qu.: 4.30
##  Median :12.60   Median :0.0000   Median :1.000   Median :16.10
##  Mean   :12.87   Mean   :0.3671   Mean   :1.177   Mean   :14.03
##  3rd Qu.:17.00   3rd Qu.:1.0000   3rd Qu.:1.600   3rd Qu.:23.20
##  Max.    :24.00   Max.    :2.0000   Max.    :2.800   Max.    :42.20
##
##  PTS_TYPE  SHOT_RESULT    CLOSEST_DEFENDER   CLOSEST_DEFENDER_PLAYER_ID
##  2:57042   made  :34627   Length:73152       Length:73152
##  3:16110   missed:38525   Class :character   Class :character
##                           Mode  :character   Mode  :character
##
##
##
##
##  CLOSE_DEF_DIST  FGM            PTS         player_name
##  Min.    : 0.00   0:38525   Min.    :0.000   Length:73152
##  1st Qu.: 2.40   1:34627   1st Qu.:0.000   Class :character
##  Median : 4.00             Median :0.000   Mode  :character
##  Mean   : 4.14             Mean   :1.062
##  3rd Qu.: 5.60             3rd Qu.:2.000
##  Max.    :10.20             Max.    :3.000
##
##   player_id
##  Length:73152
##  Class :character
##  Mode  :character
##
##
##
##
```

```
#Process: Some variables are unrelated to our questions. We need to filter the dataset.
shotnona<-shotnona[,c(3,6,7,8,9,10,11,12,13,14,15,17,20)]
#check
summary(shotnona)
```

```
##  LOCATION   SHOT_NUMBER      PERIOD      GAME_CLOCK           SHOT_CLOCK
##  A:36450   Min.    : 1.000   1:20342   Length:73152       Min.    : 0.00
##  H:36702   1st Qu.: 3.000   2:18816   Class :character   1st Qu.: 8.70
##            Median : 5.000   3:18758   Mode  :character   Median :12.60
##            Mean   : 5.622   4:14872                      Mean   :12.87
##            3rd Qu.: 8.000   5:  301                      3rd Qu.:17.00
##            Max.   :15.000   6:   50                      Max.   :24.00
##                             7:   13
##     DRIBBLES         TOUCH_TIME       SHOT_DIST      PTS_TYPE  SHOT_RESULT
##  Min.   :0.0000   Min.   :0.000   Min.   : 0.00   2:57042   made  :34627
##  1st Qu.:0.0000   1st Qu.:0.800   1st Qu.: 4.30   3:16110   missed:38525
##  Median :0.0000   Median :1.000   Median :16.10
##  Mean   :0.3671   Mean   :1.177   Mean   :14.03
##  3rd Qu.:1.0000   3rd Qu.:1.600   3rd Qu.:23.20
##  Max.   :2.0000   Max.   :2.800   Max.   :42.20
##
##  CLOSEST_DEFENDER   CLOSE_DEF_DIST  player_name
##  Length:73152       Min.    : 0.00   Length:73152
##  Class :character   1st Qu.: 2.40   Class :character
##  Mode  :character   Median : 4.00   Mode  :character
##                     Mean   : 4.14
##                     3rd Qu.: 5.60
##                     Max.   :10.20
##
```

```
##9.Useful variables that are not found in the dataset
#diagnose problems:
#Our research aims to find the factors impact the shot result in a certain time period.
#But we lacked some important data such as the information of defenders.
summary(shotnona)
```


```
##  LOCATION   SHOT_NUMBER      PERIOD      GAME_CLOCK           SHOT_CLOCK
##  A:36450   Min.    : 1.000   1:20342   Length:73152       Min.    : 0.00
##  H:36702   1st Qu.: 3.000   2:18816   Class :character   1st Qu.: 8.70
##            Median : 5.000   3:18758   Mode  :character   Median :12.60
##            Mean   : 5.622   4:14872                      Mean   :12.87
##            3rd Qu.: 8.000   5:  301                      3rd Qu.:17.00
##            Max.   :15.000   6:   50                      Max.   :24.00
##                             7:   13
##     DRIBBLES         TOUCH_TIME       SHOT_DIST      PTS_TYPE  SHOT_RESULT
##  Min.   :0.0000   Min.   :0.000   Min.   : 0.00   2:57042   made  :34627
##  1st Qu.:0.0000   1st Qu.:0.800   1st Qu.: 4.30   3:16110   missed:38525
##  Median :0.0000   Median :1.000   Median :16.10
##  Mean   :0.3671   Mean   :1.177   Mean   :14.03
##  3rd Qu.:1.0000   3rd Qu.:1.600   3rd Qu.:23.20
##  Max.   :2.0000   Max.   :2.800   Max.   :42.20
##
##  CLOSEST_DEFENDER   CLOSE_DEF_DIST  player_name
##  Length:73152       Min.    : 0.00   Length:73152
##  Class :character   1st Qu.: 2.40   Class :character
##  Mode  :character   Median : 4.00   Mode  :character
##                     Mean   : 4.14
##                     3rd Qu.: 5.60
##                     Max.   :10.20
##
```

```
#Process: introducing the supplementary dataset.
##cleaning data
#overall exploration
head(players)
```

```
##            Name Games.Played  MIN  PTS FGM  FGA  FG. X3PM X3PA X3P. FTM
## 1      AJ Price           26  324  133  51  137 37.2   15   57 26.3  16
## 2  Aaron Brooks           82 1885  954 344  817 42.1  121  313 38.7 145
## 3  Aaron Gordon           47  797  243  93  208 44.7   13   48 27.1  44
## 4 Adreian Payne           32  740  213  91  220 41.4    1    9 11.1  30
## 5    Al Horford           76 2318 1156 519  965 53.8   11   36 30.6 107
## 6  Al Jefferson           65 1992 1082 486 1010 48.1    2    5 40.0 108
##   FTA  FT. OREB DREB REB AST STL BLK TOV  PF  EFF AST.TOV STL.TOV Age
## 1  24 66.7    6   26  32  46   7   0  14  15  110    3.29    0.50  29
## 2 174 83.3   32  134 166 261  54  15 157 189  791    1.66    0.34  30
## 3  61 72.1   46  123 169  33  21  22  38  83  318    0.87    0.55  20
## 4  46 65.2   48  114 162  30  19   9  44  88  244    0.68    0.43  24
## 5 141 75.9  131  413 544 244  68  98 100 121 1530    2.44    0.68  29
## 6 165 65.5   99  449 548 113  47  84  68 139 1225    1.66    0.69  30
##   Birth_Place Birthdate                  Collage Experience Height Pos
## 1          us  7-Oct-86 University of Connecticut          5  185.0  PG
## 2          us 14-Jan-85      University of Oregon          6  180.0  PG
## 3          us 16-Sep-95     University of Arizona          R  202.5  PF
## 4          us 19-Feb-91 Michigan State University          R  205.0  PF
## 5          do  3-Jun-86     University of Florida          7  205.0   C
## 6          us  4-Jan-85                                   10  205.0   C
##   Team Weight      BMI
## 1  PHO  81.45 23.79839
## 2  CHI  72.45 22.36111
## 3  ORL  99.00 24.14266
## 4  ATL 106.65 25.37775
## 5  ATL 110.25 26.23438
## 6  CHA 130.05 30.94587
```

```
str(players)
```

```
## 'data.frame':    490 obs. of  34 variables:
##  $ Name        : Factor w/ 490 levels "Aaron Brooks",..: 4 1 2 3 6 7 8 9 10 11 ...
##  $ Games.Played: int  26 82 47 32 76 65 74 27 5 69 ...
##  $ MIN         : int  324 1885 797 740 2318 1992 1744 899 14 1518 ...
##  $ PTS         : int  133 954 243 213 1156 1082 545 374 4 432 ...
##  $ FGM         : int  51 344 93 91 519 486 195 121 1 179 ...
##  $ FGA         : int  137 817 208 220 965 1010 440 300 4 353 ...
##  $ FG.         : num  37.2 42.1 44.7 41.4 53.8 48.1 44.3 40.3 25 50.7 ...
##  $ X3PM        : int  15 121 13 1 11 2 73 26 0 1 ...
##  $ X3PA        : int  57 313 48 9 36 5 210 68 0 3 ...
##  $ X3P.        : num  26.3 38.7 27.1 11.1 30.6 40 34.8 38.2 0 33.3 ...
##  $ FTM         : int  16 145 44 30 107 108 82 106 2 73 ...
##  $ FTA         : int  24 174 61 46 141 165 101 129 2 104 ...
##  $ FT.         : num  66.7 83.3 72.1 65.2 75.9 65.5 81.2 82.2 100 70.2 ...
##  $ OREB        : int  6 32 46 48 131 99 31 19 1 142 ...
##  $ DREB        : int  26 134 123 114 413 449 173 95 0 312 ...
##  $ REB         : int  32 166 169 162 544 548 204 114 1 454 ...
##  $ AST         : int  46 261 33 30 244 113 83 82 1 32 ...
##  $ STL         : int  7 54 21 19 68 47 56 17 0 34 ...
##  $ BLK         : int  0 15 22 9 98 84 5 5 0 105 ...
##  $ TOV         : int  14 157 38 44 100 68 60 52 0 74 ...
##  $ PF          : int  15 189 83 88 121 139 148 64 1 213 ...
##  $ EFF         : int  110 791 318 244 1530 1225 569 338 3 778 ...
##  $ AST.TOV     : num  3.29 1.66 0.87 0.68 2.44 1.66 1.38 1.58 0 0.43 ...
##  $ STL.TOV     : num  0.5 0.34 0.55 0.43 0.68 0.69 0.93 0.33 0 0.46 ...
##  $ Age         : int  29 30 20 24 29 30 33 24 24 22 ...
##  $ Birth_Place : Factor w/ 42 levels "","ar","au","ba",..: 40 40 40 40 13 40 40 40 40 39 ...
##  $ Birthdate   : Factor w/ 409 levels "","1-Apr-88",..: 375 70 103 134 296 321 102 161 74 96 ...
##  $ Collage     : Factor w/ 113 levels "","Arizona State University",..: 66 90 60 33 69 1 33 65 86 7
7 ...
##  $ Experience  : Factor w/ 21 levels "0","1","10","11",..: 16 17 21 21 18 3 16 14 21 2 ...
##  $ Height      : num  185 180 202 205 205 ...
##  $ Pos         : Factor w/ 5 levels "C","PF","PG",..: 3 3 2 2 1 1 5 5 1 1 ...
##  $ Team        : Factor w/ 31 levels "","ATL","BOS",..: 25 5 23 2 2 4 19 30 6 25 ...
##  $ Weight      : num  81.5 72.5 99 106.7 110.2 ...
##  $ BMI         : num  23.8 22.4 24.1 25.4 26.2 ...
```

```
#missing values
sapply(players, function(x)(sum(is.na(x))))
```

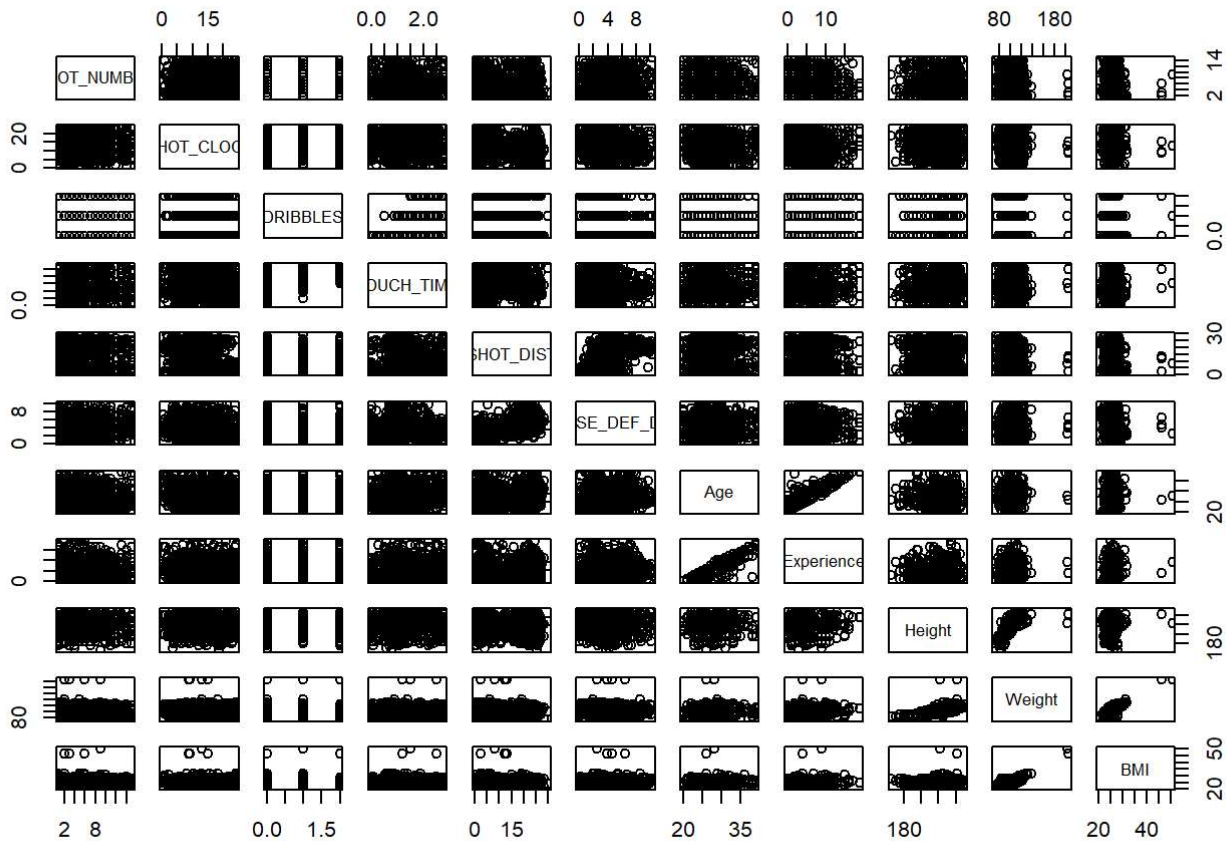
```
##          Name Games.Played           MIN           PTS           FGM
##             0             0             0             0             0
##           FGA           FG.          X3PM          X3PA          X3P.
##             0             0             0             0             0
##           FTM           FTA           FT.         OREB         DREB
##             0             0             0             0             0
##           REB           AST           STL           BLK           TOV
##             0             0             0             0             0
##            PF           EFF      AST.TOV      STL.TOV           Age
##             0             0             0             0             0
##  Birth_Place     Birthdate      Collage   Experience       Height
##             0             0             0             0             0
##           Pos          Team        Weight           BMI
##             0             0             0             0
```

```
#no missing values. we have filled the missing values manually according to external resources.
#Remove Variables that are not important for our analysis
players <- players[c(1,25,29,30,31,33,34)]
#Replace "R" to "0"  in variable "Experience"
players$Experience <- as.character(players$Experience)
players[players$Experience=="R",]$Experience="0"
players$Experience <- as.numeric(players$Experience)
#Unify the format of "Names" with dataset1
x <- str_split_fixed(players$Name, " ",2)
players$Name <- paste(x[,2],x[,1], sep=", ")

#Merge data
shotmerge<-merge(shotnona,players,by.x="CLOSEST_DEFENDER", by.y="Name")

#####III. Data exploring

#Sample visualization
set.seed(100)
sample_numeric<-shotmerge[sample(1:71934, 1000, replace = FALSE),c(3,6,7,8,9,12,14,15,16,18,19)]
pairs(sample_numeric)
```




```
scatterplotMatrix(~SHOT_NUMBER+SHOT_CLOCK+DRIBBLES+TOUCH_TIME+SHOT_DIST+CLOSE_DEF_DIST+Age+Experience+Height+Weight+BMI,
                  reg.line=lm, smooth=TRUE, spread=FALSE, span=0.5, id.n=0, diagonal = 'density', data=sample_numeric)
```

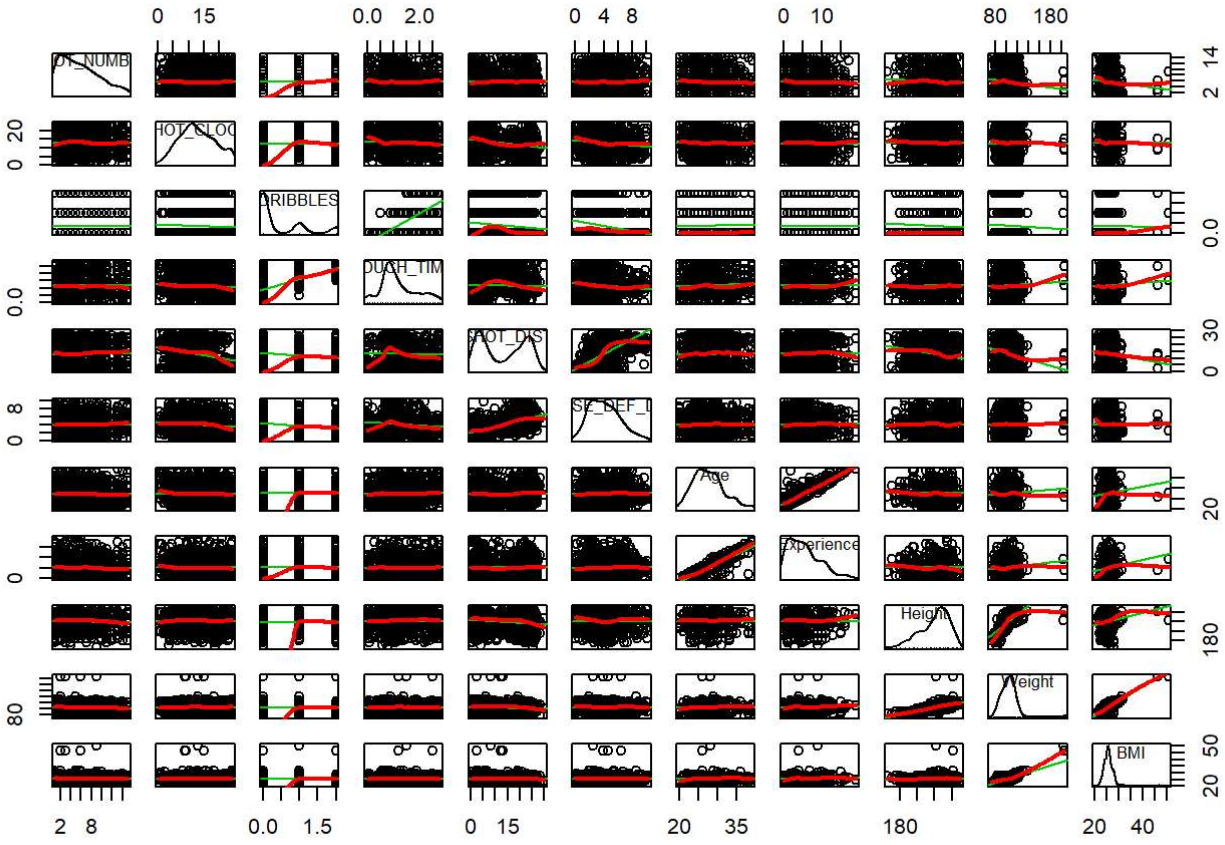
```
## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth
```




```
#####IV. Data saving
save(shotmerge, file = "shotmerge.Rdata")
```