data_prep

pmcdevitt

5 août 2017

```
setwd(home_dir)
setwd(data_dir)

shots <- read.csv("data.csv", stringsAsFactors = FALSE)
setwd(home_dir)

names(shots) <- tolower(names(shots))

for (i in 2:(length(shots)))
{
    if (class(shots[,i]) == "character")
    {
        shots[,i] <- factor (shots[,i])
    }
}
```

Some Wikipedia BBall trivia :

In the National Basketball Association (NBA), the court is 94 by 50 feet
The NBA adopted the three-point line at the start of the 1979–80 season. This is of variable distance, ranging from 22 feet (6.7 m) in the corners to 23.75 feet (7.24 m) behind the top of the key.
Kobe Bean Bryant is an American retired professional basketball player and businessman. He played his entire 20-year career with the Los Angeles Lakers of the National Basketball Association (NBA). He entered the NBA directly from high school and won five NBA championships with the Lakers.
Playing career 1996–2016
Career statistics
Points 33,643 (25.0 ppg)
Rebounds 7,047 (5.2 rpg)
Assists 6,306 (4.7 apg)

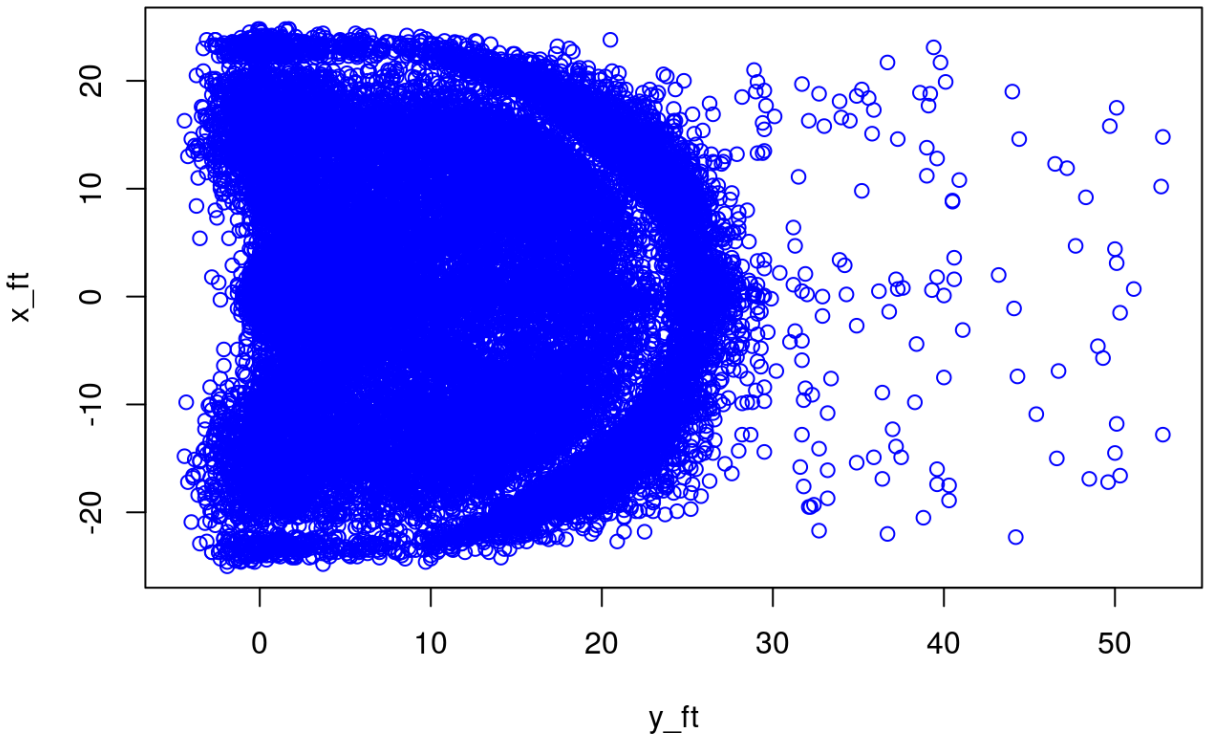
```
# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
=-=-
# ...   remove outliers ... more than 5 sigma from mean value
# ...   -=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
=-=-

    lst <- length(shots)    #

    for (i in 1 : lst)
    {
        if(class(shots[,i]) == "integer" || class(shots[,i]) == "numeric")
        {
            shots[,i][which(scale(shots[,i]) > 5)] <- NA
            shots[,i][which(scale(shots[,i]) < -5)] <- NA
        }
    }
```

```
    summary_tbl <- data.frame(x = character(0), stats = character(0))

    for (i in 2 : length(shots))
    {
        if(class(shots[,i]) == "integer" || class(shots[,i]) == "numeric")
        {
            new_row <- data.frame(x = names(shots[i]),
                                  stats = sprintf (
                                    "| %8d | %8d | %8.1f | %8.1f | %8.1f | %
8.1f | %8.3f | ",
                                    colSums(!is.na(shots[i])),
                                    (dim(shots)[1] - colSums(!is.na(shots[i])
)),
                                    mean(shots[,i], na.rm = TRUE),
                                    median(shots[,i], na.rm = TRUE),
                                    max(shots[,i], na.rm = TRUE),
                                    min(shots[,i], na.rm = TRUE),
                                    skewness(shots[,i], na.rm = TRUE)
                                    )
                                  )
            summary_tbl <- rbind(summary_tbl, new_row)
        }
    }

    summary_tbl
```

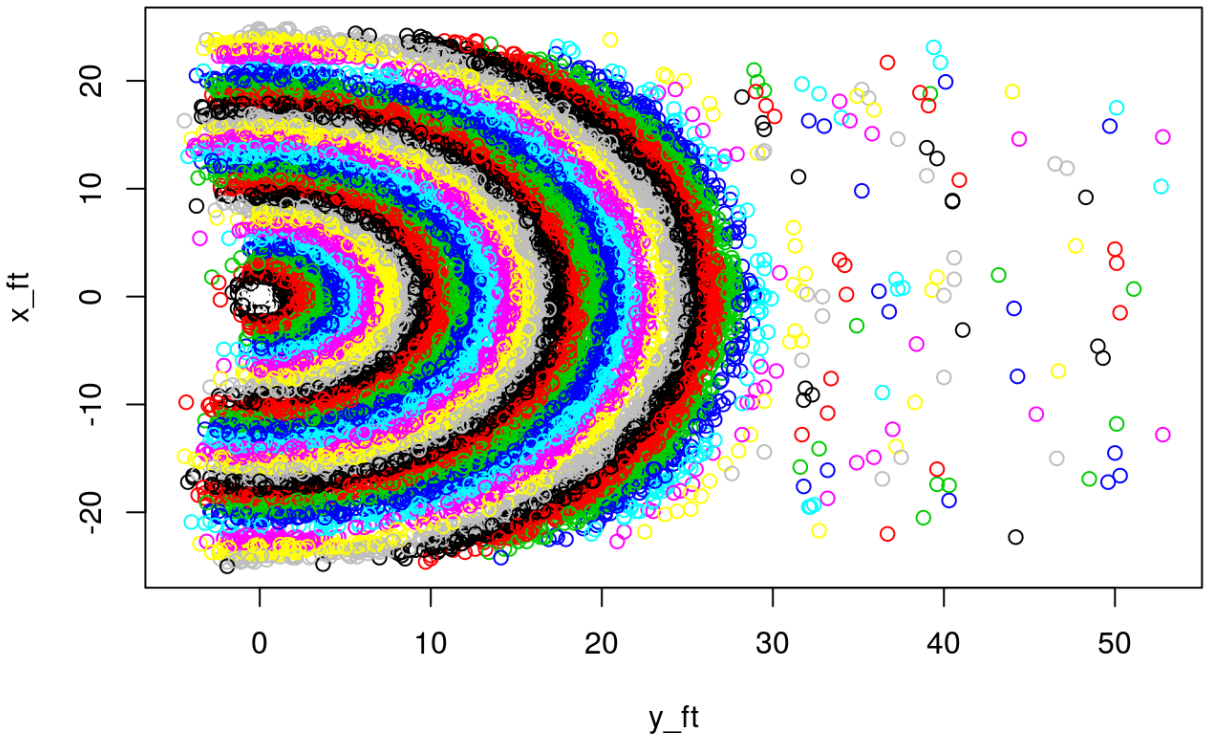
```
##                    x
## 1      game_event_id
## 2            game_id
## 3                lat
## 4              loc_x
## 5              loc_y
## 6                lon
## 7  minutes_remaining
## 8             period
## 9           playoffs
## 10 seconds_remaining
## 11    shot_distance
## 12   shot_made_flag
## 13           team_id
## 14           shot_id
##
stats
## 1                  |    30697 |        0 |    249.2 |    253.0 |    659.0
|      2.0 |    0.065 |
## 2          |    30697 |        0 | 24764065.9 | 20900354.0 | 49900088.0 |
20000012.0 |    1.705 |
## 3                  |    30655 |       42 |     34.0 |     34.0 |     34.1
|     33.5 |   -0.559 |
## 4                  |    30697 |        0 |      7.1 |      0.0 |    248.0
|   -250.0 |   -0.085 |
## 5                  |    30655 |       42 |     90.4 |     74.0 |    528.0
|    -44.0 |    0.559 |
## 6                  |    30697 |        0 |   -118.3 |   -118.3 |   -118.0
|   -118.5 |   -0.085 |
## 7                  |    30697 |        0 |      4.9 |      5.0 |     11.0
|      0.0 |    0.199 |
## 8                  |    30697 |        0 |      2.5 |      3.0 |      7.0
|      1.0 |    0.055 |
## 9                  |    30697 |        0 |      0.1 |      0.0 |      1.0
|      0.0 |    1.999 |
## 10                 |    30697 |        0 |     28.4 |     28.0 |     59.0
|      0.0 |    0.031 |
## 11                 |    30673 |       24 |     13.4 |     15.0 |     60.0
|      0.0 |   -0.036 |
## 12                 |    25697 |     5000 |      0.4 |      0.0 |      1.0
|      0.0 |    0.217 |
## 13 |    30697 |        0 | 1610612747.0 | 1610612747.0 | 1610612747.0 | 16
10612747.0 |      NaN |
## 14                 |    30697 |        0 |  15349.0 |  15349.0 |  30697.0
|      1.0 |    0.000 |
```

```
# ...   seconds_remaining is seconds remaining in the current minute
# ...   minutes_remaining is minutes remaining in the period
# ...   there are 4 (regular time) periods in a match, each of 12 minutes
# ...   create new vairable of time remaining in the match
# ...       - creates negative values for overtime periods

    shots$time_remaining <- shots$seconds_remaining +
                            shots$minutes_remaining * 60 +
                            (4 - shots$period) * 12 * 60

# ...   home / away games can be determined from 'matchup' field
# ...       @ - designates away game
# ...       vs. - designates home game

    shots$home_away <- "home"
    away_lst <- grep("@", shots$matchup, perl=TRUE, value=FALSE)
    shots$home_away[away_lst] <- "away"

# ...   shot distances appear to be in feet * 10 ... just convert to feet

    shots$x_ft <- shots$loc_x / 10
    shots$y_ft <- shots$loc_y / 10

# ...   add polar coordinate - from basket to shot point

    shots$rad <- shots$shot_distance
    shots$ang <- atan2(shots$x_ft, shots$y_ft) * 180/pi

# ...   calculate total points scored in this data set

    shots$pts_scored <- 2
    three_pt_lst <- grep("3PT", shots$shot_type, perl=TRUE, value=FALSE)
    shots$pts_scored[three_pt_lst] <- 3
    shots$pts_scored <- shots$pts_scored * shots$shot_made_flag

    total_pts_scored <- sum(shots$pts_scored, na.rm = TRUE)
```

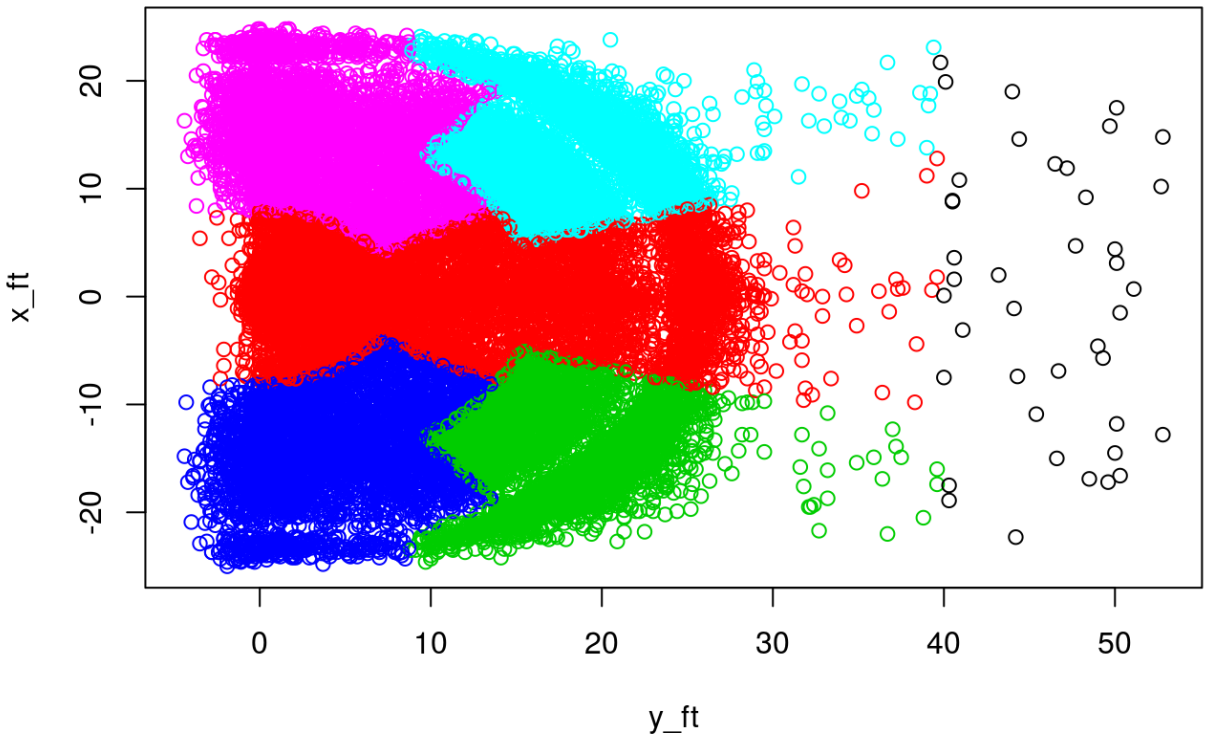
```
    plot(x_ft ~ y_ft, shots, col = "blue")
```



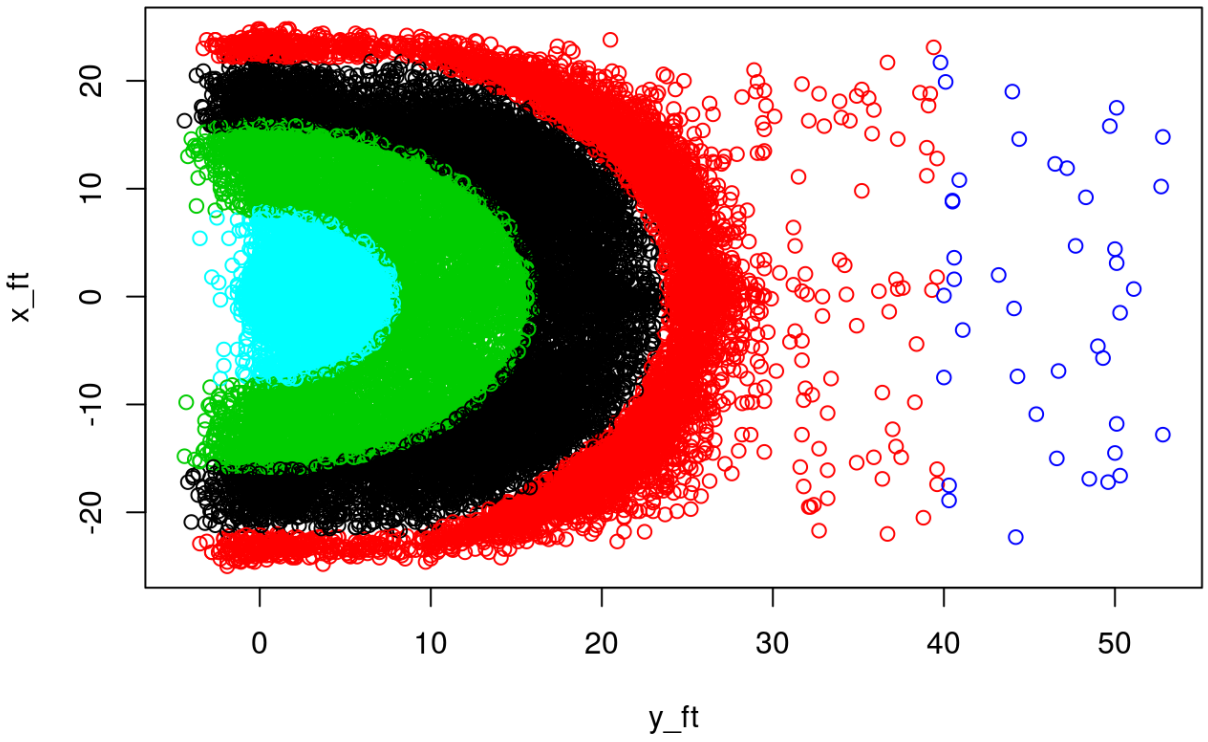

```
plot(x_ft ~ y_ft, shots, col = shot_distance)
```



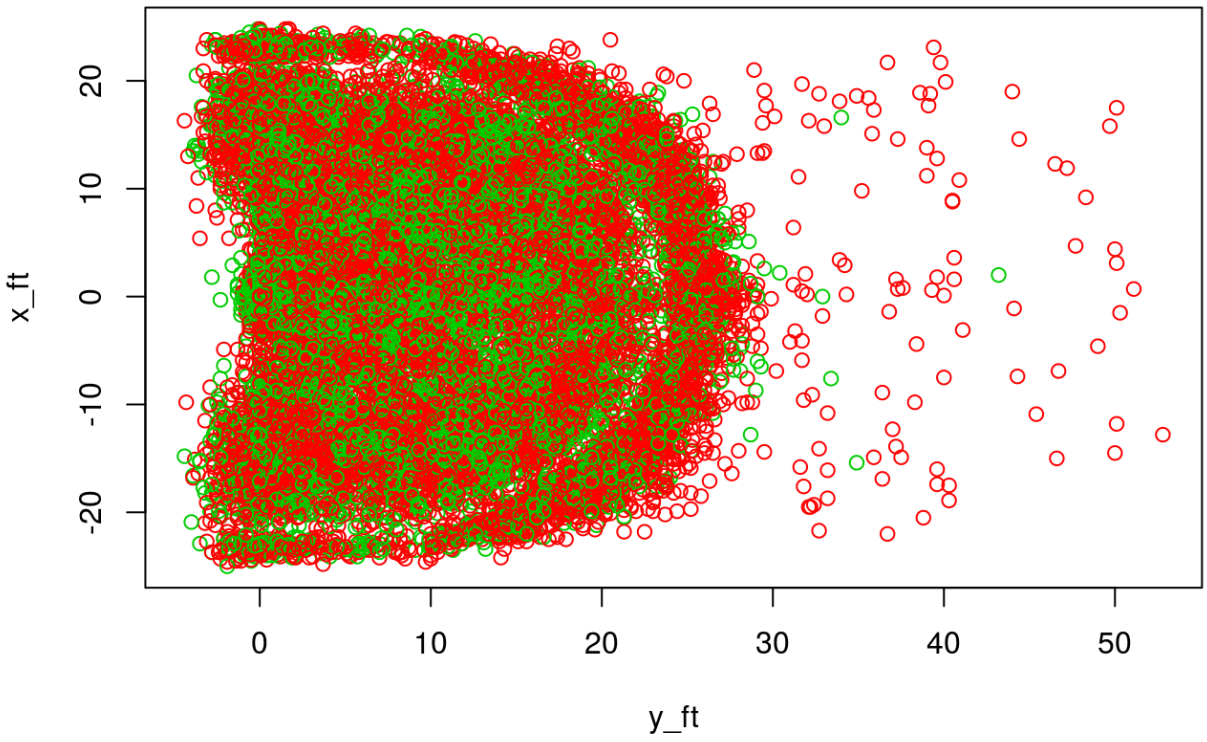

```
plot(x_ft ~ y_ft, shots, col = shot_zone_area)
```



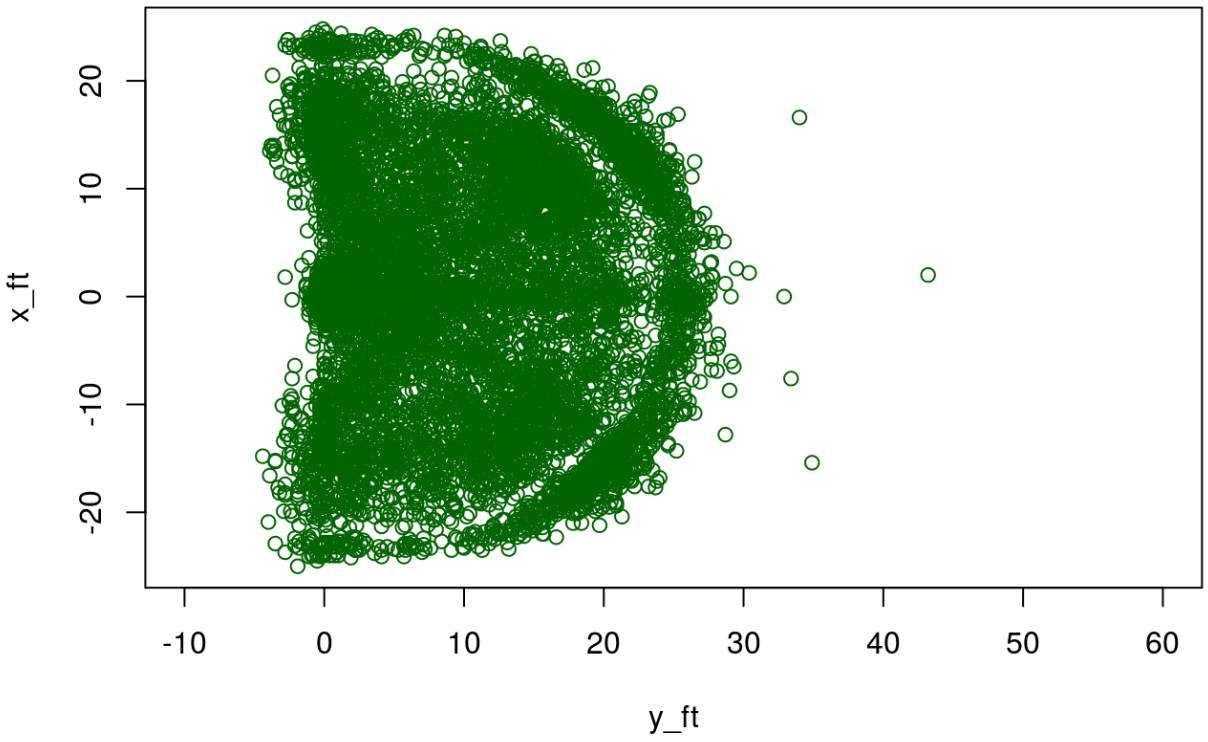

```
plot(x_ft ~ y_ft, shots, col = shot_zone_range)
```



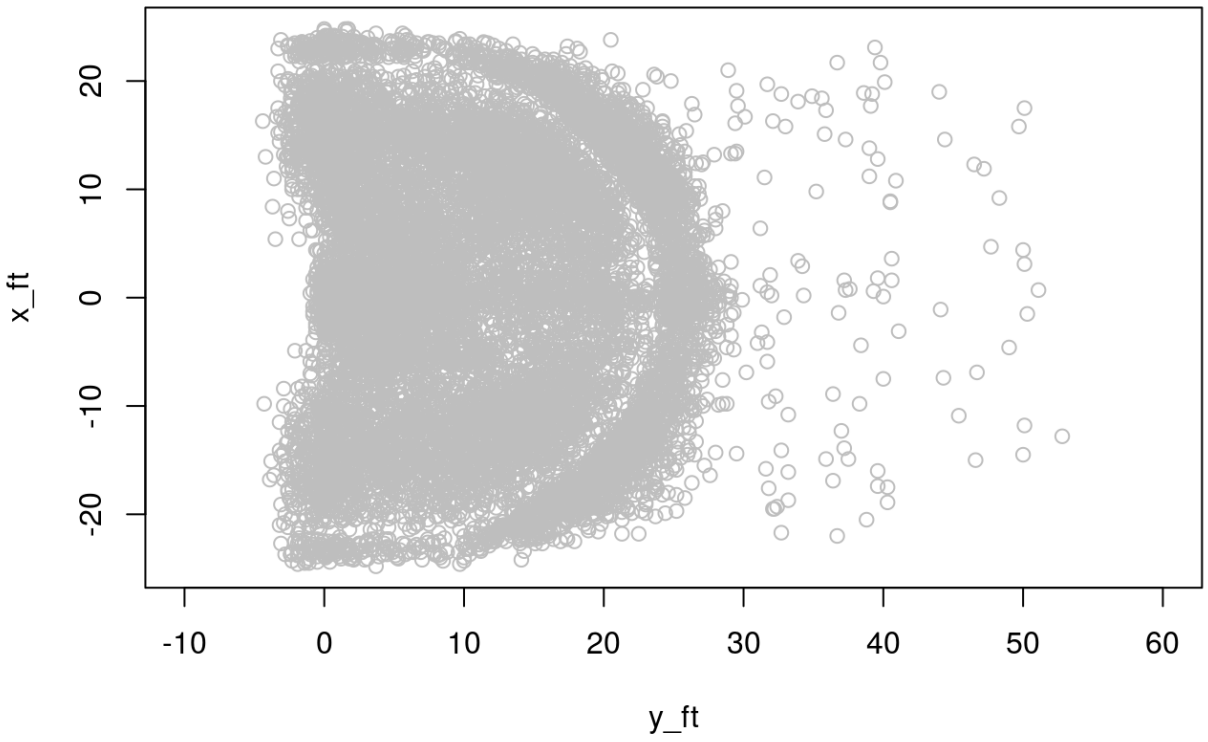

```
plot(x_ft ~ y_ft, shots, col = shot_made_flag+2)
```



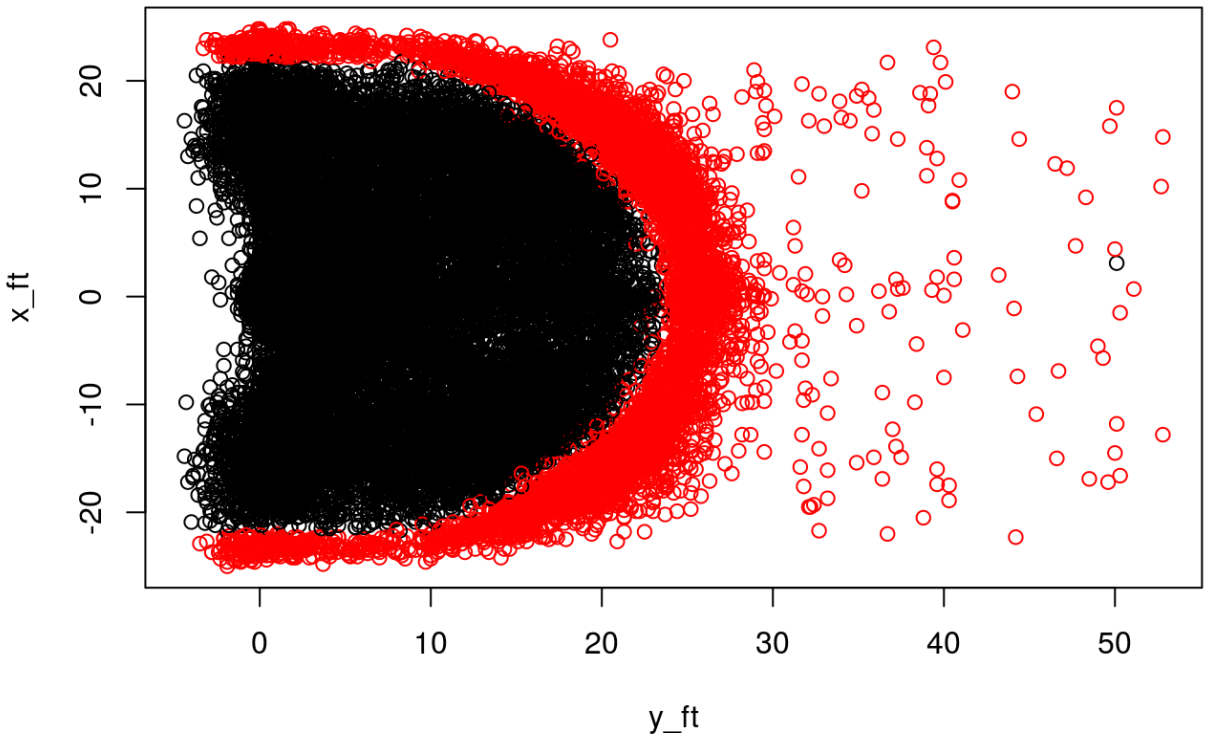

```
plot(x_ft ~ y_ft, data = shots[shots$shot_made_flag == 1,],
        col = "darkgreen",
        xlim = c(-10, 60))
```



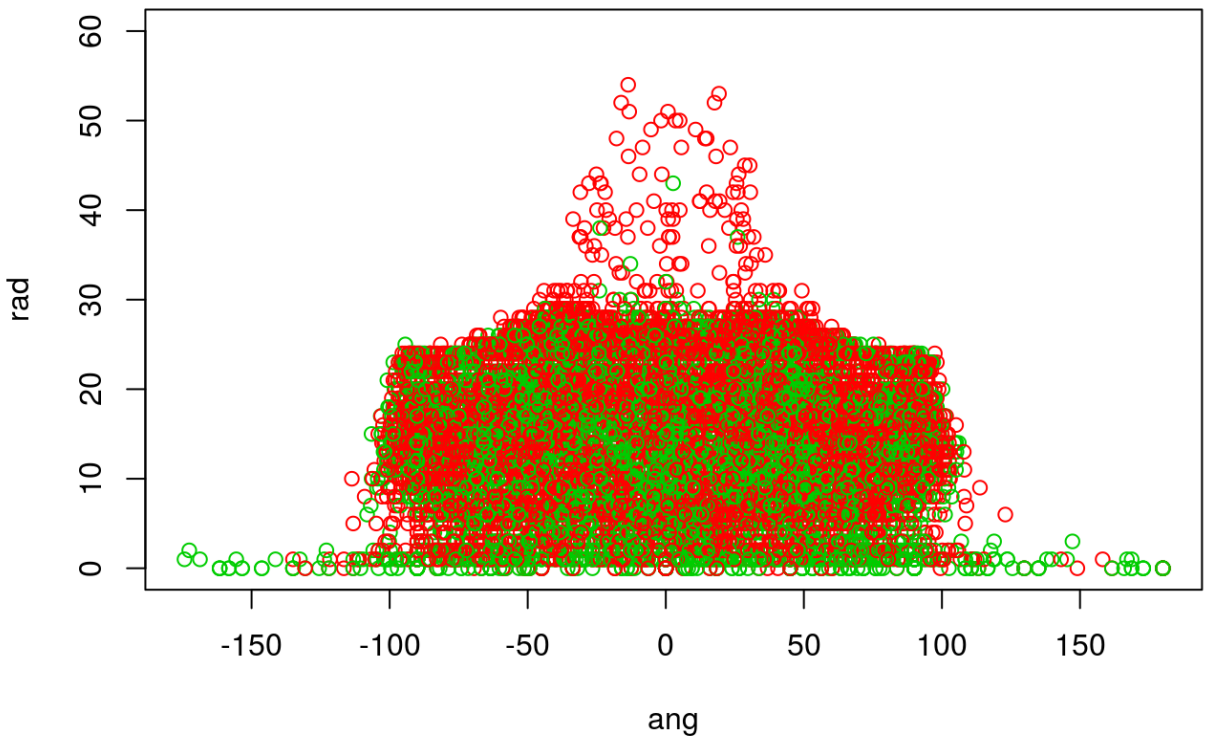

```
plot(x_ft ~ y_ft, data = shots[shots$shot_made_flag == 0,],
        col = "grey",
     xlim = c(-10, 60))
```



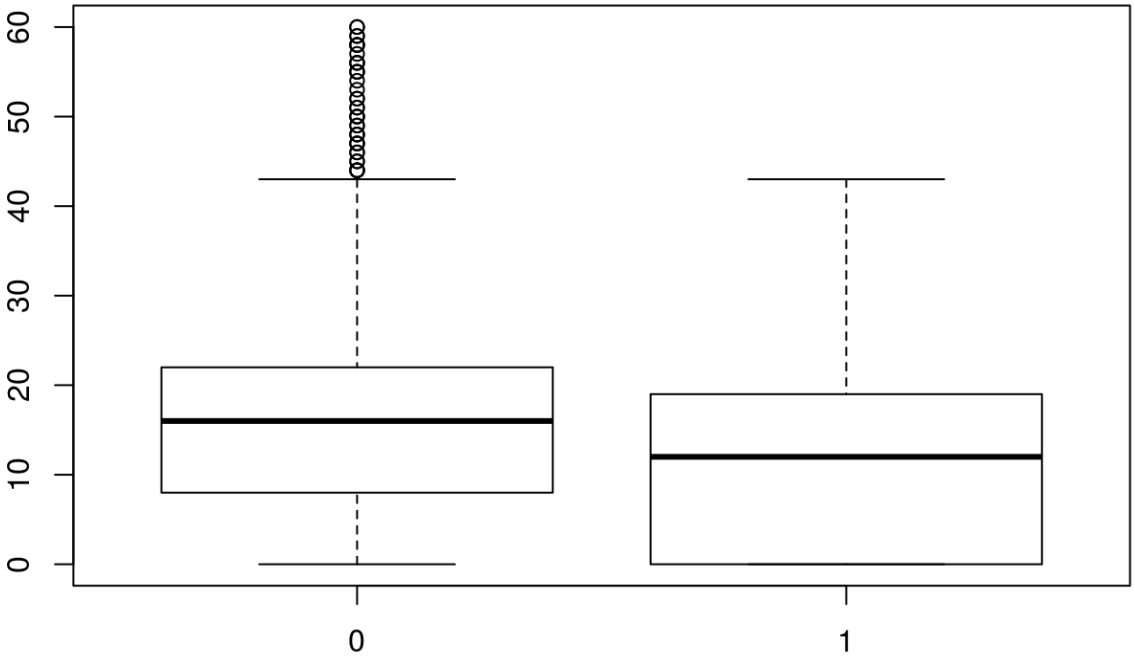

```
plot(x_ft ~ y_ft, shots, col = shot_type)
```



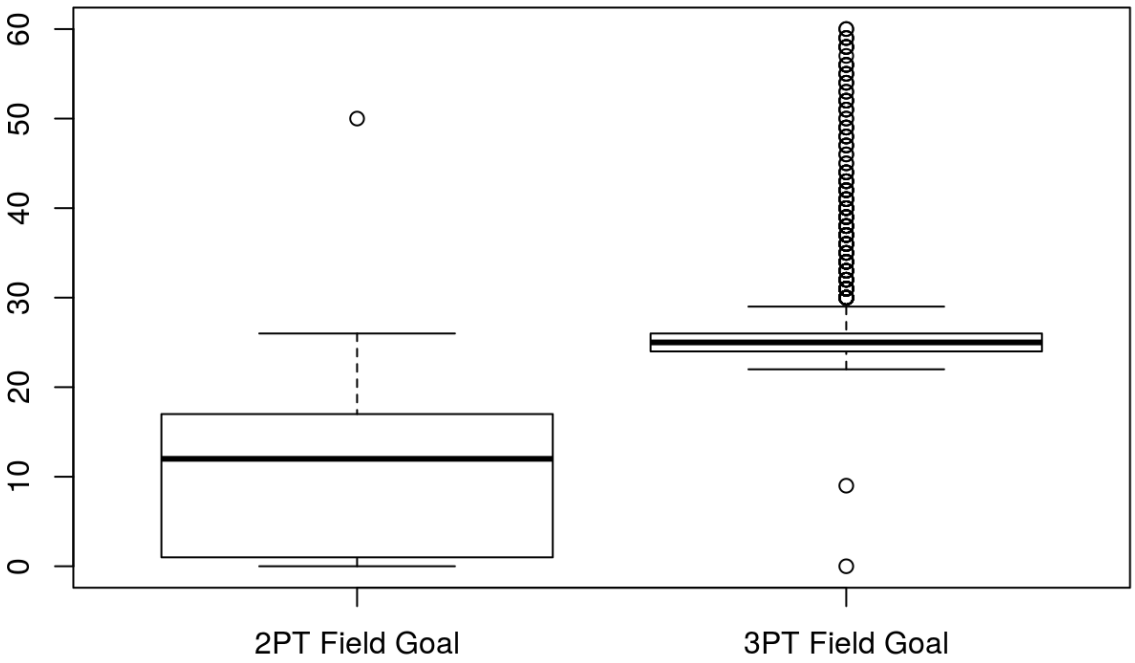

```
plot(rad ~ ang, data = shots, col = shot_made_flag+2)
```




```
boxplot(shot_distance ~ shot_made_flag, data = shots)
```




```
boxplot(shot_distance ~ shot_type, data = shots)
```


0
10
20
30
40
50
60
2PT Field Goal
3PT Field Goal