MATH2349: Data Wrangling Assignment 2

Rafeed Sultaan(s3763175)

13/10/2020

1. Executive Summary:

The data analysis process includes many steps from defining the problem statement to data pre-processing which includes getting the data ready to analyze so that we can explore the data to gain meaningful insights from it. This report provides an analysis on a NBA game data set for the season of 2019-2020 season following the 3 steps of Data Analysis: Defining a problem statement, Data Collecting and Data pre processing.

The data collection for this analysis is retrieved from Kaggle.com and R is used to carry out the data pre processing. Our data-preprocessing methodology is highlighted below:

1. The two separate data sets called “games.csv” and “games_details.csv” is imported using the ‘readr’ library .
2. The redundant columns, for example Comment,s were dropped from the data set and the untidy dataset were made tidy using the Tidy Principles.
3. The two data sets were joined together to make one data set using the primary keys GAME_ID and HOME_TEAM_ID.
4. For analysis a new variable was created called OREB_To_DREB_RATIO, which would help in the further analysis because it indicates the offensive contribution in terms of the defensive contribution of a NBA player.
5. After the above steps were implemented the missing values were analyzed in the data set and were imputed or removed accordingly.
6. Next, outliers were found out for the numeric columns and “capping” technique was applied to replace the outliers.
7. Finally, Box Cox transformations were applied to ensure the normality of the data, which makes it ready for the analysis.

From the analysis of the data set it can be concluded that the nba_games_dataset is the final cleaned data set which is free from any impurities and can be used for further analysis.

2. Required packages [R code]:

```
#Loading the required packages
library(readr)
library(dplyr)
library(tidyr)
library(car)
library(outliers)
library(stringr)
library(forecast)
library(magrittr)
```

3. Data:

The “games.csv” contains information about NBA Games during 2019 Season. It contains 23195 observations and 21 variables. It was collected from Kaggle[https://www.kaggle.com/nathanlauga/nba-games?select=games.csv (https://www.kaggle.com/nathanlauga/nba-games?select=games.csv)].

The “games.csv” contains the following attributes:

- GAME_DATE_EST(Data Type- Date) - This attribute contains The date when the match occured.
- GAMEE_ID (Data Type - double) - This attribute contains the game status. It is the primary key
- HOME_TEAM_ID (Data Type - double) - This attribute contains the unique id of the home team.
- VISITOR_TEAM_ID(Data Type - double) -This attribute contains the unique id of the visitor team.
- SEASON(Data Type - double) - It contains the NBA season information of the current match.
- TEAM_ID_home (Data Type - double) - It contains the unique id of the home team.
- PTS_home (Data Type- double)- This attribute indicates the the total points scored by the home team.
- FG_PCT_home (Data Type-double)- This attribute indicates the the assists by the home team.
- FT_PCT_home (Data Type-double) - This attribute indicates the free throw percentage by the home team
- FG3_PCT_home (Data Type-double) - This attribute indicates the 3 point field goal percentage by the home team
- AST_home (Data Type-double) - This attribute indicates the assists made by the home team.
- REB_home (Data Type-double) - This attribute indicates the rebounds by the home team.
- TEAM_ID_away (Data Type-double) - This attribute contains the team id of the away team.
- PTS_away (Data Type-double) - This attribute indicates the total points scored by the away team
- FG_PCT_away (Data Type-double) - This attribute indicates the Field Goal Percentage by the away team
- FT_PCT_away (Data Type-double) - This attribute indicates the free throw percentage by the away team
- FG3_PCT_away (Data Type-double) - This attribute indicates the 3 point field goal percentage by the away team
- AST_away (Data Type-double) - This attribute indicates the the assists by the away team

- REB_away (Data Type-double) - This attribute indicates the rebounds by the away team.
- HOME_TEAM_WINS (Data Type-double) - This attribute shows if the home team wins or not. 1 indicates win and 0 indicates loss

Loading the “games.csv” data set. viewing the contents using the head(). Checking the attributes and their data types using the str function

```
#Loading the data set "games.csv"
#Stripping the white spaces where it is possible
games <- read.csv("data/games.csv", strip.white = TRUE)
#Viewing the games dataset
head(games)
```

```
##   GAME_DATE_EST  GAME_ID GAME_STATUS_TEXT HOME_TEAM_ID VISITOR_TEAM_ID SEASON
## 1    2020-03-01 21900895            Final   1610612766      1610612749   2019
## 2    2020-03-01 21900896            Final   1610612750      1610612742   2019
## 3    2020-03-01 21900897            Final   1610612746      1610612755   2019
## 4    2020-03-01 21900898            Final   1610612743      1610612761   2019
## 5    2020-03-01 21900899            Final   1610612758      1610612765   2019
## 6    2020-03-01 21900900            Final   1610612740      1610612747   2019
##   TEAM_ID_home PTS_home FG_PCT_home FT_PCT_home FG3_PCT_home AST_home REB_home
## 1   1610612766       85       0.354       0.900        0.229       22       47
## 2   1610612750       91       0.364       0.400        0.310       19       57
## 3   1610612746      136       0.592       0.805        0.542       25       37
## 4   1610612743      133       0.566       0.700        0.500       38       41
## 5   1610612758      106       0.407       0.885        0.257       18       51
## 6   1610612740      114       0.421       0.818        0.219       24       52
##   TEAM_ID_away PTS_away FG_PCT_away FT_PCT_away FG3_PCT_away AST_away REB_away
## 1   1610612749       93       0.402       0.762        0.226       20       61
## 2   1610612742      111       0.468       0.632        0.275       28       56
## 3   1610612755      130       0.505       0.650        0.488       27       37
## 4   1610612761      118       0.461       0.897        0.263       24       36
## 5   1610612765      100       0.413       0.667        0.429       23       42
## 6   1610612747      122       0.515       0.900        0.371       23       36
##   HOME_TEAM_WINS
## 1              0
## 2              0
## 3              1
## 4              1
## 5              1
## 6              0
```

```
#Viewing the data types of  attributes in the "games.csv" dataset
str(games)
```

```
## 'data.frame':    23195 obs. of  21 variables:
##  $ GAME_DATE_EST   : chr  "2020-03-01" "2020-03-01" "2020-03-01" "2020-03-01" ...
##  $ GAME_ID         : int  21900895 21900896 21900897 21900898 21900899 21900900 21900901 21900887 21900888 21900889 ...
##  $ GAME_STATUS_TEXT: chr  "Final" "Final" "Final" "Final" ...
##  $ HOME_TEAM_ID    : int  1610612766 1610612750 1610612746 1610612743 1610612758 1610612740 1610612744 1610612752 1610612
737 1610612748 ...
##  $ VISITOR_TEAM_ID : int  1610612749 1610612742 1610612755 1610612761 1610612765 1610612747 1610612764 1610612741 1610612
757 1610612751 ...
##  $ SEASON          : int  2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 ...
##  $ TEAM_ID_home    : int  1610612766 1610612750 1610612746 1610612743 1610612758 1610612740 1610612744 1610612752 1610612
737 1610612748 ...
##  $ PTS_home        : num  85 91 136 133 106 114 110 125 129 116 ...
##  $ FG_PCT_home     : num  0.354 0.364 0.592 0.566 0.407 0.421 0.472 0.553 0.548 0.451 ...
##  $ FT_PCT_home     : num  0.9 0.4 0.805 0.7 0.885 0.818 0.708 0.697 0.864 0.833 ...
##  $ FG3_PCT_home    : num  0.229 0.31 0.542 0.5 0.257 0.219 0.321 0.4 0.429 0.368 ...
##  $ AST_home        : num  22 19 25 38 18 24 25 29 34 27 ...
##  $ REB_home        : num  47 57 37 41 51 52 52 50 36 45 ...
##  $ TEAM_ID_away    : int  1610612749 1610612742 1610612755 1610612761 1610612765 1610612747 1610612764 1610612741 1610612
757 1610612751 ...
##  $ PTS_away        : num  93 111 130 118 100 122 124 115 117 113 ...
##  $ FG_PCT_away     : num  0.402 0.468 0.505 0.461 0.413 0.515 0.488 0.461 0.5 0.465 ...
##  $ FT_PCT_away     : num  0.762 0.632 0.65 0.897 0.667 0.9 0.889 0.696 0.714 0.739 ...
##  $ FG3_PCT_away    : num  0.226 0.275 0.488 0.263 0.429 0.371 0.667 0.486 0.286 0.364 ...
##  $ AST_away        : num  20 28 27 24 23 23 24 26 14 30 ...
##  $ REB_away        : num  61 56 37 36 42 36 34 33 42 44 ...
##  $ HOME_TEAM_WINS  : int  0 0 1 1 1 0 0 1 1 1 ...
```

```
#Checking the dimension of the "games.csv" dataset
dim(games)
```

```
## [1] 23195    21
```

The “games_details.csv” contains extensive game details information about NBA Games during 2019-2020 Season. It contains 576782 observations and observations and 28 variables.It was collected from Kaggle[https://www.kaggle.com/nathanlauga/nba-games?select=games_details.csv (https://www.kaggle.com/nathanlauga/nba-games?select=games_details.csv)].

The “games_details.csv” contains the following attributes:

- GAME_ID (Data Type- double) - The primary key of dataset. This indicates the game id of the match.

- TEAM_ID (Data Type- double) - This attribute indicates the team id of the player.
- TEAM_ABBREVIATION(Data Type- character) - This attribute indicates the abbreviation of the team.
- TEAM_CITY- This attributes indicates the city of the team.
- PLAYER_ID - This attribute indicates the player id.
- PLAYER_NAME (Data Type - Character)- This attribute indicates the players name.
- START_POSITION(Data Type - Character) - This attribute indicates the starting position of the player.
- COMMENT(Data Type-character) - This attribute is unnecessary. It contains the comments about the player.
- MINS(DATA TYPE - S3:hms)- This attribute contains the total minutes player. But this attribute is untidy because hours and minutes are 2 different attributes.
- FGM (Data Type- double)- This attribute indicates the field goals made by the player.
- FGA (Data Type- double) - This attribute indicates the field goal attempted by the player.
- FG_PCT (Data Type- double) - This attribute indicates the field goal percentage of the player.
- FG3M (Data Type- double)- This attribute indicates the 3 point field goal made by the player.
- FG3A (Data Type- double) - This attribute indicates the 3 point field goal attempted by the player.
- FG3_PCT (Data Type- double)- This attribute indicates the 3 point field goal percentage of the player
- FTM (Data Type- double) - This attribute indicates the free throw made by the player.
- FTA (Data Type- double)- This attribute indicates the free throw attempted by the player.
- FT_PCT (Data Type- double) - This attribute indicates the field goal percentage of the player.
- OREB (Data Type- double) - This attribute indicates the offensive rebounds made by the player.
- DREB (Data Type- double) - This attribute indicates the defensive rebounds made by the player.
- REB (Data Type- double) - This attribute indicates the total rebounds made by the player.
- AST (Data Type- double) - This attribute indicates the total steals made by the player.
- STL (Data Type- double) - This attribute indicates the total steal made by the player.
- BLK (Data Type- double)- This attribute indicates the total number of blocks made by the player.
- TO (Data Type- double)- This attribute indicates the total turnovers made by the player.
- PF - (Data Type- double) - This attribute indicates the total personal fouls made by the player.
- PTS (Data Type- double) - This attribute indicates the total points made by the player.
- PLUS_MINUS (Data Type- double) - This attribute indicates the plus/minus attribute to quantify the team's performance when the player was on the team.

```
#Loading the "games_details.csv" data set
games_details <- read.csv("data/games_details.csv", strip.white = TRUE)
head(games_details)
```

```
##    GAME_ID    TEAM_ID TEAM_ABBREVIATION TEAM_CITY PLAYER_ID
## 1 21900895 1610612749               MIL Milwaukee    202083
## 2 21900895 1610612749               MIL Milwaukee    203507
## 3 21900895 1610612749               MIL Milwaukee    201572
## 4 21900895 1610612749               MIL Milwaukee   1628978
## 5 21900895 1610612749               MIL Milwaukee    202339
## 6 21900895 1610612749               MIL Milwaukee   1626192
##             PLAYER_NAME START_POSITION COMMENT   MIN FGM FGA FG_PCT FG3M FG3A
## 1       Wesley Matthews              F         27:08   3  11  0.273    2    7
## 2 Giannis Antetokounmpo              F         34:55  17  28  0.607    1    4
## 3           Brook Lopez              C         26:25   4  11  0.364    1    5
## 4      Donte DiVincenzo              G         27:35   1   5  0.200    0    3
## 5          Eric Bledsoe              G         22:17   2   8  0.250    0    1
## 6       Pat Connaughton                        24:52   2   5  0.400    1    4
##   FG3_PCT FTM FTA FT_PCT OREB DREB REB AST STL BLK TO PF PTS PLUS_MINUS
## 1   0.286   0   0  0.000    4    4   8   2   2   0  0  0   8         11
## 2   0.250   6   7  0.857    2   18  20   6   1   0  3  2  41         22
## 3   0.200   7   9  0.778    2    5   7   0   0   3  0  2  16         16
## 4   0.000   0   0  0.000    1    6   7   5   0   1  2  0   2         14
## 5   0.000   0   0  0.000    1    0   1   2   1   0  3  2   4          6
## 6   0.250   1   2  0.500    2    3   5   1   0   0  1  2   6          0
```

```
#Viewing the attributes in the "games_details.csv" data set
str(games_details)
```

```
## 'data.frame':    576782 obs. of  28 variables:
##  $ GAME_ID          : int  21900895 21900895 21900895 21900895 21900895 21900895 21900895 21900895 21900895 21900895 ...
##  $ TEAM_ID          : int  1610612749 1610612749 1610612749 1610612749 1610612749 1610612749 1610612749 1610612749 161061
2749 1610612749 ...
##  $ TEAM_ABBREVIATION: chr  "MIL" "MIL" "MIL" "MIL" ...
##  $ TEAM_CITY        : chr  "Milwaukee" "Milwaukee" "Milwaukee" "Milwaukee" ...
##  $ PLAYER_ID        : int  202083 203507 201572 1628978 202339 1626192 201577 1628425 101107 201588 ...
##  $ PLAYER_NAME      : chr  "Wesley Matthews" "Giannis Antetokounmpo" "Brook Lopez" "Donte DiVincenzo" ...
##  $ START_POSITION   : chr  "F" "F" "C" "G" ...
##  $ COMMENT          : chr  "" "" "" "" ...
##  $ MIN              : chr  "27:08" "34:55" "26:25" "27:35" ...
##  $ FGM              : num  3 17 4 1 2 2 1 1 0 4 ...
##  $ FGA              : num  11 28 11 5 8 5 5 2 1 11 ...
##  $ FG_PCT           : num  0.273 0.607 0.364 0.2 0.25 0.4 0.2 0.5 0 0.364 ...
##  $ FG3M             : num  2 1 1 0 0 1 0 1 0 1 ...
##  $ FG3A             : num  7 4 5 3 1 4 0 2 1 4 ...
##  $ FG3_PCT          : num  0.286 0.25 0.2 0 0 0.25 0 0.5 0 0.25 ...
##  $ FTM              : num  0 6 7 0 0 1 0 0 0 2 ...
##  $ FTA              : num  0 7 9 0 0 2 0 0 0 3 ...
##  $ FT_PCT           : num  0 0.857 0.778 0 0 0.5 0 0 0 0.667 ...
##  $ OREB             : num  4 2 2 1 1 2 1 0 0 2 ...
##  $ DREB             : num  4 18 5 6 0 3 2 3 2 3 ...
##  $ REB              : num  8 20 7 7 1 5 3 3 2 5 ...
##  $ AST              : num  2 6 0 5 2 1 0 0 2 2 ...
##  $ STL              : num  2 1 0 0 1 0 0 0 1 2 ...
##  $ BLK              : num  0 0 3 1 0 0 1 0 1 0 ...
##  $ TO               : num  0 3 0 2 3 1 2 1 1 3 ...
##  $ PF               : num  0 2 2 0 2 2 1 0 1 1 ...
##  $ PTS              : num  8 41 16 2 4 6 2 3 0 11 ...
##  $ PLUS_MINUS       : num  11 22 16 14 6 0 -12 -8 -11 2 ...
```

```
#Checking the dimension of the "games_details.csv" data set
dim(games_details)
```

```
## [1] 576782     28
```

We are renaming column TEAM_ID to HOME_TEAM_ID in “games_details.csv” data set , so that it becomes parts of the primary key, before we perform the join operation.

```
colnames(games_details)[2] <- "HOME_TEAM_ID"
head(games_details)
```

```
##    GAME_ID HOME_TEAM_ID TEAM_ABBREVIATION TEAM_CITY PLAYER_ID
## 1 21900895   1610612749               MIL Milwaukee    202083
## 2 21900895   1610612749               MIL Milwaukee    203507
## 3 21900895   1610612749               MIL Milwaukee    201572
## 4 21900895   1610612749               MIL Milwaukee   1628978
## 5 21900895   1610612749               MIL Milwaukee    202339
## 6 21900895   1610612749               MIL Milwaukee   1626192
##             PLAYER_NAME START_POSITION COMMENT   MIN FGM FGA FG_PCT FG3M FG3A
## 1       Wesley Matthews              F         27:08   3  11  0.273    2    7
## 2 Giannis Antetokounmpo              F         34:55  17  28  0.607    1    4
## 3           Brook Lopez              C         26:25   4  11  0.364    1    5
## 4      Donte DiVincenzo              G         27:35   1   5  0.200    0    3
## 5          Eric Bledsoe              G         22:17   2   8  0.250    0    1
## 6       Pat Connaughton                        24:52   2   5  0.400    1    4
##   FG3_PCT FTM FTA FT_PCT OREB DREB REB AST STL BLK TO PF PTS PLUS_MINUS
## 1   0.286   0   0  0.000    4    4   8   2   2   0  0  0   8         11
## 2   0.250   6   7  0.857    2   18  20   6   1   0  3  2  41         22
## 3   0.200   7   9  0.778    2    5   7   0   0   3  0  2  16         16
## 4   0.000   0   0  0.000    1    6   7   5   0   1  2  0   2         14
## 5   0.000   0   0  0.000    1    0   1   2   1   0  3  2   4          6
## 6   0.250   1   2  0.500    2    3   5   1   0   0  1  2   6          0
```

Performing the join operation operation of the two data sets: “games.csv” dataset and “games_details.csv”. The merged data set contains 288643 observations and 47 attributes found using dim() function before any kind of data pre-processing.

```
#Performing the Join operation of the data section
nba_games_dataset <- games_details %>% inner_join(games, by = c('GAME_ID','HOME_TEAM_ID'))
head(nba_games_dataset)
```

```
##    GAME_ID HOME_TEAM_ID TEAM_ABBREVIATION TEAM_CITY PLAYER_ID     PLAYER_NAME
## 1 21900895   1610612766               CHA Charlotte   1628970   Miles Bridges
## 2 21900895   1610612766               CHA Charlotte   1629023 P.J. Washington
## 3 21900895   1610612766               CHA Charlotte    202687 Bismack Biyombo
## 4 21900895   1610612766               CHA Charlotte   1628984 Devonte' Graham
## 5 21900895   1610612766               CHA Charlotte   1626179    Terry Rozier
## 6 21900895   1610612766               CHA Charlotte   1628998     Cody Martin
##   START_POSITION COMMENT   MIN FGM FGA FG_PCT FG3M FG3A FG3_PCT FTM FTA FT_PCT
## 1              F         35:15   3  13  0.231    1    7   0.143   0   0      0
## 2              F         31:52   5  14  0.357    1    8   0.125   1   1      1
## 3              C         22:07   2   8  0.250    0    0   0.000   4   4      1
## 4              G         32:21   7  18  0.389    3    8   0.375   0   1      0
## 5              G         36:05   6  18  0.333    0    3   0.000   1   1      1
## 6                        29:08   4   8  0.500    2    5   0.400   1   1      1
##   OREB DREB REB AST STL BLK TO PF PTS PLUS_MINUS GAME_DATE_EST GAME_STATUS_TEXT
## 1    1    3   4   2   2   2  2  2   7         -4    2020-03-01            Final
## 2    1    5   6   3   0   2  2  3  12        -13    2020-03-01            Final
## 3    4    5   9   2   0   2  1  3   8        -15    2020-03-01            Final
## 4    0    2   2   3   1   0  0  2  17        -14    2020-03-01            Final
## 5    2    1   3   4   1   0  2  2  13        -20    2020-03-01            Final
## 6    0    5   5   2   0   1  1  2  11          2    2020-03-01            Final
##   VISITOR_TEAM_ID SEASON TEAM_ID_home PTS_home FG_PCT_home FT_PCT_home
## 1      1610612749   2019   1610612766       85       0.354         0.9
## 2      1610612749   2019   1610612766       85       0.354         0.9
## 3      1610612749   2019   1610612766       85       0.354         0.9
## 4      1610612749   2019   1610612766       85       0.354         0.9
## 5      1610612749   2019   1610612766       85       0.354         0.9
## 6      1610612749   2019   1610612766       85       0.354         0.9
##   FG3_PCT_home AST_home REB_home TEAM_ID_away PTS_away FG_PCT_away FT_PCT_away
## 1        0.229       22       47   1610612749       93       0.402       0.762
## 2        0.229       22       47   1610612749       93       0.402       0.762
## 3        0.229       22       47   1610612749       93       0.402       0.762
## 4        0.229       22       47   1610612749       93       0.402       0.762
## 5        0.229       22       47   1610612749       93       0.402       0.762
## 6        0.229       22       47   1610612749       93       0.402       0.762
##   FG3_PCT_away AST_away REB_away HOME_TEAM_WINS
## 1        0.226       20       61              0
## 2        0.226       20       61              0
## 3        0.226       20       61              0
## 4        0.226       20       61              0
## 5        0.226       20       61              0
## 6        0.226       20       61              0
```

```
#Checking the dimensions of the merged dataset
dim(nba_games_dataset)
```

```
## [1] 288643     47
```

4. Tidy & Manipulate Data I:

The “games.csv” data set is not tidy because it contains 2 attributes “AWAY_TEAM_ID” and “TEAM_ID_away” which are duplicates of each other.Therefore, we are dropping “Team_ID_away” from the data set.

According to the Tidy Principles:Each variable forms a column, whereas variables here are forming multiple columns. Therefore, I am dropping “Team_ID_home” and “Team_ID_away” attributes from the “nba_games_dataset” dataset.

```
#Dropping TEAM_ID_HOME and TEAM_ID_AWAY
nba_games_dataset <- nba_games_dataset %>% subset( select= -c(TEAM_ID_home,TEAM_ID_away) )
head(nba_games_dataset)
```

```
##    GAME_ID HOME_TEAM_ID TEAM_ABBREVIATION TEAM_CITY PLAYER_ID     PLAYER_NAME
## 1 21900895   1610612766               CHA Charlotte   1628970   Miles Bridges
## 2 21900895   1610612766               CHA Charlotte   1629023 P.J. Washington
## 3 21900895   1610612766               CHA Charlotte    202687 Bismack Biyombo
## 4 21900895   1610612766               CHA Charlotte   1628984 Devonte' Graham
## 5 21900895   1610612766               CHA Charlotte   1626179    Terry Rozier
## 6 21900895   1610612766               CHA Charlotte   1628998     Cody Martin
##   START_POSITION COMMENT   MIN FGM FGA FG_PCT FG3M FG3A FG3_PCT FTM FTA FT_PCT
## 1              F         35:15   3  13  0.231    1    7   0.143   0   0      0
## 2              F         31:52   5  14  0.357    1    8   0.125   1   1      1
## 3              C         22:07   2   8  0.250    0    0   0.000   4   4      1
## 4              G         32:21   7  18  0.389    3    8   0.375   0   1      0
## 5              G         36:05   6  18  0.333    0    3   0.000   1   1      1
## 6                        29:08   4   8  0.500    2    5   0.400   1   1      1
##   OREB DREB REB AST STL BLK TO PF PTS PLUS_MINUS GAME_DATE_EST GAME_STATUS_TEXT
## 1    1    3   4   2   2   2  2  2   7         -4    2020-03-01            Final
## 2    1    5   6   3   0   2  2  3  12        -13    2020-03-01            Final
## 3    4    5   9   2   0   2  1  3   8        -15    2020-03-01            Final
## 4    0    2   2   3   1   0  0  2  17        -14    2020-03-01            Final
## 5    2    1   3   4   1   0  2  2  13        -20    2020-03-01            Final
## 6    0    5   5   2   0   1  1  2  11          2    2020-03-01            Final
##   VISITOR_TEAM_ID SEASON PTS_home FG_PCT_home FT_PCT_home FG3_PCT_home AST_home
## 1      1610612749   2019       85       0.354         0.9        0.229       22
## 2      1610612749   2019       85       0.354         0.9        0.229       22
## 3      1610612749   2019       85       0.354         0.9        0.229       22
## 4      1610612749   2019       85       0.354         0.9        0.229       22
## 5      1610612749   2019       85       0.354         0.9        0.229       22
## 6      1610612749   2019       85       0.354         0.9        0.229       22
##   REB_home PTS_away FG_PCT_away FT_PCT_away FG3_PCT_away AST_away REB_away
## 1       47       93       0.402       0.762        0.226       20       61
## 2       47       93       0.402       0.762        0.226       20       61
## 3       47       93       0.402       0.762        0.226       20       61
## 4       47       93       0.402       0.762        0.226       20       61
## 5       47       93       0.402       0.762        0.226       20       61
## 6       47       93       0.402       0.762        0.226       20       61
##   HOME_TEAM_WINS
## 1              0
## 2              0
## 3              0
## 4              0
## 5              0
## 6              0
```

The “games_details.csv” data set is not tidy because “COMMENT” is not really a variable. According to Tidy Principle,

1. Each variable should form a column, whereas comment is not a variable
2. Comment is absent in almost all the rows, for it to be considered as part of a table.

To tidy this problem, we drop “COMMENT” column from the joined “nba_games_dataset” dataset.

```
#Dropping the comment column
nba_games_dataset <- nba_games_dataset %>% subset( select= -c(COMMENT) )
head(nba_games_dataset)
```

```
##    GAME_ID HOME_TEAM_ID TEAM_ABBREVIATION TEAM_CITY PLAYER_ID     PLAYER_NAME
## 1 21900895   1610612766               CHA Charlotte   1628970   Miles Bridges
## 2 21900895   1610612766               CHA Charlotte   1629023 P.J. Washington
## 3 21900895   1610612766               CHA Charlotte    202687 Bismack Biyombo
## 4 21900895   1610612766               CHA Charlotte   1628984 Devonte' Graham
## 5 21900895   1610612766               CHA Charlotte   1626179    Terry Rozier
## 6 21900895   1610612766               CHA Charlotte   1628998     Cody Martin
##   START_POSITION   MIN FGM FGA FG_PCT FG3M FG3A FG3_PCT FTM FTA FT_PCT OREB
## 1              F 35:15   3  13  0.231    1    7   0.143   0   0      0    1
## 2              F 31:52   5  14  0.357    1    8   0.125   1   1      1    1
## 3              C 22:07   2   8  0.250    0    0   0.000   4   4      1    4
## 4              G 32:21   7  18  0.389    3    8   0.375   0   1      0    0
## 5              G 36:05   6  18  0.333    0    3   0.000   1   1      1    2
## 6                29:08   4   8  0.500    2    5   0.400   1   1      1    0
##   DREB REB AST STL BLK TO PF PTS PLUS_MINUS GAME_DATE_EST GAME_STATUS_TEXT
## 1    3   4   2   2   2  2  2   7         -4    2020-03-01            Final
## 2    5   6   3   0   2  2  3  12        -13    2020-03-01            Final
## 3    5   9   2   0   2  1  3   8        -15    2020-03-01            Final
## 4    2   2   3   1   0  0  2  17        -14    2020-03-01            Final
## 5    1   3   4   1   0  2  2  13        -20    2020-03-01            Final
## 6    5   5   2   0   1  1  2  11          2    2020-03-01            Final
##   VISITOR_TEAM_ID SEASON PTS_home FG_PCT_home FT_PCT_home FG3_PCT_home AST_home
## 1      1610612749   2019       85       0.354         0.9        0.229       22
## 2      1610612749   2019       85       0.354         0.9        0.229       22
## 3      1610612749   2019       85       0.354         0.9        0.229       22
## 4      1610612749   2019       85       0.354         0.9        0.229       22
## 5      1610612749   2019       85       0.354         0.9        0.229       22
## 6      1610612749   2019       85       0.354         0.9        0.229       22
##   REB_home PTS_away FG_PCT_away FT_PCT_away FG3_PCT_away AST_away REB_away
## 1       47       93       0.402       0.762        0.226       20       61
## 2       47       93       0.402       0.762        0.226       20       61
## 3       47       93       0.402       0.762        0.226       20       61
## 4       47       93       0.402       0.762        0.226       20       61
## 5       47       93       0.402       0.762        0.226       20       61
## 6       47       93       0.402       0.762        0.226       20       61
##   HOME_TEAM_WINS
## 1              0
## 2              0
## 3              0
## 4              0
## 5              0
## 6              0
```

“MIN” is the total time played by a player. Where the minutes are stored as hours and minutes are concatenated by “:” According to Tidy Principle, 1. Each variable should form a column, whereas MIN contains multiples variables stored in one column.

To tidy this problem, we divide “MIN” column into “HOURS” and “MINUTES” COLUMNS using “:” as the separator.

```
#Separating MIN into HOURS and MINUTES TO MAKE the merged dataset tidy
nba_games_dataset <- nba_games_dataset %>% separate(MIN, into = c("HOURS", "MINUTES"), sep = ":")
head(nba_games_dataset)
```

```
##    GAME_ID HOME_TEAM_ID TEAM_ABBREVIATION TEAM_CITY PLAYER_ID     PLAYER_NAME
## 1 21900895   1610612766               CHA Charlotte   1628970   Miles Bridges
## 2 21900895   1610612766               CHA Charlotte   1629023 P.J. Washington
## 3 21900895   1610612766               CHA Charlotte    202687 Bismack Biyombo
## 4 21900895   1610612766               CHA Charlotte   1628984 Devonte' Graham
## 5 21900895   1610612766               CHA Charlotte   1626179    Terry Rozier
## 6 21900895   1610612766               CHA Charlotte   1628998     Cody Martin
##   START_POSITION HOURS MINUTES FGM FGA FG_PCT FG3M FG3A FG3_PCT FTM FTA FT_PCT
## 1              F    35      15   3  13  0.231    1    7   0.143   0   0      0
## 2              F    31      52   5  14  0.357    1    8   0.125   1   1      1
## 3              C    22      07   2   8  0.250    0    0   0.000   4   4      1
## 4              G    32      21   7  18  0.389    3    8   0.375   0   1      0
## 5              G    36      05   6  18  0.333    0    3   0.000   1   1      1
## 6                   29      08   4   8  0.500    2    5   0.400   1   1      1
##   OREB DREB REB AST STL BLK TO PF PTS PLUS_MINUS GAME_DATE_EST GAME_STATUS_TEXT
## 1    1    3   4   2   2   2  2  2   7         -4    2020-03-01            Final
## 2    1    5   6   3   0   2  2  3  12        -13    2020-03-01            Final
## 3    4    5   9   2   0   2  1  3   8        -15    2020-03-01            Final
## 4    0    2   2   3   1   0  0  2  17        -14    2020-03-01            Final
## 5    2    1   3   4   1   0  2  2  13        -20    2020-03-01            Final
## 6    0    5   5   2   0   1  1  2  11          2    2020-03-01            Final
##   VISITOR_TEAM_ID SEASON PTS_home FG_PCT_home FT_PCT_home FG3_PCT_home AST_home
## 1      1610612749   2019       85       0.354         0.9        0.229       22
## 2      1610612749   2019       85       0.354         0.9        0.229       22
## 3      1610612749   2019       85       0.354         0.9        0.229       22
## 4      1610612749   2019       85       0.354         0.9        0.229       22
## 5      1610612749   2019       85       0.354         0.9        0.229       22
## 6      1610612749   2019       85       0.354         0.9        0.229       22
##   REB_home PTS_away FG_PCT_away FT_PCT_away FG3_PCT_away AST_away REB_away
## 1       47       93       0.402       0.762        0.226       20       61
## 2       47       93       0.402       0.762        0.226       20       61
## 3       47       93       0.402       0.762        0.226       20       61
## 4       47       93       0.402       0.762        0.226       20       61
## 5       47       93       0.402       0.762        0.226       20       61
## 6       47       93       0.402       0.762        0.226       20       61
##   HOME_TEAM_WINS
## 1              0
## 2              0
## 3              0
## 4              0
## 5              0
## 6              0
```

For tidying the character variables, The strip.white = True was used while reading the data set which removes any white space in the data set but the data set contains newline character in the characters. This ensures that there are no white spaces in the character columns.

```
nba_games_dataset$TEAM_ABBREVIATION <- str_replace_all(nba_games_dataset$TEAM_ABBREVIATION, "[\r\n]" , "")
nba_games_dataset$TEAM_CITY <- str_replace_all(nba_games_dataset$TEAM_CITY, "[\r\n]" , "")
nba_games_dataset$PLAYER_NAME <- str_replace_all(nba_games_dataset$PLAYER_NAME, "[\r\n]" , "")
nba_games_dataset$GAME_DATE_EST <- str_replace_all(nba_games_dataset$GAME_DATE_EST, "[\r\n]" , "")
```

5. Tidy & Manipulate Data II:

Mutating a new variable called “OREB_To_DREB_RATIO” that measures offensive rebounds as a comparison to defensive rebounds. The value “0.00001” is added in the denominator so that ratio doesn’t turn to the infinity value.

```
#This disables the scientific notation
options(scipen = 999)
#Mutating the variable "OREB_To_DREB_RATIO"
nba_games_dataset <-  nba_games_dataset %>% mutate(OREB_To_DREB_RATIO = round((OREB/(DREB+0.00001)),2))
head(nba_games_dataset)
```

```
##    GAME_ID HOME_TEAM_ID TEAM_ABBREVIATION TEAM_CITY PLAYER_ID     PLAYER_NAME
## 1 21900895   1610612766               CHA Charlotte   1628970   Miles Bridges
## 2 21900895   1610612766               CHA Charlotte   1629023 P.J. Washington
## 3 21900895   1610612766               CHA Charlotte    202687 Bismack Biyombo
## 4 21900895   1610612766               CHA Charlotte   1628984 Devonte' Graham
## 5 21900895   1610612766               CHA Charlotte   1626179    Terry Rozier
## 6 21900895   1610612766               CHA Charlotte   1628998     Cody Martin
##   START_POSITION HOURS MINUTES FGM FGA FG_PCT FG3M FG3A FG3_PCT FTM FTA FT_PCT
## 1              F    35      15   3  13  0.231    1    7   0.143   0   0      0
## 2              F    31      52   5  14  0.357    1    8   0.125   1   1      1
## 3              C    22      07   2   8  0.250    0    0   0.000   4   4      1
## 4              G    32      21   7  18  0.389    3    8   0.375   0   1      0
## 5              G    36      05   6  18  0.333    0    3   0.000   1   1      1
## 6                   29      08   4   8  0.500    2    5   0.400   1   1      1
##   OREB DREB REB AST STL BLK TO PF PTS PLUS_MINUS GAME_DATE_EST GAME_STATUS_TEXT
## 1    1    3   4   2   2   2  2  2   7         -4    2020-03-01            Final
## 2    1    5   6   3   0   2  2  3  12        -13    2020-03-01            Final
## 3    4    5   9   2   0   2  1  3   8        -15    2020-03-01            Final
## 4    0    2   2   3   1   0  0  2  17        -14    2020-03-01            Final
## 5    2    1   3   4   1   0  2  2  13        -20    2020-03-01            Final
## 6    0    5   5   2   0   1  1  2  11          2    2020-03-01            Final
##   VISITOR_TEAM_ID SEASON PTS_home FG_PCT_home FT_PCT_home FG3_PCT_home AST_home
## 1      1610612749   2019       85       0.354         0.9        0.229       22
## 2      1610612749   2019       85       0.354         0.9        0.229       22
## 3      1610612749   2019       85       0.354         0.9        0.229       22
## 4      1610612749   2019       85       0.354         0.9        0.229       22
## 5      1610612749   2019       85       0.354         0.9        0.229       22
## 6      1610612749   2019       85       0.354         0.9        0.229       22
##   REB_home PTS_away FG_PCT_away FT_PCT_away FG3_PCT_away AST_away REB_away
## 1       47       93       0.402       0.762        0.226       20       61
## 2       47       93       0.402       0.762        0.226       20       61
## 3       47       93       0.402       0.762        0.226       20       61
## 4       47       93       0.402       0.762        0.226       20       61
## 5       47       93       0.402       0.762        0.226       20       61
## 6       47       93       0.402       0.762        0.226       20       61
##   HOME_TEAM_WINS OREB_To_DREB_RATIO
## 1              0               0.33
## 2              0               0.20
## 3              0               0.80
## 4              0               0.00
## 5              0               2.00
## 6              0               0.00
```

6. Understand:

Displaying the attributes in the data set to understand the data types in the merged data set

```
#Displaying the datatype of the attributes before data type conversions
str(nba_games_dataset)
```

```
## 'data.frame':    288643 obs. of  46 variables:
##  $ GAME_ID           : int  21900895 21900895 21900895 21900895 21900895 21900895 21900895 21900895 21900895 21900895 ...
##  $ HOME_TEAM_ID      : int  1610612766 1610612766 1610612766 1610612766 1610612766 1610612766 1610612766 1610612766 16106
12766 1610612766 ...
##  $ TEAM_ABBREVIATION : chr  "CHA" "CHA" "CHA" "CHA" ...
##  $ TEAM_CITY         : chr  "Charlotte" "Charlotte" "Charlotte" "Charlotte" ...
##  $ PLAYER_ID         : int  1628970 1629023 202687 1628984 1626179 1628998 1629667 1626195 1628997 201587 ...
##  $ PLAYER_NAME       : chr  "Miles Bridges" "P.J. Washington" "Bismack Biyombo" "Devonte' Graham" ...
##  $ START_POSITION    : chr  "F" "F" "C" "G" ...
##  $ HOURS             : chr  "35" "31" "22" "32" ...
##  $ MINUTES           : chr  "15" "52" "07" "21" ...
##  $ FGM               : num  3 5 2 7 6 4 1 4 2 NA ...
##  $ FGA               : num  13 14 8 18 18 8 2 9 6 NA ...
##  $ FG_PCT            : num  0.231 0.357 0.25 0.389 0.333 0.5 0.5 0.444 0.333 NA ...
##  $ FG3M              : num  1 1 0 3 0 2 0 0 1 NA ...
##  $ FG3A              : num  7 8 0 8 3 5 1 1 2 NA ...
##  $ FG3_PCT           : num  0.143 0.125 0 0.375 0 0.4 0 0 0.5 NA ...
##  $ FTM               : num  0 1 4 0 1 1 0 2 0 NA ...
##  $ FTA               : num  0 1 4 1 1 1 0 2 0 NA ...
##  $ FT_PCT            : num  0 1 1 0 1 1 0 1 0 NA ...
##  $ OREB              : num  1 1 4 0 2 0 0 3 1 NA ...
##  $ DREB              : num  3 5 5 2 1 5 1 10 3 NA ...
##  $ REB               : num  4 6 9 2 3 5 1 13 4 NA ...
##  $ AST               : num  2 3 2 3 4 2 1 4 1 NA ...
##  $ STL               : num  2 0 0 1 1 0 0 2 1 NA ...
##  $ BLK               : num  2 2 2 0 0 1 1 0 0 NA ...
##  $ TO                : num  2 2 1 0 2 1 0 1 1 NA ...
##  $ PF                : num  2 3 3 2 2 2 1 0 3 NA ...
##  $ PTS               : num  7 12 8 17 13 11 2 10 5 NA ...
##  $ PLUS_MINUS        : num  -4 -13 -15 -14 -20 2 0 11 13 NA ...
##  $ GAME_DATE_EST     : chr  "2020-03-01" "2020-03-01" "2020-03-01" "2020-03-01" ...
##  $ GAME_STATUS_TEXT  : chr  "Final" "Final" "Final" "Final" ...
##  $ VISITOR_TEAM_ID   : int  1610612749 1610612749 1610612749 1610612749 1610612749 1610612749 1610612749 1610612749 16106
12749 1610612749 ...
##  $ SEASON            : int  2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 ...
##  $ PTS_home          : num  85 85 85 85 85 85 85 85 85 85 ...
##  $ FG_PCT_home       : num  0.354 0.354 0.354 0.354 0.354 0.354 0.354 0.354 0.354 0.354 ...
##  $ FT_PCT_home       : num  0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 ...
##  $ FG3_PCT_home      : num  0.229 0.229 0.229 0.229 0.229 0.229 0.229 0.229 0.229 0.229 ...
##  $ AST_home          : num  22 22 22 22 22 22 22 22 22 22 ...
##  $ REB_home          : num  47 47 47 47 47 47 47 47 47 47 ...
##  $ PTS_away          : num  93 93 93 93 93 93 93 93 93 93 ...
##  $ FG_PCT_away       : num  0.402 0.402 0.402 0.402 0.402 0.402 0.402 0.402 0.402 0.402 ...
##  $ FT_PCT_away       : num  0.762 0.762 0.762 0.762 0.762 0.762 0.762 0.762 0.762 0.762 ...
##  $ FG3_PCT_away      : num  0.226 0.226 0.226 0.226 0.226 0.226 0.226 0.226 0.226 0.226 ...
##  $ AST_away          : num  20 20 20 20 20 20 20 20 20 20 ...
##  $ REB_away          : num  61 61 61 61 61 61 61 61 61 61 ...
##  $ HOME_TEAM_WINS    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ OREB_To_DREB_RATIO: num  0.33 0.2 0.8 0 2 0 0 0.3 0.33 NA ...
```

GAME_ID, HOME_TEAM_ID,VISITOR_TEAM_ID, PLAYER_ID are not factors in the merged dataset, which needs to be converted into a factor because Identification numbers are factors.

Checking the Data Type

```
a <- nba_games_dataset$GAME_ID  %>% class()
b <- nba_games_dataset$HOME_TEAM_ID  %>% class()
c <- nba_games_dataset$VISITOR_TEAM_ID  %>% class()
d <- nba_games_dataset$PLAYER_ID %>% class()
print(c(a,b,c,d))
```

```
## [1] "integer" "integer" "integer" "integer"
```

After Data Type Conversion

```
#Conversion to Factor
nba_games_dataset$GAME_ID <- nba_games_dataset$GAME_ID  %>% as.factor()
nba_games_dataset$HOME_TEAM_ID <- nba_games_dataset$HOME_TEAM_ID  %>% as.factor()
nba_games_dataset$VISITOR_TEAM_ID <- nba_games_dataset$VISITOR_TEAM_ID  %>% as.factor()
nba_games_dataset$PLAYER_ID <- nba_games_dataset$PLAYER_ID %>% as.integer() %>% as.factor()
a <- nba_games_dataset$GAME_ID  %>% class()
b <- nba_games_dataset$HOME_TEAM_ID  %>% class()
c <- nba_games_dataset$VISITOR_TEAM_ID  %>% class()
d <- nba_games_dataset$PLAYER_ID %>% class()
print(c(a,b,c,d))
```

```
## [1] "factor" "factor" "factor" "factor"
```

TEAM_CITY, TEAM_ABBREVIATION,PLAYER_NAME are all character variables. So no changes need to be made.

```
#Checking the data type
nba_games_dataset$TEAM_CITY  %>% class()
```

```
## [1] "character"
```

```
nba_games_dataset$TEAM_ABBREVIATION  %>% class()
```

```
## [1] "character"
```

```
nba_games_dataset$PLAYER_NAME  %>% class()
```

```
## [1] "character"
```

START_POSITION is a character variable in the merged dataset. In reality it is a ordered factor variable where positions are ranked by height. Guard(G)Forward(F)<Center(C). G is the smallest player on the team and C is the largest player on the team. Therefore START POSITION needs to be converted to a ordered factor variable.

Checking the data type of START_POSITION

```
nba_games_dataset$START_POSITION %>% class()
```

```
## [1] "character"
```

Converting START_POSITION to an ordered factor variable and checking the levels of the factor variable

```
nba_games_dataset$START_POSITION<-factor(nba_games_dataset$START_POSITION ,levels=c("G", "F", "C"),ordered=TRUE)
head(nba_games_dataset$START_POSITION)
```

```
## [1] F    F    C    G    G    <NA>
## Levels: G < F < C
```

```
#Checking the levels of the factor variable
levels(nba_games_dataset$START_POSITION)
```

```
## [1] "G" "F" "C"
```

HOURS and MINUTES are all characters variables in the merged data set.

```
#Checking the data type
nba_games_dataset$HOURS  %>% class()
```

```
## [1] "character"
```

```
nba_games_dataset$MINUTES  %>% class()
```

```
## [1] "character"
```

HOURS and MINUTES are all numeric variables, more specifically, Integer variables logically. Converting the data type of HOURS MINUTES TO numeric variables

```
nba_games_dataset$HOURS<-nba_games_dataset$HOURS  %>% as.numeric()
nba_games_dataset$MINUTES<-nba_games_dataset$MINUTES  %>% as.numeric()
#Checking the data type
nba_games_dataset$HOURS  %>% class()
```

```
## [1] "numeric"
```

```
nba_games_dataset$MINUTES  %>% class()
```

```
## [1] "numeric"
```

FGM,FGA,FG3M,FTM,FTA,OREB,DREB,REB,AST,STL,BLK,TO,
PF,PTS,PLUS_MINUS,PTS_home,AST_home,REB_home,PTS_away,AST_away,REB_away are double variables(i.e. numeric) in the merged dataset. Therefore no data type conversions needs to be made.

```
#Checking the data type
a <- nba_games_dataset$FGM  %>% class()
b <- nba_games_dataset$FGA  %>% class()
c <- nba_games_dataset$FG3M  %>% class()
d <- nba_games_dataset$FTM  %>% class()
e <- nba_games_dataset$FTA  %>% class()
f <- nba_games_dataset$OREB  %>% class()
g <- nba_games_dataset$DREB  %>% class()
h <- nba_games_dataset$REB  %>% class()
i <- nba_games_dataset$AST%>% class()
j <- nba_games_dataset$STL  %>% class()
k <- nba_games_dataset$BLK  %>% class()
l <- nba_games_dataset$TO%>% class()
m <- nba_games_dataset$PF %>% class()
n <- nba_games_dataset$PTS %>% class()
o <- nba_games_dataset$PLUS_MINUS %>% class()
p <- nba_games_dataset$PTS_home %>% class()
q <- nba_games_dataset$AST_home%>% class()
r <- nba_games_dataset$REB_home %>% class()
s <- nba_games_dataset$PTS_away %>% class()
t <- nba_games_dataset$AST_away%>% class()
u <- nba_games_dataset$REB_away %>% class()
print(c(a,b,c,d,e,f,g,h,i,j,l,m,n,o,p,q,r,s,t,u))
```

```
##  [1] "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
##  [8] "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
## [15] "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
```

FG_PCT,FG3_PCT,FT_PCT,FT_PCT_home,FG3_PCT_home,FT_PCT_away,FG3_PCT_away are all double variables (i.e.) because they contain percentage value.Therefore, no data type conversion need to be made.

```
#Checking the data type
a <- nba_games_dataset$FG_PCT  %>% typeof()
b <- nba_games_dataset$FG3_PCT %>% typeof()
c <- nba_games_dataset$FT_PCT  %>% typeof()
d <- nba_games_dataset$FT_PCT_home  %>% typeof()
e <- nba_games_dataset$FG3_PCT_home  %>% typeof()
f <- nba_games_dataset$FT_PCT_away  %>% typeof()
g <- nba_games_dataset$FG3_PCT_away  %>% typeof()

print(c(a,b,c,d,e,f,g))
```

```
## [1] "double" "double" "double" "double" "double" "double" "double"
```

SEASON is a numeric variable in the merged dataset. In reality is a factor variable.

```
#Checking the datatype of SEASON Variable
nba_games_dataset$SEASON %>% class()
```

```
## [1] "integer"
```

Converting the SEASON variable to factor variable

```
nba_games_dataset$SEASON<-nba_games_dataset$SEASON %>% as.factor()
#Checking the data type after data type conversion
nba_games_dataset$SEASON %>% class()
```

```
## [1] "factor"
```

GAME_STATUS_TEXT is a character variable in the merged dataset. However, it is a factor variable with “FINAL” being one of the states. The other states might include “HALF-TIME” and others

```
#Checking the data type of GAME_STATUS_TEXT
nba_games_dataset$GAME_STATUS_TEXT %>% class()
```

```
## [1] "character"
```

Converting the GAME_STATUS_TEXT variable to factor variable

```
#Converting the GAME_STATUS_TEXT variable to factor variable
nba_games_dataset$GAME_STATUS_TEXT<-nba_games_dataset$GAME_STATUS_TEXT %>% as.factor()
#Checking the data type of of GAME_STATUS_TEXT
nba_games_dataset$GAME_STATUS_TEXT %>% class()
```

```
## [1] "factor"
```

HOME_TEAM_WINS is a numeric variable is the merged dataset. However it is encoded ordered factor where 0 means loss and 1 means WIN. The hierarchy is LOSS<WIN

```
#Checking the data type of HOME_TEAM_WINS
nba_games_dataset$HOME_TEAM_WINS %>% class()
```

```
## [1] "integer"
```

Changing it to the right hierarchy.

```
nba_games_dataset$HOME_TEAM_WINS <- factor(nba_games_dataset$HOME_TEAM_WINS,levels = c(0,1),labels = c("LOSS","WIN"),ordered
= TRUE)
levels(nba_games_dataset$HOME_TEAM_WINS )
```

```
## [1] "LOSS" "WIN"
```

```
#Checking the data type
nba_games_dataset$HOME_TEAM_WINS %>% class()
```

```
## [1] "ordered" "factor"
```

GAME_DATE_EST is a character in the merged data set. So the string must be converted to date.

```
#Checking the data type of GAME_DATE_EST
nba_games_dataset$GAME_DATE_EST  %>% class()
```

```
## [1] "character"
```

```
#Converting the string to date
nba_games_dataset$GAME_DATE_EST <- nba_games_dataset$GAME_DATE_EST  %>% as.Date(format = "%YYYY/%MM/%DD")
#Checking the data type
nba_games_dataset$GAME_DATE_EST  %>% class()
```

```
## [1] "Date"
```

The mutated variable OREB_To_DREB_RATIO is a ratio, the data type must be “double” which it already is in the original data set.

```
#Checking the data type of Offensive_Rebound_To_Defensive_Rebound_Ratio
nba_games_dataset$OREB_To_DREB_RATIO %>% typeof()
```

```
## [1] "double"
```

7. Feature Selection:

Due to the constraints of the maximum page limit of 25 pages, we are selecting important features to reduce the number of features in the merged data set. FG_PCT,FGM and FGA are multi-collinear features since FG_PCT can be derived from FGM and FGA, therefore dropping FGM and FGA. FG_3PCT,FG3M,FG3A are multi-collinear features since FG_3PCT can be derived from FG3M and FG3A, therefore dropping FG3M and FG3A. FT_PCT,FTA,FTM are multi-collinear features since FT_PCT can be derived from FTM and FTA, therefore dropping FTM and FTA. There is a high degree of multi-collinearity between REB,OREB,DREB since REB = (OREB+DREB), therefore dropping OREB and DREB from the merged data set.

GAME_ID,PLAYER_ID,HOME_TEAM_ID and VISITOR_TEAM_ID are identifiers which have a sequential pattern. If this kind of identifiers are inserted into a model, it will introduce biases and learning wrong patterns. Therefore these attributes are dropped from the merged data set.

```
#Performing Feature Selection by dropping columns which might be problematic for machine learning
nba_games_dataset <- nba_games_dataset %>% subset( select= -c(OREB,DREB,FGM,FGA,FG3M,FG3A,FTM,FTA,GAME_ID,PLAYER_ID,HOME_TEA
M_ID,VISITOR_TEAM_ID) )
```

8. Scan I:

We are scanning for any “NA” values in the data set, which are the missing values,as it can affect the results of analysis.

```
# creating a dataframe with the sum of na values in the data set
a<-  data.frame(colSums(is.na (nba_games_dataset  )))
a
```

```
##                   colSums.is.na.nba_games_dataset..
## TEAM_ABBREVIATION                                 0
## TEAM_CITY                                         0
## PLAYER_NAME                                       0
## START_POSITION                               177959
## HOURS                                         46304
## MINUTES                                       58253
## FG_PCT                                        46304
## FG3_PCT                                       46304
## FT_PCT                                        46304
## REB                                           46304
## AST                                           46304
## STL                                           46304
## BLK                                           46304
## TO                                            46304
## PF                                            46304
## PTS                                           46304
## PLUS_MINUS                                    58253
## GAME_DATE_EST                                288643
## GAME_STATUS_TEXT                                  0
## SEASON                                            0
## PTS_home                                          0
## FG_PCT_home                                       0
## FT_PCT_home                                       0
## FG3_PCT_home                                      0
## AST_home                                          0
## REB_home                                          0
## PTS_away                                          0
## FG_PCT_away                                       0
## FT_PCT_away                                       0
## FG3_PCT_away                                      0
## AST_away                                          0
## REB_away                                          0
## HOME_TEAM_WINS                                    0
## OREB_To_DREB_RATIO                            46304
```

From the output of the columns and their missing values are as follows:

START_POSITION 177959,
MINUTES 58253,
FG_PCT 46304,
FG3_PCT 46304,
FT_PCT 46304,
REB 46304,

As it is seen that major columns have a missing values of 46304 which would not add any value and if replaced by any other value it would give wrong results while analysis and hence omitting these rows.

```
 nba_games_dataset <- nba_games_dataset %>% drop_na(REB)
#Checking the dimensions after dropping the all the rows with NA values
dim(nba_games_dataset)
```

```
## [1] 242339     34
```

```
# checking the missing values
b<- data.frame( colSums(is.na(nba_games_dataset)))
b
```

```
##                   colSums.is.na.nba_games_dataset..
## TEAM_ABBREVIATION                                 0
## TEAM_CITY                                         0
## PLAYER_NAME                                       0
## START_POSITION                               131655
## HOURS                                             0
## MINUTES                                       11949
## FG_PCT                                            0
## FG3_PCT                                           0
## FT_PCT                                            0
## REB                                               0
## AST                                               0
## STL                                               0
## BLK                                               0
## TO                                                0
## PF                                                0
## PTS                                               0
## PLUS_MINUS                                    11949
## GAME_DATE_EST                                242339
## GAME_STATUS_TEXT                                  0
## SEASON                                            0
## PTS_home                                          0
## FG_PCT_home                                       0
## FT_PCT_home                                       0
## FG3_PCT_home                                      0
## AST_home                                          0
## REB_home                                          0
## PTS_away                                          0
## FG_PCT_away                                       0
## FT_PCT_away                                       0
## FG3_PCT_away                                      0
## AST_away                                          0
## REB_away                                          0
## HOME_TEAM_WINS                                    0
## OREB_To_DREB_RATIO                                0
```

We are left with missing values from

START_POSITION - 131655, MINUTES - 11949,

As the time column was split into HOURS and MINUTES all of the columns have the same number of missing values and since the hours played for each match are almost the same, replacing it with mean values of the columns.

```
# replacing the HOURS and MINUTES  column.
nba_games_dataset$HOURS[is.na(nba_games_dataset$HOURS)] <- mean(nba_games_dataset$HOURS, na.rm =TRUE)
nba_games_dataset$MINUTES[is.na(nba_games_dataset$MINUTES)] <- mean(nba_games_dataset$MINUTES, na.rm =TRUE)
```

As the start position has many missing values and the guards(“G”) can play small Forward position which is “F”, and Center(“C”) players can also play in the power Forward position which is “F”, hence replacing the na values with “F”

```
# checking the levels of the column
levels(nba_games_dataset$START_POSITION)
```

```
## [1] "G" "F" "C"
```

```
# replacing
nba_games_dataset$START_POSITION[is.na(nba_games_dataset$START_POSITION)] <- "F"
# checking if it is factor
is.factor(nba_games_dataset$START_POSITION)
```

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

```
# Checking if all the na values were replaced
b<- data.frame( colSums(is.na(nba_games_dataset)))
b
```

```
##                   colSums.is.na.nba_games_dataset..
## TEAM_ABBREVIATION                                 0
## TEAM_CITY                                         0
## PLAYER_NAME                                       0
## START_POSITION                                    0
## HOURS                                             0
## MINUTES                                           0
## FG_PCT                                            0
## FG3_PCT                                           0
## FT_PCT                                            0
## REB                                               0
## AST                                               0
## STL                                               0
## BLK                                               0
## TO                                                0
## PF                                                0
## PTS                                               0
## PLUS_MINUS                                    11949
## GAME_DATE_EST                                242339
## GAME_STATUS_TEXT                                  0
## SEASON                                            0
## PTS_home                                          0
## FG_PCT_home                                       0
## FT_PCT_home                                       0
## FG3_PCT_home                                      0
## AST_home                                          0
## REB_home                                          0
## PTS_away                                          0
## FG_PCT_away                                       0
## FT_PCT_away                                       0
## FG3_PCT_away                                      0
## AST_away                                          0
## REB_away                                          0
## HOME_TEAM_WINS                                    0
## OREB_To_DREB_RATIO                                0
```

```
#Checking the dimension of the merged dataset
dim(nba_games_dataset)
```

```
## [1] 242339     34
```

From the last output it can be seen that the data set does not have any missing values anymore.

9. Scan II:

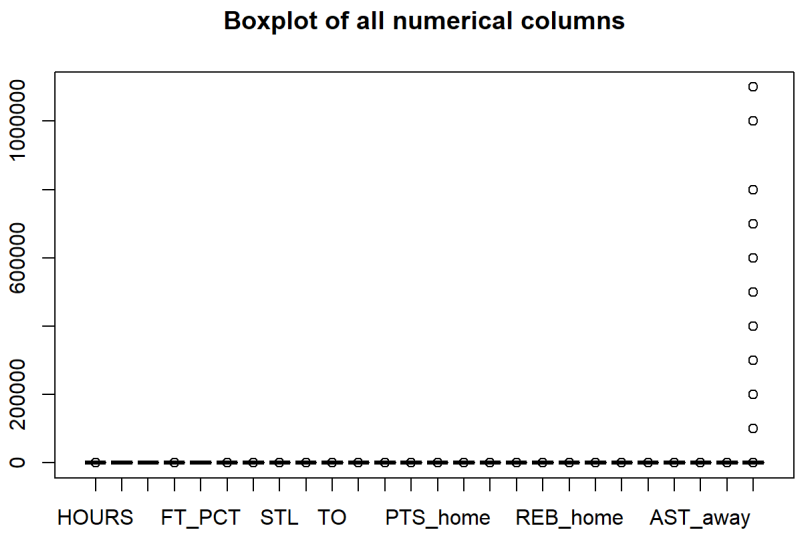
In this section we are scanning for outliers in all of the numeric columns.

```
# creating a data set with only the numerical columns called d
d<-data.frame(select_if(nba_games_dataset, is.numeric))
str(d)
```

```
## 'data.frame':    242339 obs. of  26 variables:
##  $ HOURS            : num  35 31 22 32 36 29 9 20 23 22 ...
##  $ MINUTES          : num  15 52 7 21 5 8 21 21 30 59 ...
##  $ FG_PCT           : num  0.231 0.357 0.25 0.389 0.333 0.5 0.5 0.444 0.333 0.5 ...
##  $ FG3_PCT          : num  0.143 0.125 0 0.375 0 0.4 0 0 0.5 0.5 ...
##  $ FT_PCT           : num  0 1 1 0 1 1 0 1 0 0 ...
##  $ REB              : num  4 6 9 2 3 5 1 13 4 6 ...
##  $ AST              : num  2 3 2 3 4 2 1 4 1 1 ...
##  $ STL              : num  2 0 0 1 1 0 0 2 1 1 ...
##  $ BLK              : num  2 2 2 0 0 1 1 0 0 2 ...
##  $ TO               : num  2 2 1 0 2 1 0 1 1 2 ...
##  $ PF               : num  2 3 3 2 2 2 1 0 3 3 ...
##  $ PTS              : num  7 12 8 17 13 11 2 10 5 9 ...
##  $ PLUS_MINUS       : num  -4 -13 -15 -14 -20 2 0 11 13 -9 ...
##  $ PTS_home         : num  85 85 85 85 85 85 85 85 85 91 ...
##  $ FG_PCT_home      : num  0.354 0.354 0.354 0.354 0.354 0.354 0.354 0.354 0.354 0.364 ...
##  $ FT_PCT_home      : num  0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.4 ...
##  $ FG3_PCT_home     : num  0.229 0.229 0.229 0.229 0.229 0.229 0.229 0.229 0.229 0.31 ...
##  $ AST_home         : num  22 22 22 22 22 22 22 22 22 19 ...
##  $ REB_home         : num  47 47 47 47 47 47 47 47 47 57 ...
##  $ PTS_away         : num  93 93 93 93 93 93 93 93 93 111 ...
##  $ FG_PCT_away      : num  0.402 0.402 0.402 0.402 0.402 0.402 0.402 0.402 0.402 0.468 ...
##  $ FT_PCT_away      : num  0.762 0.762 0.762 0.762 0.762 0.762 0.762 0.762 0.762 0.632 ...
##  $ FG3_PCT_away     : num  0.226 0.226 0.226 0.226 0.226 0.226 0.226 0.226 0.226 0.275 ...
##  $ AST_away         : num  20 20 20 20 20 20 20 20 28 ...
##  $ REB_away         : num  61 61 61 61 61 61 61 61 56 ...
##  $ OREB_To_DREB_RATIO: num  0.33 0.2 0.8 0 2 0 0 0.3 0.33 1 ...
```

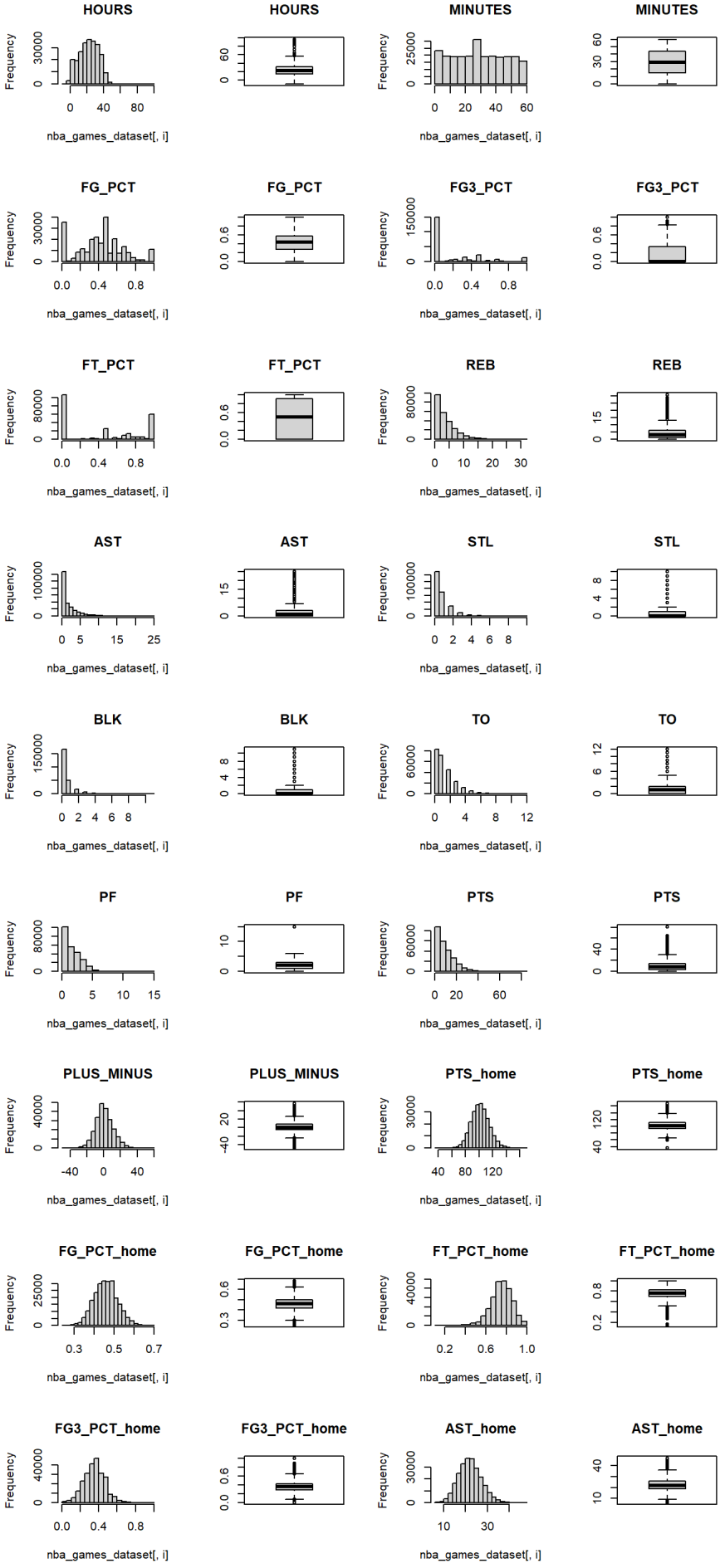
Plotting the box-plot to visualize the outliers

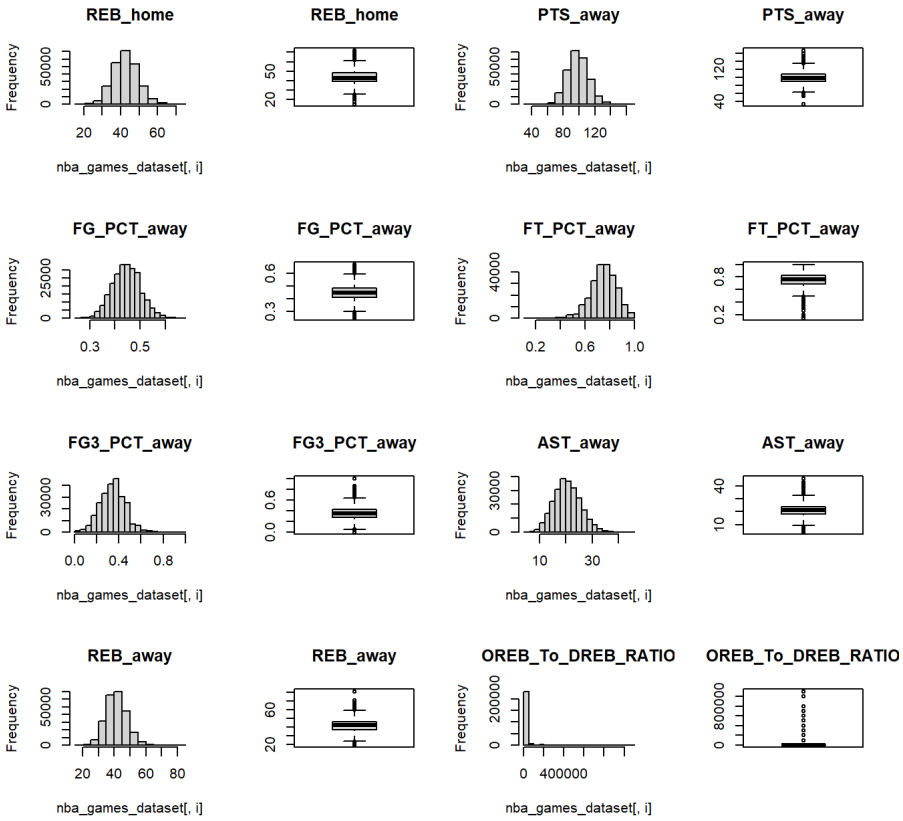
```
#Plotting the boxplot
boxplot(d,main="Boxplot of all numerical columns")
```



Since the box-plot does not show detailed display all the columns and the outliers clearly, we will have to plot individual box plot to identify any kind of outliers. We have also plotted the histograms of the data set to better understand the outliers.

```
par(mfrow=c(3,4))
# plotting the histogram
for(i in colnames(nba_games_dataset[,c("HOURS","MINUTES","FG_PCT","FG3_PCT","FT_PCT","REB","AST","STL","BLK","TO","PF","PTS"
,"PLUS_MINUS","PTS_home","FG_PCT_home","FT_PCT_home","FG3_PCT_home","AST_home","REB_home","PTS_away","FG_PCT_away","FT_PCT_a
way","FG3_PCT_away","AST_away","REB_away","OREB_To_DREB_RATIO")]))
{
  hist(nba_games_dataset[,i],main = colnames(nba_games_dataset[i]))
  boxplot(nba_games_dataset[i])
  title(colnames(nba_games_dataset[i]))
}
```


HOURS
HOURS
MINUTES
MINUTES
FG_PCT
FG_PCT
FG3_PCT
FG3_PCT
FT_PCT
FT_PCT
REB
REB
AST
AST
STL
STL
BLK
BLK
TO
TO
PF
PF
PTS
PTS
PLUS_MINUS
PLUS_MINUS
PTS_home
PTS_home
FG_PCT_home
FG_PCT_home
FT_PCT_home
FT_PCT_home
FG3_PCT_home
FG3_PCT_home
AST_home
AST_home
Frequency
nba_games_dataset[, i]



From the box-plots and histograms it can be seen that all the numerical columns except PLUS_MINUS,PTS_HOme,FG_PCT_home,REB_home,PTS_away and FG_PCT_away are all skewed. Since PLUS_MINUS,PTS_HOme,FG_PCT_home,REB_home,PTS_away and FG_PCT_away is normally distributed, we could have used the z-score to only identify the outliers. But we have chosen “Capping” method, as it does not require for the columns to be normally distributed, allowing us to us this method for all the numerical columns regardless of the distribution of the data.

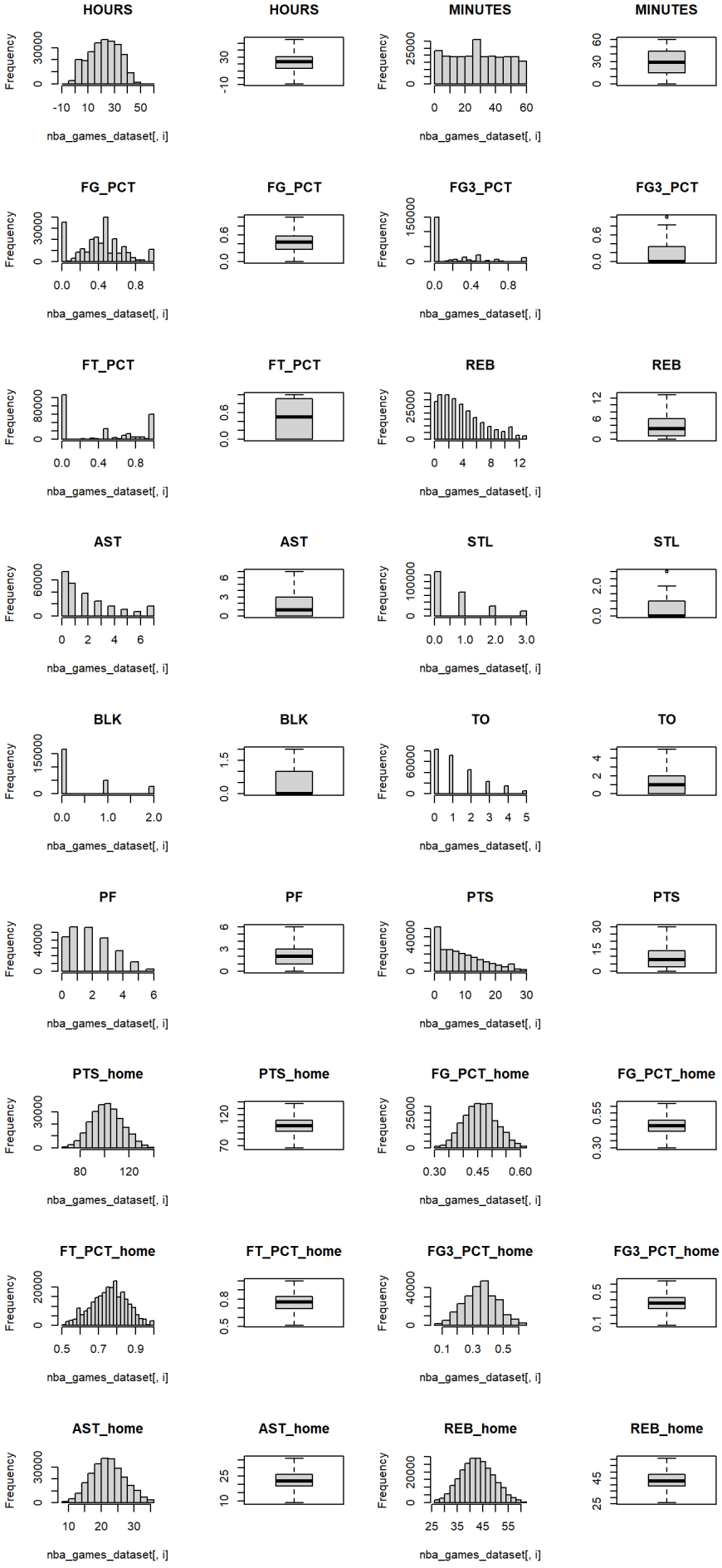
As we can see from the box-plot, there are multiple columns with outliers, HOURS,FG3_PCT,REB,AST,STL,BLK,TO,PF,PTS,PLUS_MINUS,PTS_home,FG_PCT_home,FT_PCT_home,FG3_PCT,AST_home,REB_home,PTS_away,FG_ and REB_away. Eliminating these outliers would not be right for analysis since these outliers do not present any meaningful information and would not affect the data analysis of the data. The advantage of “Capping” method is that it replaces the outliers with the best value near it.This makes this method an appropriate method for outlier detection of all numerical columns and imputations with appropriate values.

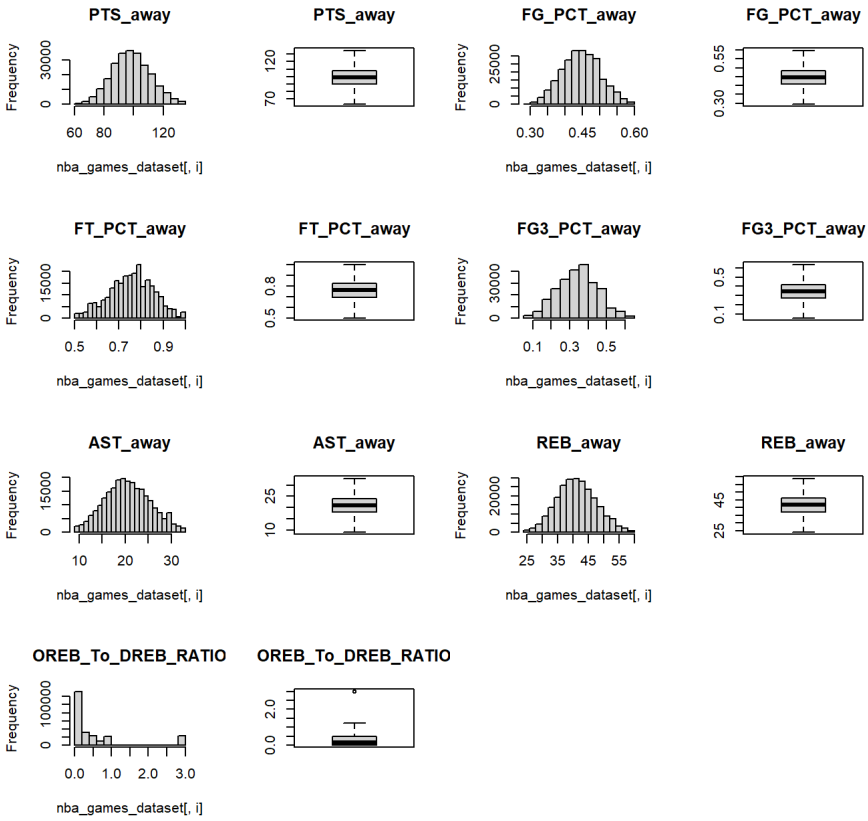
```
# Defining the function for the "capping" method based on the upper and lower fences of the boxplot
cap <-function(x){ quantiles <- quantile( x, c(.05,0.25,0.75,.95) )
x[ x < quantiles[2] -1.5*IQR(x) ] <- quantiles[1]
x[ x > quantiles[3] +1.5*IQR(x) ] <- quantiles[4]
x }

par(mfrow=c(3,4))

#Selecting the numerical columns
for (i in colnames(nba_games_dataset[,c("HOURS","MINUTES","FG_PCT","FG3_PCT","FT_PCT","REB","AST","STL","BLK","TO","PF","PTS","PTS_home","FG_PCT_home","FT_PCT_home","FG3_PCT_home","AST_home","REB_home","PTS_away","FG_PCT_away","FT_PCT_away","FG3_PCT_away","AST_away","REB_away","OREB_To_DREB_RATIO")])){
  nba_games_dataset[,c("HOURS","MINUTES","FG_PCT","FG3_PCT","FT_PCT","REB","AST","STL","BLK","TO","PF","PTS","PTS_home","FG_PCT_home","FT_PCT_home","FG3_PCT_home","AST_home","REB_home","PTS_away","FG_PCT_away","FT_PCT_away","FG3_PCT_away","AST_away","REB_away","OREB_To_DREB_RATIO")] <- sapply( nba_games_dataset[,c("HOURS","MINUTES","FG_PCT","FG3_PCT","FT_PCT","REB","AST","STL","BLK","TO","PF","PTS","PTS_home","FG_PCT_home","FT_PCT_home","FG3_PCT_home","AST_home","REB_home","PTS_away","FG_PCT_away","FT_PCT_away","FG3_PCT_away","AST_away","REB_away","OREB_To_DREB_RATIO")], FUN = cap)
}


#Plot the histograms and box-plots
for(i in colnames(nba_games_dataset[,c("HOURS","MINUTES","FG_PCT","FG3_PCT","FT_PCT","REB","AST","STL","BLK","TO","PF","PTS","PTS_home","FG_PCT_home","FT_PCT_home","FG3_PCT_home","AST_home","REB_home","PTS_away","FG_PCT_away","FT_PCT_away","FG3_PCT_away","AST_away","REB_away","OREB_To_DREB_RATIO")])){
  hist(nba_games_dataset[,i],main = colnames(nba_games_dataset[i]))
  boxplot(nba_games_dataset[i])
  title(colnames(nba_games_dataset[i]))
}
```


HOURS
HOURS
MINUTES
MINUTES
FG_PCT
FG_PCT
FG3_PCT
FG3_PCT
FT_PCT
FT_PCT
REB
REB
AST
AST
STL
STL
BLK
BLK
TO
TO
PF
PF
PTS
PTS
PTS_home
PTS_home
FG_PCT_home
FG_PCT_home
FT_PCT_home
FT_PCT_home
FG3_PCT_home
FG3_PCT_home
AST_home
AST_home
REB_home
REB_home
Frequency
nba_games_dataset[, i]



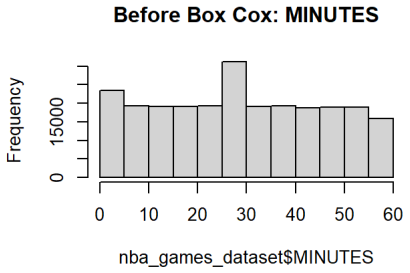
From the output of the box-plots above, it can be seen that there are no outliers remaining in almost all of the numerical columns, except for columns FG3_PCT,STL and OREB_To_DREB RATIO.. It can be also seen that after the outliers are removed, all the numerical columns except columns MINUTES,FG_PCT,FG3PCT,FT_PCT,AST,STL,TO,PF,PTS and OREB_To_DREB RATIO, have become normally distributed.

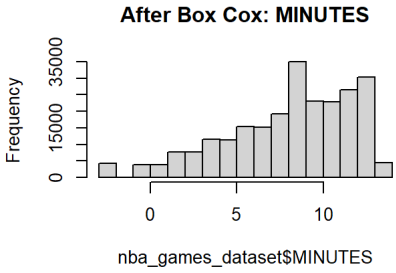
10. Transform

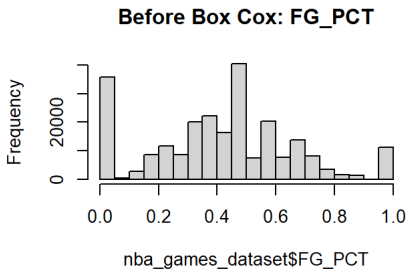
In this section, we are applying Box-Cox transformation to the columns to reduce the skewness of the numerical columns.Box-Cox Transformation is used instead it can be used to reduce both left and right skewness of the data. FG3_PCT, FT_PCT, AST, STL, TO,PF,PTS and OREB_To_DREB_RATIO columns which are not normally distributed.This is because most machine learning models, for example regression models, require the data to be normally distributed.

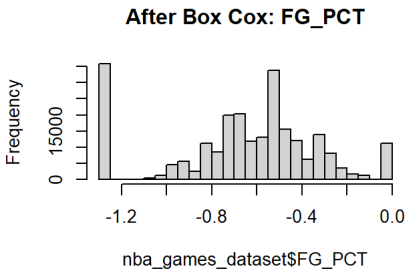
Box-Cox transformation is applied to reduce the skewness of the columns that are not normally distributed.

```
#Box Cox Transformation of all skewed numerical columns
#The code for Box-Cox Transformation
par(mfrow=c(2,2))
#Plotting the histogram of MINUTES before and after BOX COX transformation
hist(nba_games_dataset$MINUTES,main = "Before Box Cox: MINUTES")
nba_games_dataset$MINUTES<-BoxCox(nba_games_dataset$MINUTES,lambda = "auto")
hist(nba_games_dataset$MINUTES,main = "After Box Cox: MINUTES")
#Plotting the histogram of FG_PCT before and after BOX COX transformation
hist(nba_games_dataset$FG_PCT, main = "Before Box Cox: FG_PCT")
nba_games_dataset$FG_PCT<-BoxCox(nba_games_dataset$FG_PCT,lambda = "auto")
hist(nba_games_dataset$FG_PCT, main = "After Box Cox: FG_PCT")
```

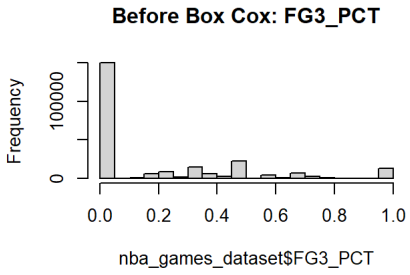


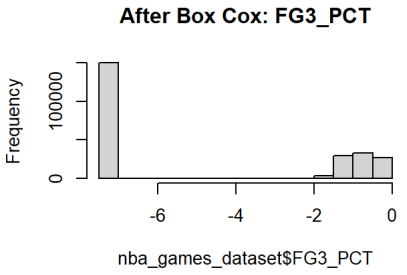


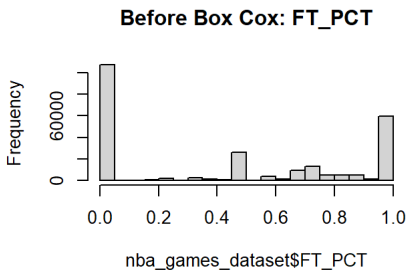


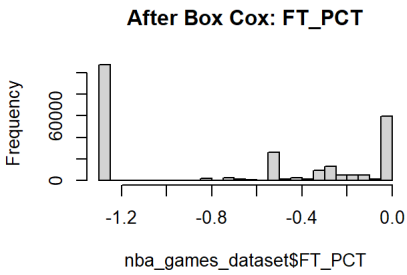



```
#Plotting the histogram of FG3_PCT before and after BOX COX transformation
hist(nba_games_dataset$FG3_PCT, main = "Before Box Cox: FG3_PCT")
nba_games_dataset$FG3_PCT<-BoxCox(nba_games_dataset$FG3_PCT,lambda = "auto")
hist(nba_games_dataset$FG3_PCT, main="After Box Cox: FG3_PCT")
#Plotting the histogram of FT_PCT before and after BOX COX transformation
hist(nba_games_dataset$FT_PCT, main = "Before Box Cox: FT_PCT")
nba_games_dataset$FT_PCT<-BoxCox(nba_games_dataset$FT_PCT,lambda = "auto")
hist(nba_games_dataset$FT_PCT, main="After Box Cox: FT_PCT")
```

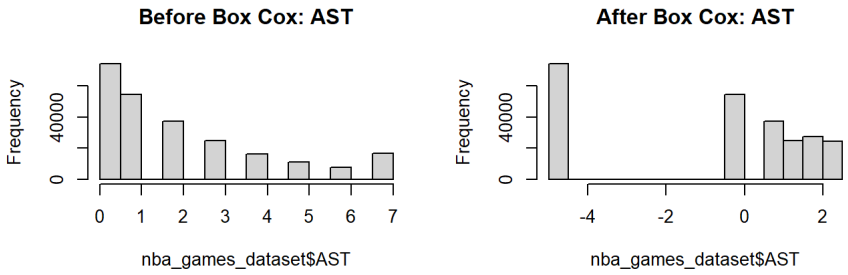


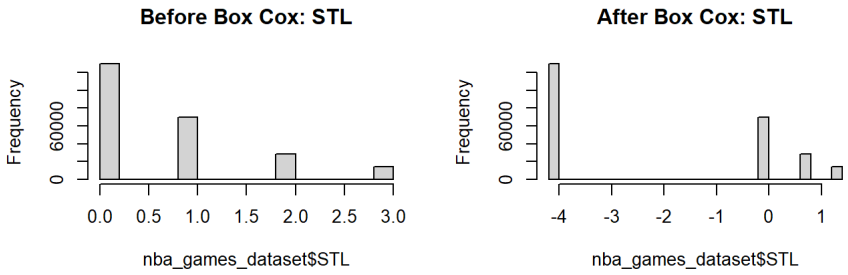




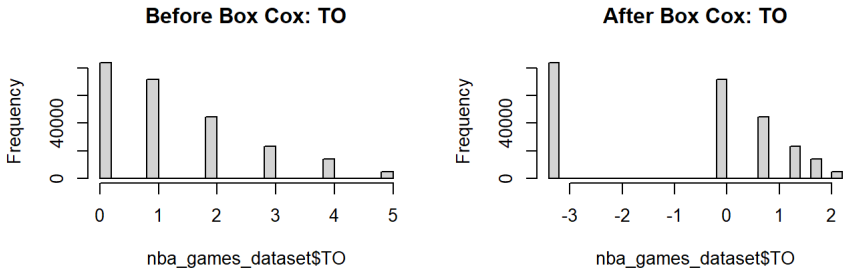


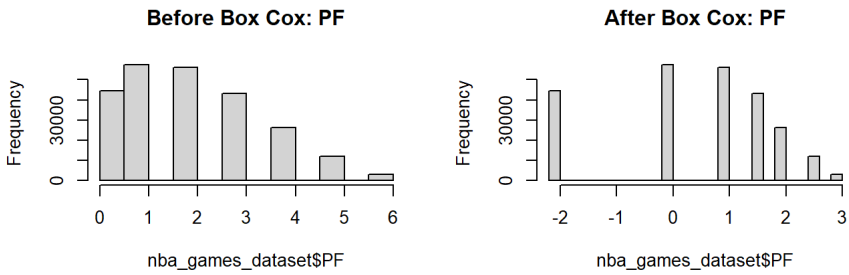

```
#Plotting the histogram of AST before and after BOX COX transformation
hist(nba_games_dataset$AST, main = "Before Box Cox: AST")
nba_games_dataset$AST<-BoxCox(nba_games_dataset$AST,lambda = "auto")
hist(nba_games_dataset$AST, main="After Box Cox: AST")
#Plotting the histogram of STL before and after BOX COX transformation
hist(nba_games_dataset$STL, main = "Before Box Cox: STL")
nba_games_dataset$STL<-BoxCox(nba_games_dataset$STL,lambda = "auto")
hist(nba_games_dataset$STL, main="After Box Cox: STL")
```



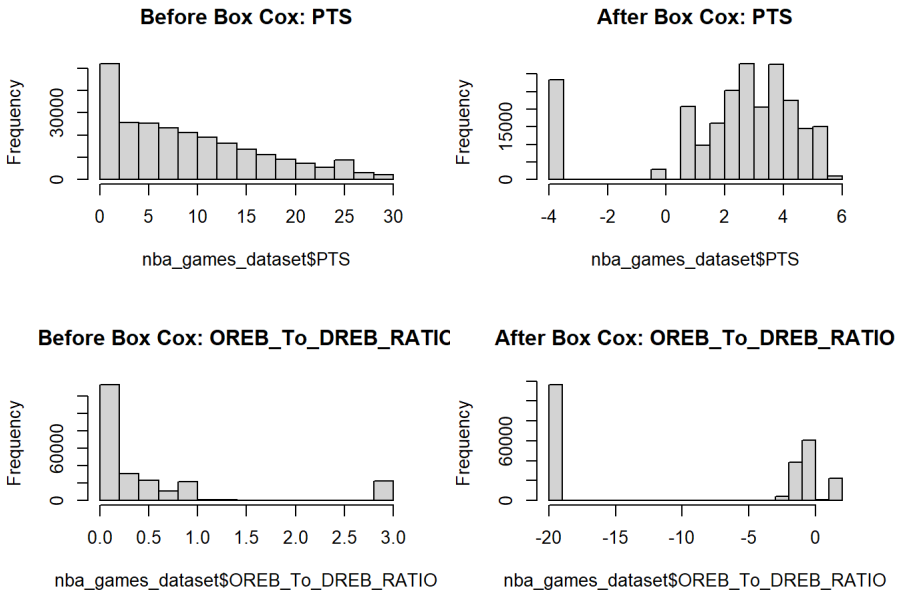



```
#Plotting the histogram of TO before and after BOX COX transformation
hist(nba_games_dataset$TO, main = "Before Box Cox: TO")
nba_games_dataset$TO<-BoxCox(nba_games_dataset$TO,lambda = "auto")
hist(nba_games_dataset$TO, main="After Box Cox: TO")
#Plotting the histogram of PF before and after BOX COX transformation
hist(nba_games_dataset$PF, main = "Before Box Cox: PF")
nba_games_dataset$PF<-BoxCox(nba_games_dataset$PF,lambda = "auto")
hist(nba_games_dataset$PF, main="After Box Cox: PF")
```






```
#Plotting the histogram of PTS before and after BOX COX transformation
hist(nba_games_dataset$PTS, main = "Before Box Cox: PTS")
nba_games_dataset$PTS<-BoxCox(nba_games_dataset$PTS,lambda = "auto")
hist(nba_games_dataset$PTS, main="After Box Cox: PTS")
#Plotting the histogram of OREB_To_DREB_RATIO before and after BOX COX transformation
hist(nba_games_dataset$OREB_To_DREB_RATIO, main = "Before Box Cox: OREB_To_DREB_RATIO")
nba_games_dataset$OREB_To_DREB_RATIO<-BoxCox(nba_games_dataset$OREB_To_DREB_RATIO,
                                             lambda = "auto")
hist(nba_games_dataset$OREB_To_DREB_RATIO,main="After Box Cox: OREB_To_DREB_RATIO")
```



From the comparison of histograms before and after Box-Cox transformation, it can be seen that the skewness of the numerical columns have decreased across most of the numerical columns.

11. Conclusion

The final “nba_games_dataset” is clean data set with no missing values with 242339 observations and 34 variables. Almost all of the outliers were removed and replaced by the capping method. Finally, transformation was applied to reduce the skewness, of all numerical variables which were not normally distributed. In short, the data was cleaned in order to ensure the data set does not have any discrepancies relating to the NBA games 2019-2020 data set.

References

1. En.wikipedia.org. 2020. Rules Of Basketball. [online] Available at: https://en.wikipedia.org/wiki/Rules_of_basketball (https://en.wikipedia.org/wiki/Rules_of_basketball) [Accessed 14 October 2020].
2. Kaggle.com. n.d. NBA Games Data. [online] Available at: https://www.kaggle.com/nathanlauga/nba-games?select=games.csv (https://www.kaggle.com/nathanlauga/nba-games?select=games.csv) [Accessed 13 October 2020].
3. Kaggle.com. n.d. NBA Games Data. [online] Available at: https://www.kaggle.com/nathanlauga/nba-games?select=games_details.csv (https://www.kaggle.com/nathanlauga/nba-games?select=games_details.csv) [Accessed 13 October 2020].
4. Ladd, T., 2020. NBA Official Basketball Rules And Regulations For Beginners. [online] Sportsierra. Available at: https://sportsierra.com/nba-official-basketball-rules-and-regulations/#:~:text=%20NBA%20basketball%20rules%20%201%20Regulation%20NBA,4%20Fighting%20and%20flagrant%20fouls.%20%20More%20 (https://sportsierra.com/nba-official-basketball-rules-and-regulations/#:~:text=%20NBA%20basketball%20rules%20%201%20Regulation%20NBA,4%20Fighting%20and%20flagrant%20fouls.%20%20More%20) [Accessed 14 October 2020].