# Project 1

**Sachin Samal**

**Sept 09, 2021**

**Data Mining with R** This is my first project in R. In this project, I have decided to work on a dataset of results of *380 Premier League matches in a single season of the year 2017/2018*. It is time to examine the performance of each team and predict which team has the most successful home and away season overall. Since the main goal is to predict the result of each football game, the most important question that we should first ask is what features would likely impose a direct impact on the result of a football game. Has the successful team at home managed to succeed away too? Is the title winner team successful at both the venues or just heavily reliant on home advantage? The list can go on and on. So let us start the fun part as we explore the dataset.

Lets load some data

```
library(readxl)
Results <- read_excel("Results.xlsx")
str(Results)
```

```
## tibble [380 x 6] (S3: tbl_df/tbl/data.frame)
##  $ Home_team : chr [1:380] "Arsenal" "Watford" "Chelsea" "Crystal Palace" ...
##  $ Away_team : chr [1:380] "Leicester City" "Liverpool" "Burnley" "Huddersfield Town" ...
##  $ Home_goal : num [1:380] 4 3 2 0 1 0 1 0 0 4 ...
##  $ Away_goals: num [1:380] 3 3 3 3 0 0 0 2 2 0 ...
##  $ Result    : chr [1:380] "H" "D" "A" "A" ...
##  $ Season    : chr [1:380] "2017-2018" "2017-2018" "2017-2018" "2017-2018" ...
```

```
summary(Results$Home_goal)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   1.000   1.532   2.000   7.000
```

### STATISTICS CALCULATIONS

```
mean(Results$Home_goal)
```

```
## [1] 1.531579
```

```
sd(Results$Home_goal)
```

```
## [1] 1.340087
```

### DESCRIPTIVE STATISTICS

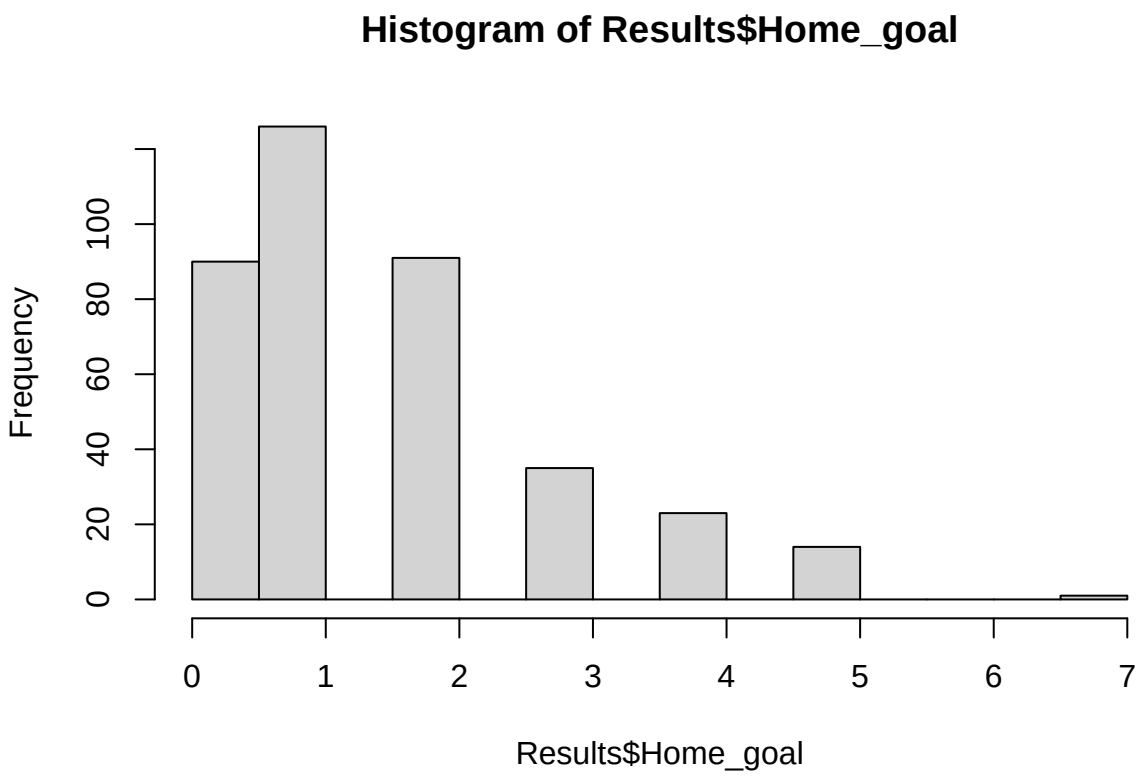
The summary shows the descriptive statistics for the qualitative data analytic variable. The average Home_goal was 1.5(where standard deviation(sd) = 1.3), with a minimum goal score of 1 and maximum goal score of 7 in the whole season 2017-2018.

### GRAPHICAL DISPLAYS

### HISTOGRAM

A histogram contains rectangular area to display the statistical information which is proportional to the frequency of a variable and its width in successive numerical intervals. It requires only 1 numeric variable as input. This function automatically cut the variable in bins and count the number of data point per bin.

```
#histogram graph
hist(Results$Home_goal)
```



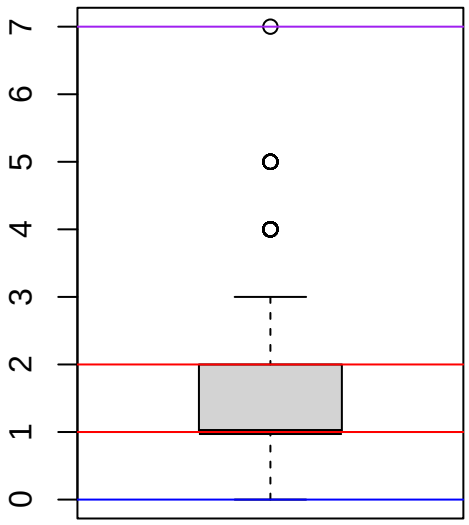

**From the histogram, there seems to be a couple of observations higher than all other observations (see the bar on the right side of the plot). From the figure, we can say that the distribution is right tailed. The no. of times 1 or 2 Home goals were scored is comparatively higher. There is a conclusive evidence from the graph that the Home team has managed to score the maximum score of 7 goals.**

### BOX PLOT

In addition to histograms, boxplots are also useful to detect potential outliers.

Using R base:

```
#boxplot
par(mfrow = c(1, 2))
boxplot(Results$Home_goal, name= "Home goals")
abline(h = min(Results$Home_goal), col = "Blue")
abline(h = max(Results$Home_goal), col = "Purple")
abline(h = median(Results$Home_goal), col = "Dark Green")
abline(h = quantile(Results$Home_goal, c(0.25, 0.75)), col = "Red")
```




A boxplot helps to visualize a quantitative variable by displaying five common location summary (minimum, median, first and third quartiles and maximum) and any observation that was classified as a suspected outlier using the interquartile range (IQR) criterion. Based on this criterion, there are about 3 potential outliers (see the 3 points above the vertical line, at the top of the boxplot).

It is also possible to extract the values of the potential outliers based on the IQR criterion thanks to the function

```
boxplot.stats(Results$Home_goal)$out
```

```
##  [1] 4 4 4 5 4 5 4 7 4 4 4 4 4 5 4 5 4 4 5 5 5 4 4 4 5 4 5 4 4 4 4 5 4 5 5 5 4 5
```

## CHECKING NORMALITY IN R (VISUAL METHODS)

**Density plot** and **Q-Q plot** can be used to check normality visually.

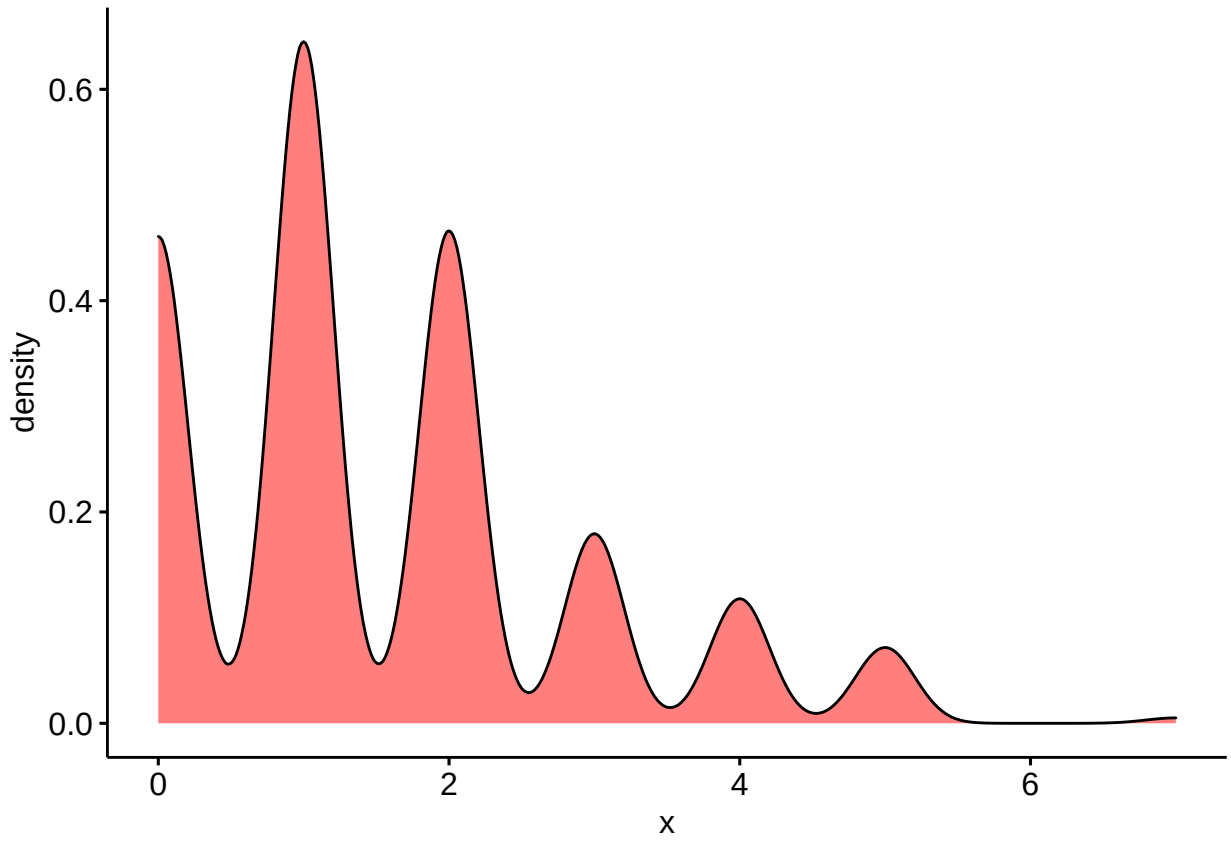
## DENSITY PLOT

The density plot provides a visual judgment about whether the distribution is bell shaped.

```
library("ggpubr")
```

```
## Loading required package: ggplot2
```

```
##Density plot
ggdensity(Results$Home_goal, fill = "Red")
```

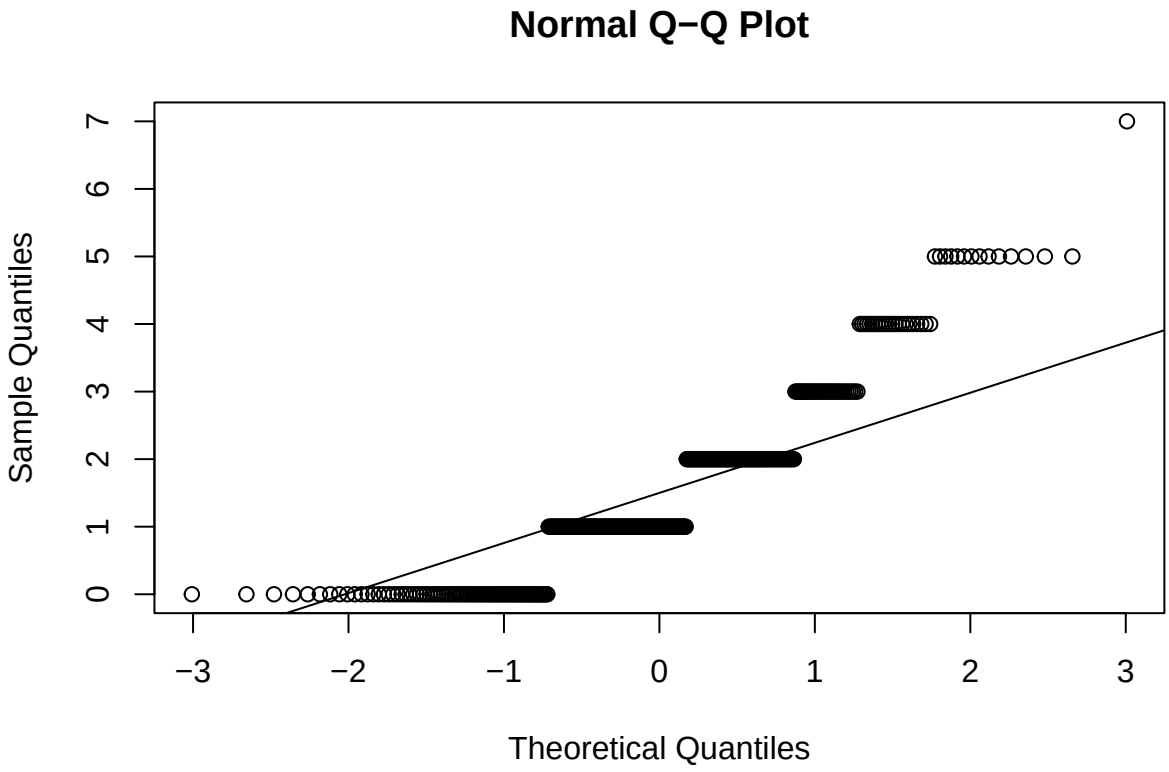



Since a relatively small number of data points in normally distributed data fall in the few highest and few lowest quantiles, we are more likely to see the results of random fluctuations at the extreme ends. We now understand that the Premier League Results data is not precisely normal, but not too far off.

### QQ PLOT

QQ plot (or quantile-quantile plot) draws the correlation between a given sample and the normal distribution. A 45-degree reference line is also plotted. In a QQ plot, each observation is plotted as a single dot. If the data are normal, the dots should form a straight line.

```
#qqnormal
qqnorm(Results$Home_goal)
qqline(Results$Home_goal)
```




This dataset is not normally distributed, but doesn't look that far off. The Q-Q plot clearly shows that the quantile points do not lie on the theoretical normal line. We see that the sample values are generally lower than the normal values for quantiles along the smaller side of the distribution. The points fall along a line in the middle of the graph, but curve off in the extremities. Normal Q-Q plots that exhibit this behavior usually mean that the data have more extreme values than would be expected if they truly came from a Normal distribution.

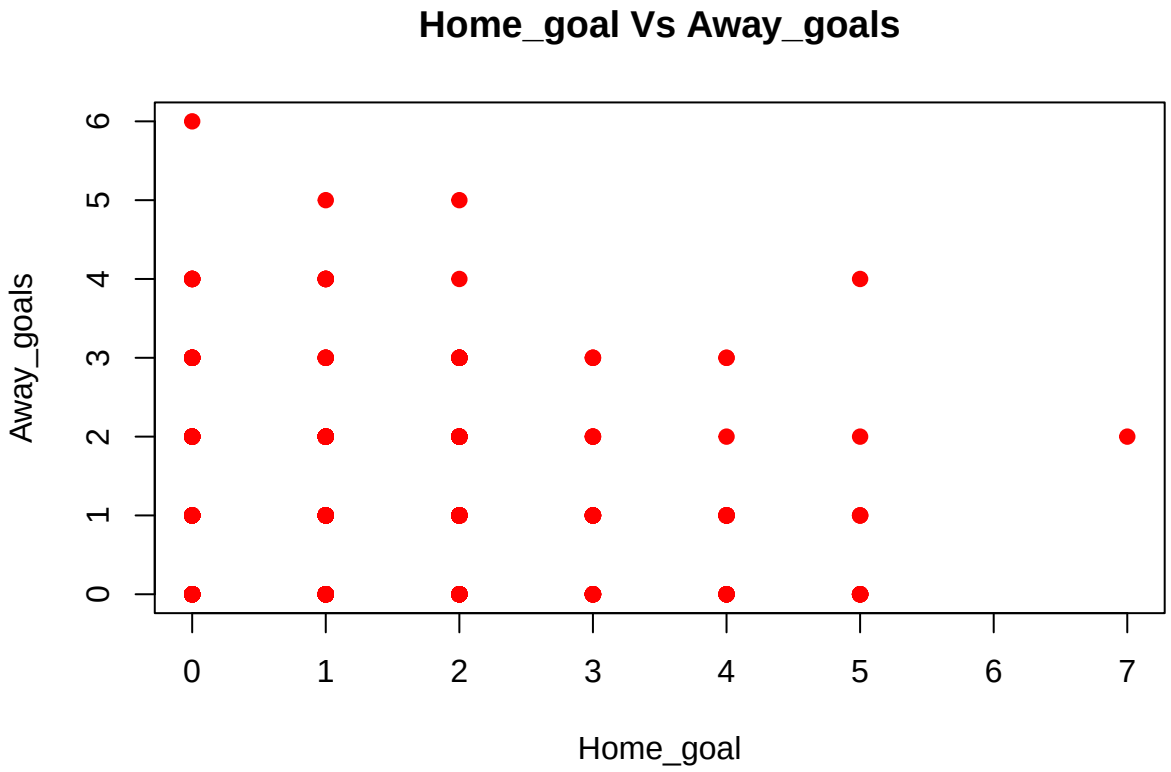
### SHAPE OF DISTRIBUTION

**From the histogram and density plot, we can say that the distribution is right tailed, This means that the mean is in the right edge of the graph and greater than the median.**

### SCATTER PLOT

A scatter plot shows the relationship between two quantitative variables measured for the same individuals. The values of one variable appear on the horizontal axis(here(x)=Home_goal), and the values of the other variable appear on the vertical axis(here(y)=Away_goals).

```
#scatter plot
plot(x=Results$Home_goal, y=Results$Away_goals, type="p",
xlab="Home_goal",
ylab="Away_goals",
main="Home_goal Vs Away_goals"
,pch=19,
col = "red",
cex=1)
```




Here, from the scatter plot diagram with points, we can see that the trend is going downwards while going from left to right. It means that the Home_goals were scored more than the Away_goal.

### CORRELATION IN SCATTER PLOT IN R

```
#finding correlation
cor(x=Results$Home_goal, y=Results$Away_goals)
```

```
## [1] -0.1300056
```

It looks like there is a moderately strong relationship in the scatter plot which is true for the correlation cause it matches to what we see in the plot.

### FREQUENCY

In statistics, frequency or absolute frequency indicates the number of occurrences of a data value or the number of times a data value occurs. These frequencies are often plotted on bar graphs or histograms to compare the data values.

### ABSOLUTE FREQUENCY

Absolute frequency shows the number of times the value is repeated in the data vector.

```
#frequency table
table(Results$Away_goals)
```

```
##
##   0   1   2   3   4   5   6
## 136 127  65  33  16   2   1
```

There has been 136 cases when the Away teams have failed to score and open up their account. They have managed to score 6 goals (max) just once and 5 goals for a couple of occasions. Also, the Away teams have failed to score the maximum value of the dataset i.e. 7 (goals).

### RELATIVE FREQUENCY

In R language, table() function and length of data vector is used together to find relative frequency of data vector.

```
table(Results$Away_goals)/length(Results$Away_goals)
```

```
##
##           0           1           2           3           4           5
## 0.357894737 0.334210526 0.171052632 0.086842105 0.042105263 0.005263158
##           6
## 0.002631579
```

### TWO-WAY TABLE IN R

A two-way table is a table that describes two categorical data variables together, and R gives you a whole toolset to work with two-way tables. They contain the number of cases for each combination of the categories in both variables. The analysis of categorical data always starts with tables. But first, we have to create the tables.

```
#table function-two way table
table(Results$Home_team, Results$Result)
```

```
##
##                            A  D  H
##   AFC Bournemouth           7  5  7
##   Arsenal                   2  2 15
##   Brighton and Hove Albion  4  8  7
##   Burnley                   7  5  7
##   Chelsea                   4  4 11
##   Crystal Palace            7  5  7
##   Everton                   5  4 10
##   Huddersfield Town         8  5  6
##   Leicester City            6  6  7
##   Liverpool                 0  7 12
##   Manchester City           1  2 16
##   Manchester United         2  2 15
##   Newcastle United          7  4  8
##   Southampton               8  7  4
##   Stoke City                9  5  5
##   Swansea City             10  3  6
##   Tottenham Hotspur         2  4 13
##   Watford                   6  6  7
##   West Bromwich Albion      7  9  3
##   West Ham United           6  6  7
```

From this two-way frequency table, *Manchester City, Manchester United, Arsenal, Tottenham HotSpur, Liverpool, Chelsea* are mostly successful as Home team with over 10 wins playing as a host on their stadium. On the other hand, *WestBromwich Albion, Stoke City, Southampton* seems to be the unsuccessful teams on their home ocassions failing to win over 5 times in total of 19 occasions.

```
#frequency table
table(Results$Home_goal)
```

```
##
##   0   1   2   3   4   5   7
##  90 126  91  35  23  14   1
```

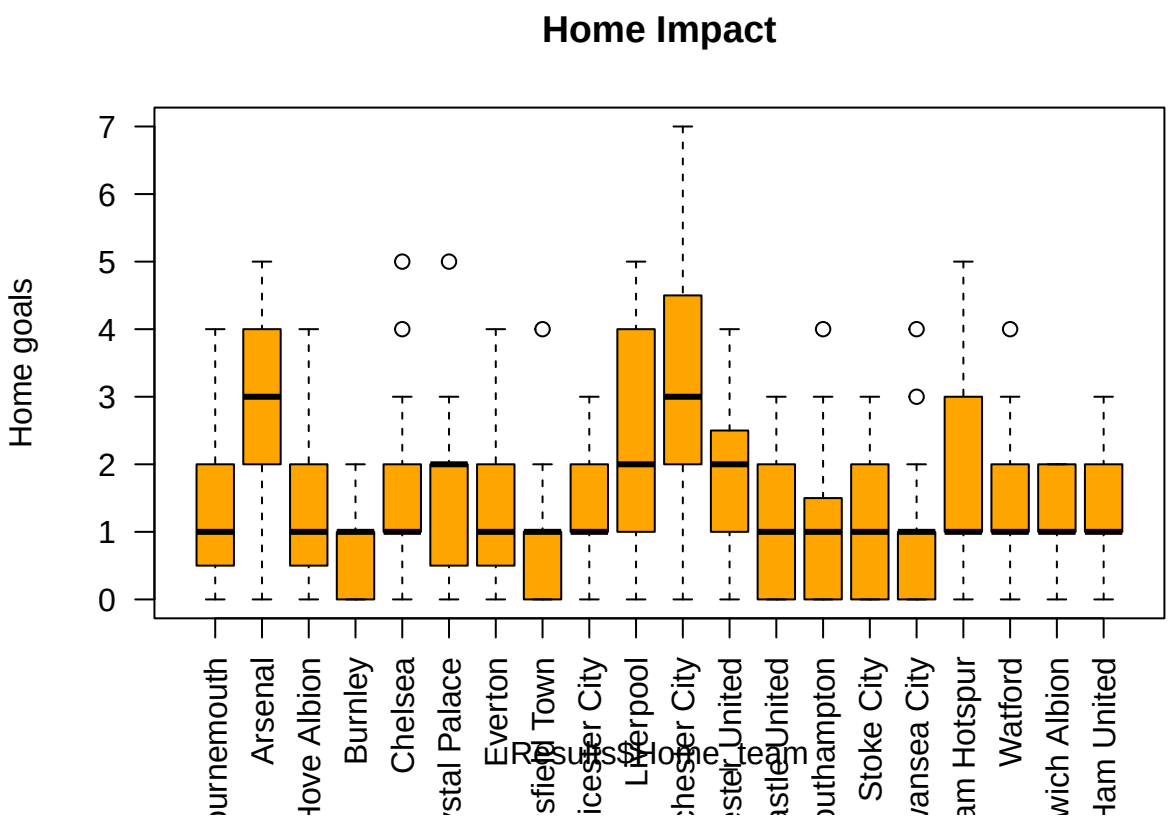
There has been 90 cases when the Home teams have failed to score and open up their account. They have managed to score 5 goals 14 times and 4 goals for 23 occasions. Also, the Home teams have succeed to score

the maximum value of the dataset i.e. 7 (goals). Thus, Home teams seems to have more goal scores on their account than Away teams.

### PARALLEL BOX PLOT

I have chosen side by side box plot *( also known as parallel box or comparative box plot)* to visually display comparing the levels of one categorical variable by means of a quantative variable.

```
#side by side plot
boxplot(Results$Home_goal~Results$Home_team, col="Orange", main="Home Impact", ylab="Home goals", las=2)
```




As we can see from the graph, *Manchester City* fascinates everyone's eyes in the middle of the side by side plot. Its median is pointing at 3, which is slightly on the left from it's center and three times better than that of the dataset. It's tail at the top ranges to point 7(max value of dataset) on the y-axis which indicates it's maximum value. Also, the sight of the plot indicates that it has been a solid and consistent team throughout the season (2017-18).

Other than that *Manchester United, Liverpool, Chelsea* have median ranging over 2. And looks balanced according to the plot.

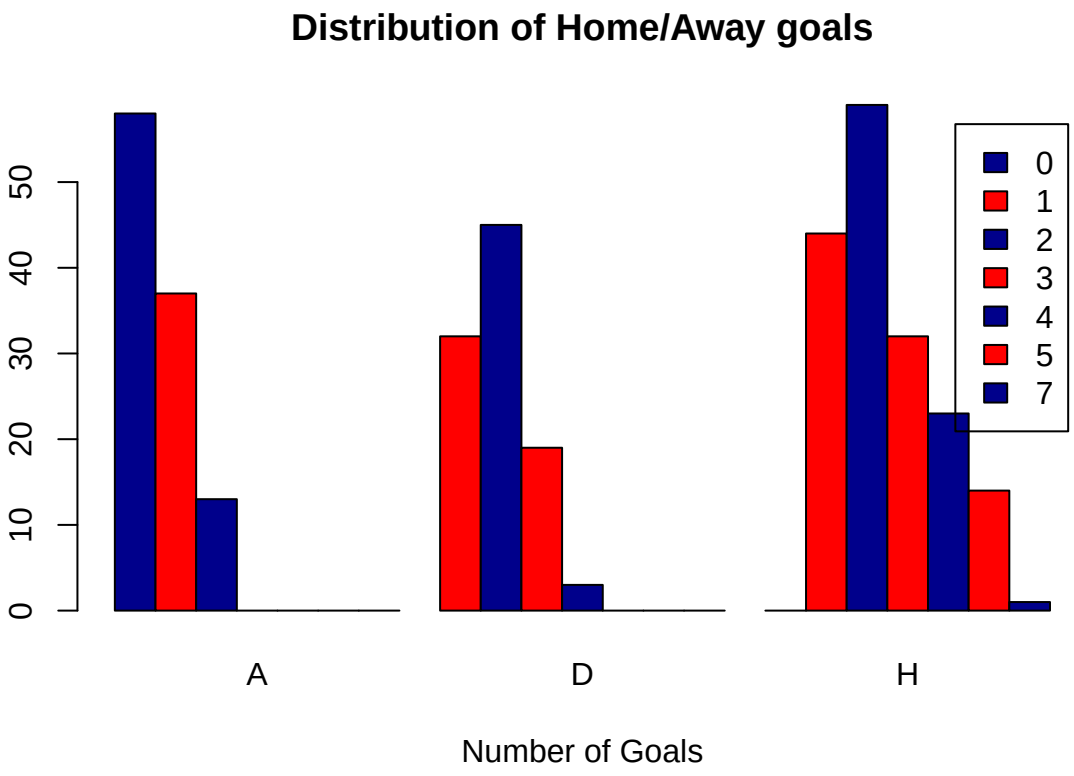
*Chelsea and Swansea City* have two outliers each and *Crystal Palace, Southampton, Watford, Huddersfield* have one each. They could be either on the scoring or conceding side of significant amount of goals in the season.

*Burnley, Swansea City, Stoke City, WestBromich Albion* are the teams that look poor on the basis of their home performances. These teams failed to take the home advantage and make home impact as per the plot.

### VISUALIZATION AND STATISTICAL COMPUTATION

For visualization, I have chosen Bar graph to compare the amount of home goals and its impact on the result over the season. On my x-axis, I have the results and its frequency on the y-axis.

```
#bar-graph
counts <- table(Results$Home_goal, Results$Result)
barplot(counts, main="Distribution of Home/Away goals",
  xlab="Number of Goals", col=c("darkblue","red"),
  legend = rownames(counts), beside=TRUE)
```




Here in the bar-graph;

A = Away_team (it means the case when Away_team is the winner)

H = Home_team (it means the case when Home_team is the winner)

D= Draw (its the case when Away_team and Home_team scored same number of goals in the match)

We can clearly see from the graph that Home teams have been able to produce more productive results than the away teams. The case when Away teams have won has lesser no. of blocks in the graph than the Home team. The Home teams have produced more no. of goals which has been closer or above the mean score of the entire season.

#Statistical Computation

```
#difference=Homegoals - Awaygoals
sum(Results$Home_goal)
```

```
## [1] 582
```

```
sum(Results$Away_goals)
```

```
## [1] 436
```

The number of Home goals is 146 more than that of the Away goals. This shows the dominance of the team hosting on their stadium.

**DATA ANALYSIS**

In this project I have included a dataset of results of 380 Premier League matches in a single season of the year 2017/2018. It can be found in (https://www.premierleague.com/results.)

Here in this dataset, I have "Home_team", "Away_team", "Season", "Results" as four categorical (qualitative) variable. "Home_goal" and "Away_goals" are the two quantitative vaiable which have been up and down in the progress of the season.

When looking for datasets, the main things I took into consideration are the usefulness, quality, and format of the data. For this project, I have pulled data from (https://www.premierleague.com/results.) as it is a well-established source and has the necessary stats in a relatively well-organized format, which will only require some light cleaning. Since, I only had a total of 6 variabes out of which 5 have been used to compare and relate with each other, I did not have to bother cleaning much as the other one variable is the year of the season (2017-2018) which remain unchanged and unaffected within the dataset.

**CONCLUSION**

By doing exploratory data analysis, I have discovered that a winning team tends to be successful at both the venues. There are many other features and relations that we can explore. For example, it would be interesting to see how the game evolves. There have been many cases when the teams playing at away venues have failed but made up at their home venues. "Burnley" seems to be the only outlier team in this case whose away results are vastly better than its home results. Overall, it is supposed to be a successful franchise with it's rank secured within top 10. All information generated can help us to build up machine learning models to predict the results of football games.