rsoccer_markdown

Akhila

01/05/2019

R Markdown

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

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

Including Plots

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot. You can also embed plots, for example:

```
## Installing package into 'C:/Users/AKMANI/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## package 'pscl' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\AKMANI\AppData\Local\Temp\RtmpyugJTc\downloaded_packages
```

```
## Installing package into 'C:/Users/AKMANI/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## package 'RSQLite' successfully unpacked and MD5 sums checked
##
```

```
## The downloaded binary packages are in
##  C:\Users\AKMANI\AppData\Local\Temp\RtmpyugJTc\downloaded_packages
```

```
## Installing package into 'C:/Users/AKMANI/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## package 'stringr' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\AKMANI\AppData\Local\Temp\RtmpyugJTc\downloaded_packages
```

```
## Installing package into 'C:/Users/AKMANI/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## package 'plyr' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\AKMANI\AppData\Local\Temp\RtmpyugJTc\downloaded_packages
```

```
## Installing package into 'C:/Users/AKMANI/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## package 'dplyr' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\AKMANI\AppData\Local\Temp\RtmpyugJTc\downloaded_packages
```

```
## Installing package into 'C:/Users/AKMANI/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
## package 'ggplot2' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\AKMANI\AppData\Local\Temp\RtmpyugJTc\downloaded_packages
```

```
## Installing package into 'C:/Users/AKMANI/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## package 'gplots' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\AKMANI\AppData\Local\Temp\RtmpyugJTc\downloaded_packages
```

```
## Installing package into 'C:/Users/AKMANI/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## package 'DBI' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\AKMANI\AppData\Local\Temp\RtmpyugJTc\downloaded_packages
```

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

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

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
## Registered S3 methods overwritten by 'ggplot2':
##   method         from 
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
## 
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
## 
##     lowess
```

```
## Downloading package from url: https://cran.r-project.org/src/contrib/Archive/RSQLite/RSQLite_2.1.0.tar.gz
```

```
## 
## 


   checking for file 'C:\Users\AKMANI\AppData\Local\Temp\RtmpyugJTc\remotes96001fefdb7\RSQLite/DESCRIPTION' ...

   checking for file 'C:\Users\AKMANI\AppData\Local\Temp\RtmpyugJTc\remotes96001fefdb7\RSQLite/DESCRIPTION' ...

v  checking for file 'C:\Users\AKMANI\AppData\Local\Temp\RtmpyugJTc\remotes96001fefdb7\RSQLite/DESCRIPTION' (895m
s)
## 



-  preparing 'RSQLite': (1.3s)
## 

   checking DESCRIPTION meta-information ...
   checking DESCRIPTION meta-information ...

v  checking DESCRIPTION meta-information
##

-  cleaning src
##




   checking vignette meta-information ...

   checking vignette meta-information ...

v  checking vignette meta-information (838ms)
##




-  checking for LF line-endings in source and make files and shell scripts (655ms)
##




-  checking for empty or unneeded directories (1.5s)
##




-  building 'RSQLite_2.1.0.tar.gz'
##


##
```

```
## Installing package into 'C:/Users/AKMANI/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

1. The first leagues of Spain, England, Germany and Italy are considered the four most

attractive football leagues in Europe.

a. In which of the four leagues do on average score the most or the fewest goals

per game

```
    country <- country %>% rename(country_id=id)
    country_match <-merge(country,match,by=c("country_id"),all=TRUE)

   match_output1 <- country_match %>%
    group_by(league_id,match_api_id)%>%
       filter(str_detect(name, "Spain") | str_detect(name, "England") | str_detect(name, "Germany") | str_detect
(name,              "Italy"))%>%
       select(league_id,match_api_id,home_team_goal,away_team_goal,name)%>%
       mutate(goal_score = home_team_goal+away_team_goal)%>%
    ungroup()%>%
      group_by(league_id,name)%>%
      summarise(average_score=mean(goal_score),sum=sum(goal_score),n=n())%>%
      arrange(desc(average_score))

   head(data.frame(match_output1))
```

```
##   league_id    name average_score  sum    n
## 1      7809 Germany      2.901552 7103 2448
## 2     21518   Spain      2.767105 8412 3040
## 3      1729 England      2.710526 8240 3040
## 4     10257   Italy      2.616838 7895 3017
```

b. Compare the average, median, standard deviation, variance, range and

interquartile distance of goals scored per match between the four most

attractive European leagues and the remaining leagues.

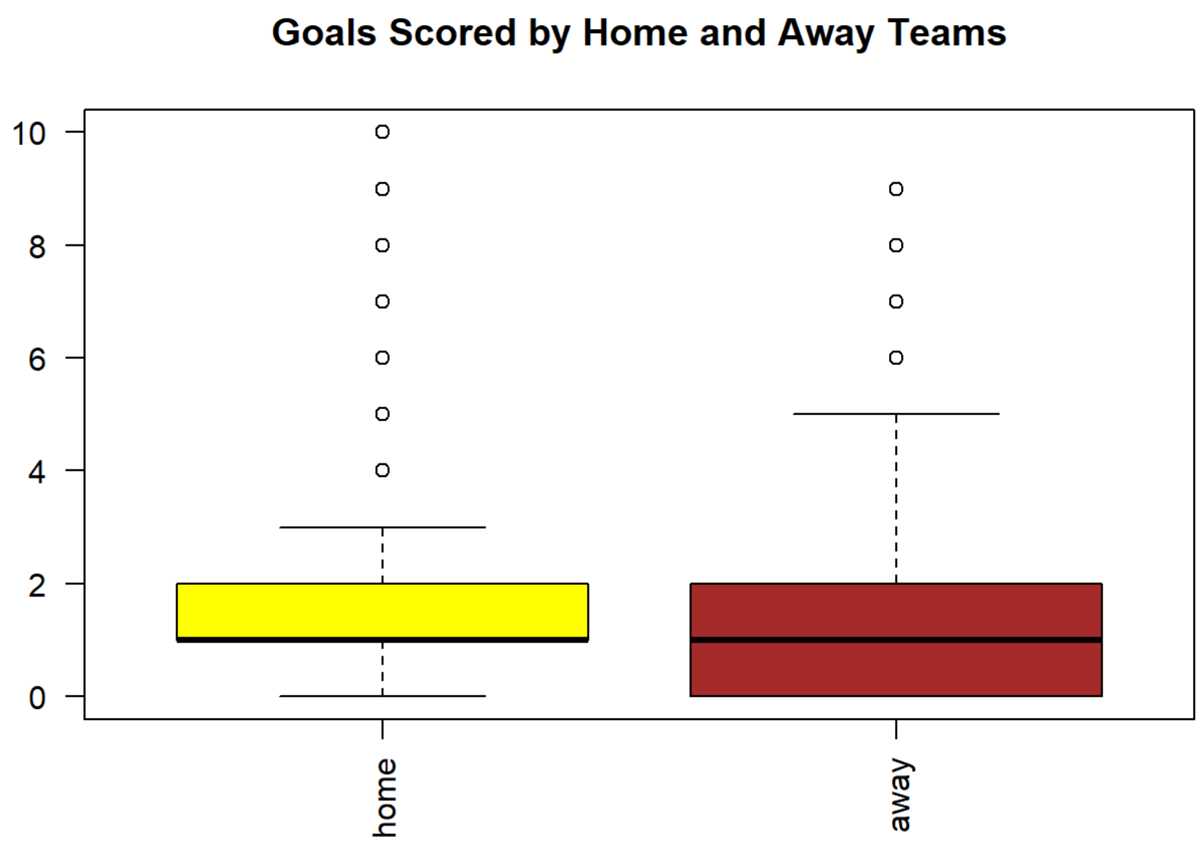
```
    match_output1 <- country_match %>%
     mutate(name =  ifelse(name %in% c("Spain","England","Germany","Italy") , "top league",
                            "Others"))

   statistical_data_match <- match_output1 %>%
     group_by(league_id,match_api_id)%>%
     select(league_id,match_api_id,home_team_goal,away_team_goal,name)%>%
     mutate(goal_score = home_team_goal+away_team_goal)%>%
     ungroup()%>%
     group_by(name)%>%
     summarise(average_score=mean(goal_score),median(goal_score),
               sd(goal_score),var(goal_score),range=max(goal_score)-min(goal_score),
               IQR(goal_score),sum=sum(goal_score),n=n())

    statistical_data_match %>%
      select(1:9)
```

```
## # A tibble: 2 x 9
##   name  average_score `median(goal_sc~ `sd(goal_score)` `var(goal_score~
##   <chr>         <dbl>            <dbl>            <dbl>            <dbl>
## 1 Othe~          2.68                2             1.65             2.74
## 2 top ~          2.74                3             1.69             2.87
## # ... with 4 more variables: range <int>, `IQR(goal_score)` <dbl>,
## #   sum <int>, n <int>
```

2. Is there really a home advantage? Use a box plot to show the number of goals scored

by home and away teams.

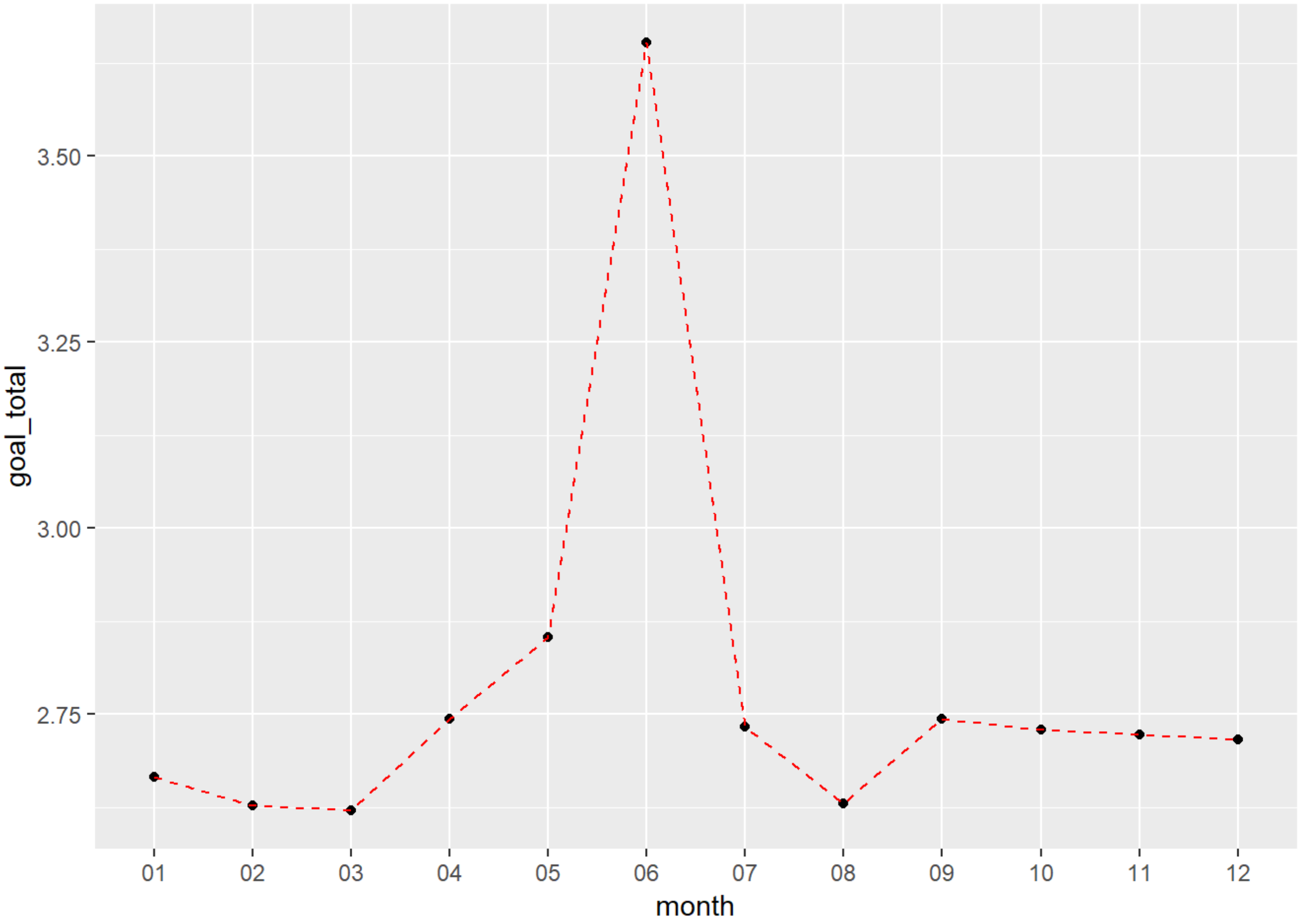
```
goals_home_away <- match %>%
     summarise(home_team_goals = sum(home_team_goal),away_team_goals = sum(away_team_goal))

 home <- country_match$home_team_goal
   away <- country_match$away_team_goal
   boxplot(home,away,
           main = "Goals Scored by Home and Away Teams",
           at = c(1,2),
           names = c("home", "away"),
           las = 2,
           col = c("yellow","brown"),
           border = "black",
           vertical = FALSE,
           notch = FALSE)
```



# 3. “All soccer players are fair-weather players!” Check the assertion with a line chart: Do # on average more goals fall per game in the summer months than in the rest of the # year?

```
match_addmonth <- match %>%
     select(date,match_api_id,home_team_goal,away_team_goal) %>%
     mutate(goal_total=home_team_goal+away_team_goal,month = date)

   match_addmonth$month <- format(as.Date(match_addmonth$month), "%m")

   ggploting <- ggplot(match_addmonth,aes(x=month, y=goal_total))
```

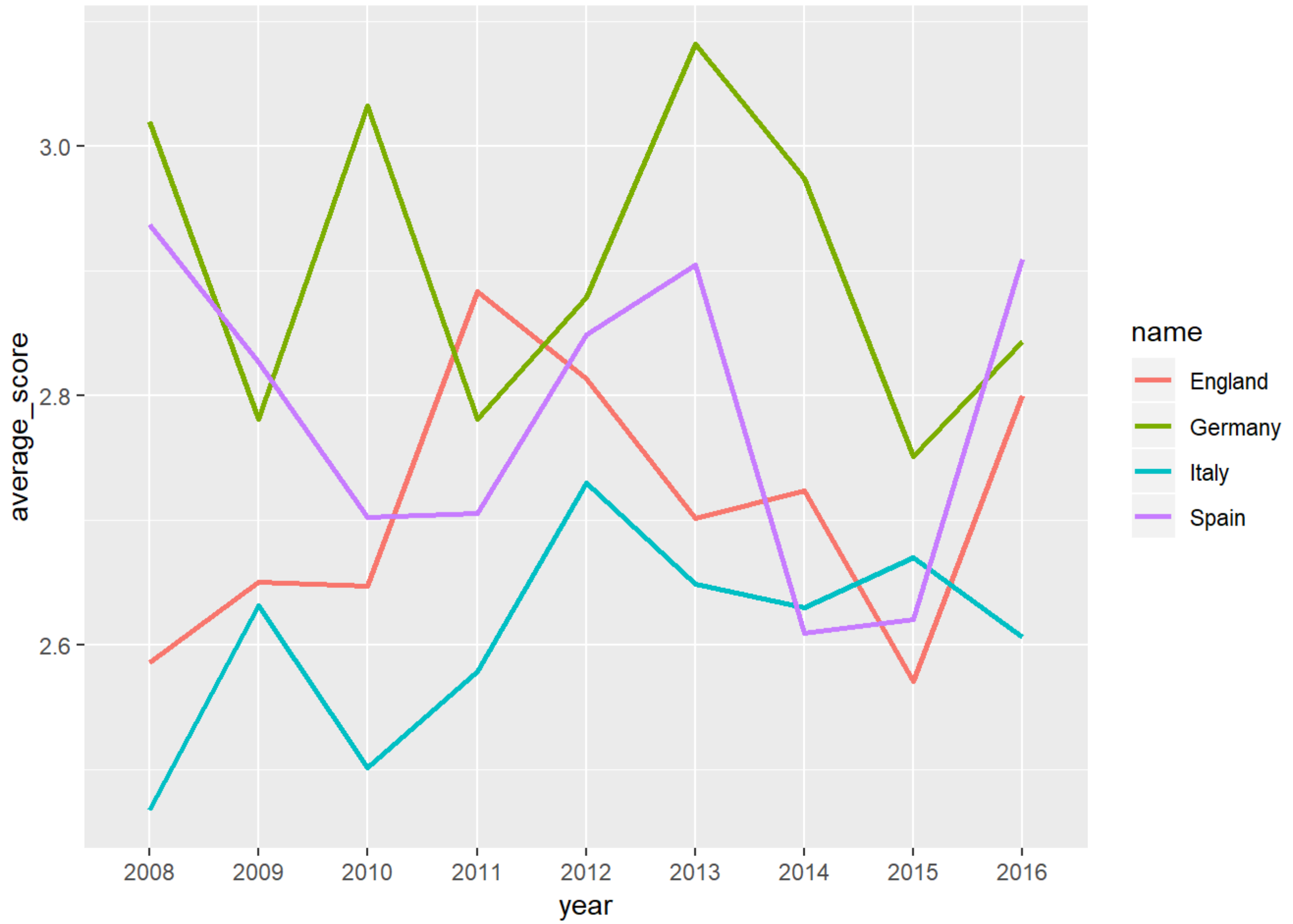
```
ggploting+
  stat_summary(fun.y=mean,geom="point")+
  stat_summary(fun.y=mean,geom="line",aes(group=1),color="red",linetype="dashed")
```



```
#match_addmonth <- match_addmonth %>%
# mutate(seasondiff =  ifelse(month %in% c("06","07","08") , "Summer",
#                              ifelse(month %in% c("01","02","03","04","05"), "Others","Others")))
#attach(match_addmonth)
#plotmeans(goal_total ~ seasondiff, data = match_addmonth)
```

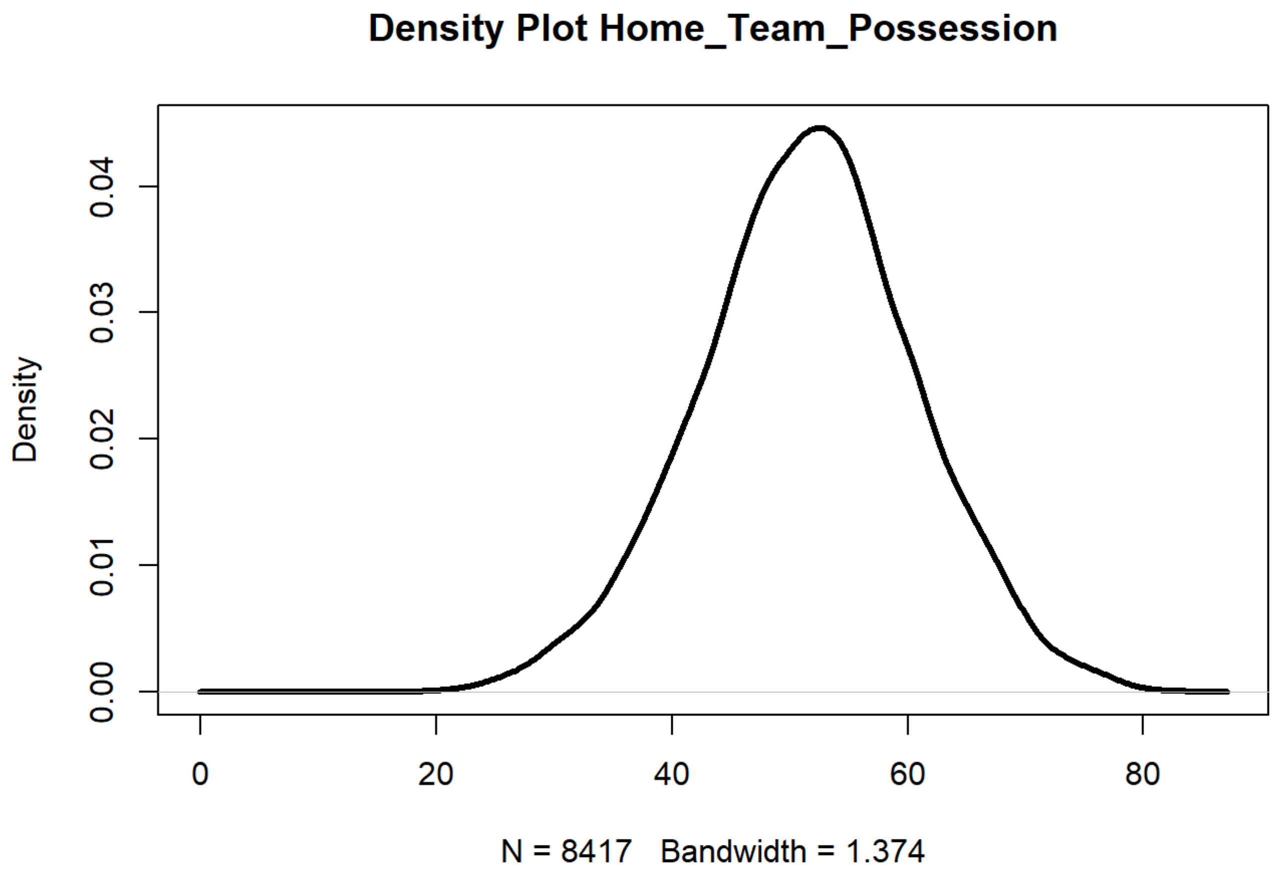
4. Display the average goals scored per game for the top 4 leagues per year from 2008 to 2016

```
match_addyear <- country_match %>%
  select(league_id,name,match_api_id,date,home_team_goal,away_team_goal) %>%
  mutate(goal_total=home_team_goal+away_team_goal,year = date)

match_addyear$year <- format(as.Date( match_addyear$year), "%Y")

match_byyear <- match_addyear %>%
  group_by(league_id,match_api_id)%>%
  filter(str_detect(name, "Spain")|str_detect(name,"England")|str_detect(name,"Germany")|str_detect(name,"Italy"))%>%
  select(league_id,match_api_id,home_team_goal,away_team_goal,name,date,year)%>%
  mutate(goal_score = home_team_goal+away_team_goal)%>%
  ungroup()%>%
  group_by(name,year)%>%
  summarise(average_score=mean(goal_score),sum=sum(goal_score),n=n())%>%
  arrange(desc(average_score))

ggplot(match_byyear, aes(x=year, y=average_score,group=name, color=name)) + geom_line(size=1)
```

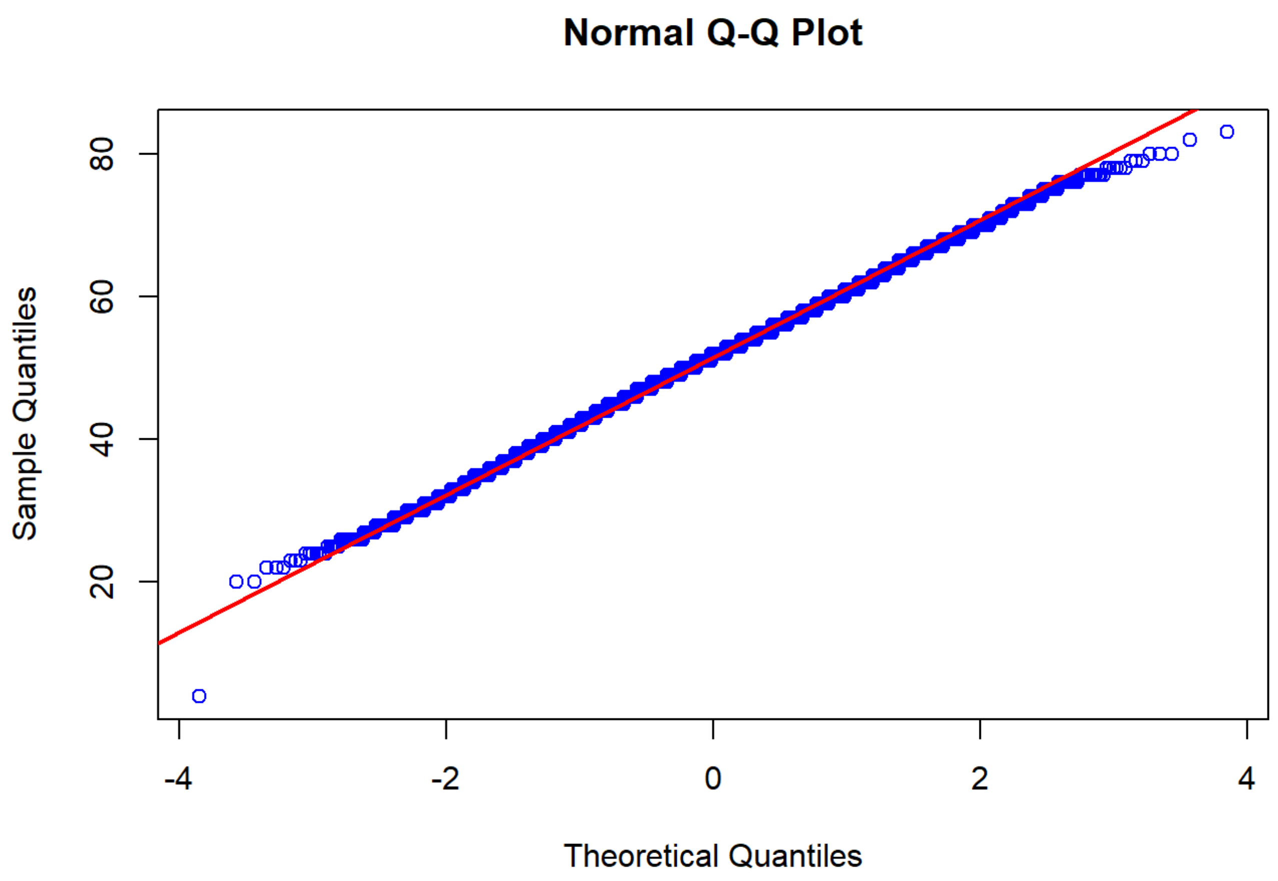


# 5. Use an estimated density function curve AND a QQ-plot to check whether the # home_team_possession variable is (approximately) normally distributed.

```
plot(density(na.omit(match$home_team_possession)),main ="Density Plot Home_Team_Possession",lwd =3 )
```



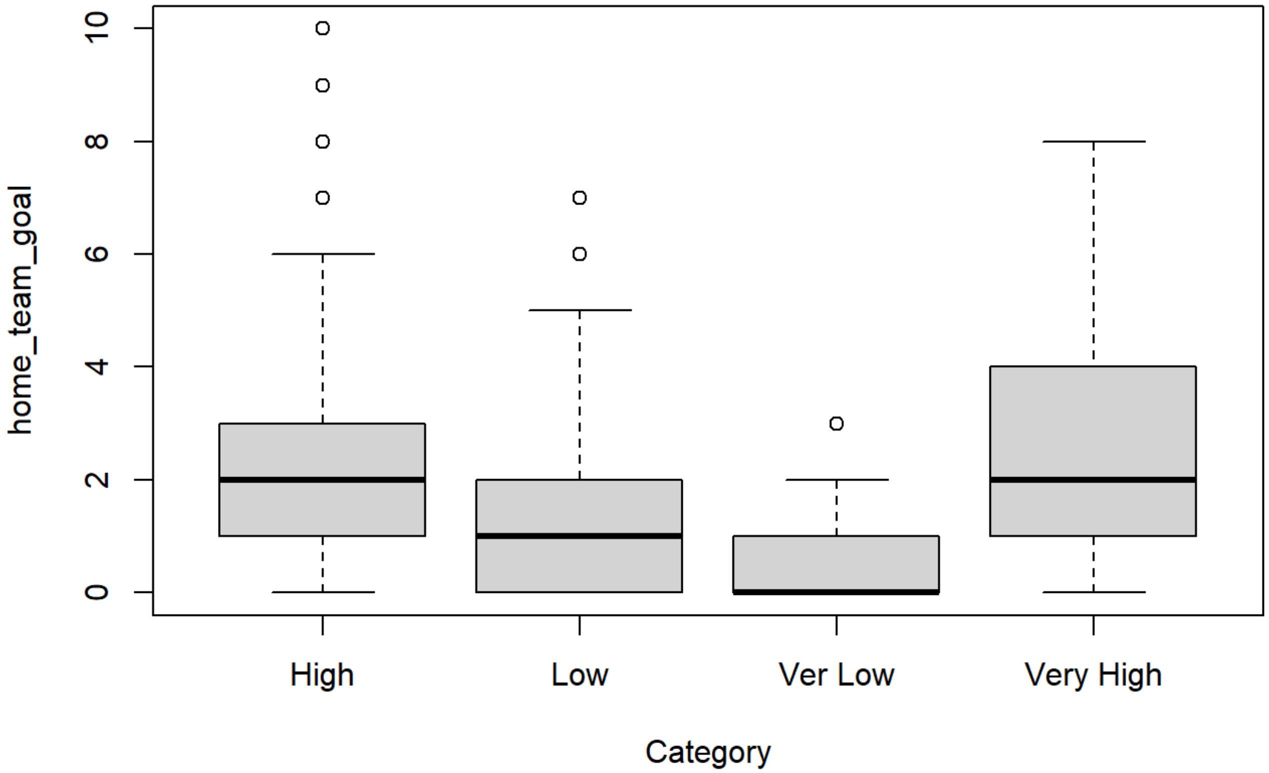
```
{qqnorm(match$home_team_possession,col ="blue")
qqline(match$home_team_possession,col ="red",lwd =2)}

#log to get closer to the line to see the distribution clearly
log.home_team_possession <-match$home_team_possession
{qqnorm(log.home_team_possession,col ="blue")
qqline(match$home_team_possession,col ="red",lwd =2)}
```



# 6. Use a box plot to show whether there is a correlation between ball ownership # (home_team_possession) and the number of goals (home_team_goals) scored per # game for home teams. Create four categories of ball ownership shares: very low # (≤25%), low (25% < x ≤50%), high (50% < x ≤75%), and very high (x > 75%).

```
match_corr <- match %>%
     select(home_team_possession,home_team_goal)
 match_corr <- subset(match_corr, home_team_possession != "NA")

 match_corr$home_team_possession <- as.numeric(match_corr$home_team_possession)
```

```
match_corr <- match_corr %>%
  mutate(category =  ifelse(home_team_possession %in% c(0:25) , "Ver Low",
                              ifelse(home_team_possession %in% c(26:50), "Low",
                                     ifelse(home_team_possession %in% c(51:75), "High","Very High"))))
attach(match_corr)
boxplot(home_team_goal~factor(category),xlab = "Category", data = match_corr, col = "lightgray")
```


home_team_goal
High
Low
Ver Low
Very High
Category