Sports Analytics-Real Madrid

Sai Abhilash Ghanta

4/29/2016

Step 1: Collecting the data:

The data is collected from http://football-data.co.uk/spainm.php where three seasons data is selected. The data gives us information about the primary division football (Soccer) match data. There are around 1098 instances and 24 attributes and the attribute names are in short form. The full names of the attribute names can be found at http://football-data.co.uk/notes.txt.

Step 2: Pre-processing:

The three seasons data are merged into one excel file and unnecessary attributes were removed. Because of anomalies in the date attribute of matches, I needed to change them to month name using the following steps.

- Used Replace function to remove 0 from month numbers. (03-3,09-9)
- Used =MID(CELL,FIND("/",CELL)+1, FIND("/",CELL, FIND("/",CELL)+1)-FIND("/",CELL)-1) to get month number.
- Copied all Month values and changed to Numeric value.
- Then used =TEXT(CELL, "MMM") on all rows in column to change it to month.

After changing to months, the file is saved in csv format.

Step 3: Loading in Rstudio:

Loaded the CSV file in Rstudio using readcsv format.

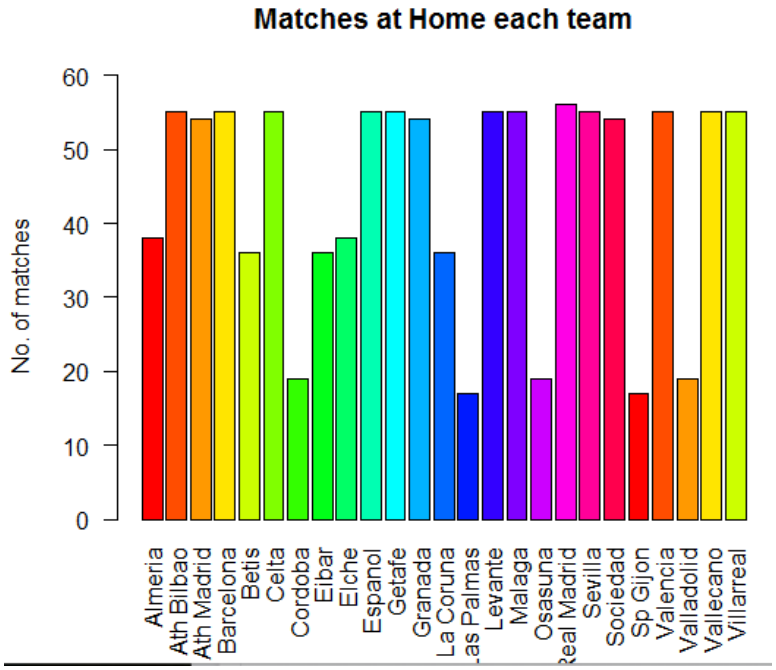
```
> setwd("~/CS544/My/Term Project")
> Laliga <- read.csv("Laliga.csv")
```

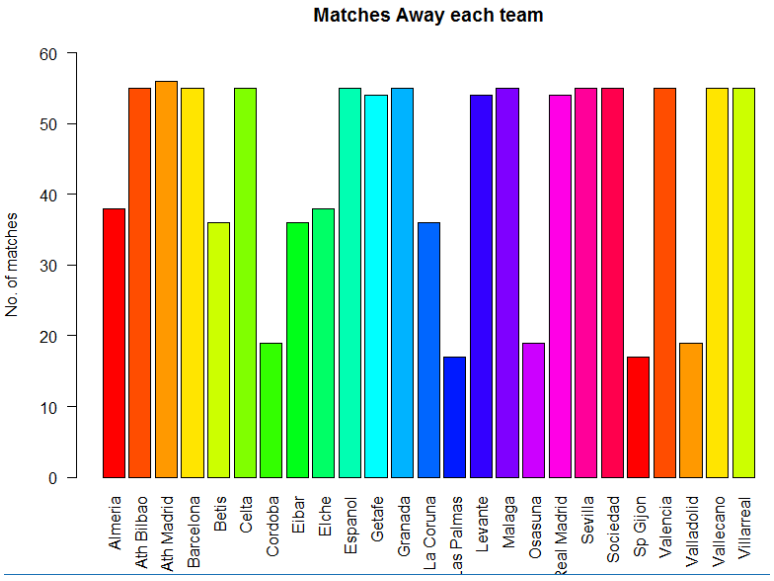
```
> summary(Laliga)
     Month           HomeTeam         AwayTeam        FTHG            FTAG        FTR
 Apr    :138   Real Madrid: 56   Ath Madrid: 56   Min.   : 0.000   Min.   :0.000   A:327
 Jan    :135   Ath Bilbao : 55   Ath Bilbao: 55   1st Qu.: 1.000   1st Qu.:0.000   D:262
 Mar    :134   Barcelona  : 55   Barcelona : 55   Median : 1.000   Median :1.000   H:509
 Feb    :123   Celta      : 55   Celta     : 55   Mean   : 1.587   Mean   :1.121
 Sep    :121   Espanol    : 55   Espanol   : 55   3rd Qu.: 2.000   3rd Qu.:2.000
 Nov    :108   Getafe     : 55   Granada   : 55   Max.   :10.000   Max.   :8.000
 (Other):339   (Other)    :767   (Other)   :767
      HTHG            HTAG            HTR           HS              AS              HST
 Min.   :0.0000   Min.   :0.0000   A:266   Min.   : 2.00   Min.   : 0.00   Min.   : 0.000
 1st Qu.:0.0000   1st Qu.:0.0000   D:426   1st Qu.:10.00   1st Qu.: 7.00   1st Qu.: 3.000
 Median :1.0000   Median :0.0000   H:406   Median :13.00   Median :10.00   Median : 4.000
 Mean   :0.7213   Mean   :0.4964           Mean   :13.54   Mean   :10.77   Mean   : 4.832
 3rd Qu.:1.0000   3rd Qu.:1.0000           3rd Qu.:16.00   3rd Qu.:14.00   3rd Qu.: 6.000
 Max.   :6.0000   Max.   :4.0000           Max.   :32.00   Max.   :35.00   Max.   :15.000

      AST              HF              AF              HC              AC              HY
 Min.   : 0.000   Min.   : 4.00   Min.   : 2.00   Min.   : 0.000   Min.   : 0.000   Min.   :0.000
 1st Qu.: 2.000   1st Qu.:11.00   1st Qu.:11.00   1st Qu.: 4.000   1st Qu.: 3.000   1st Qu.:1.000
 Median : 3.000   Median :14.00   Median :14.00   Median : 6.000   Median : 4.000   Median :2.000
 Mean   : 3.741   Mean   :13.97   Mean   :13.97   Mean   : 5.945   Mean   : 4.569   Mean   :2.443
 3rd Qu.: 5.000   3rd Qu.:17.00   3rd Qu.:17.00   3rd Qu.: 8.000   3rd Qu.: 6.000   3rd Qu.:3.000
 Max.   :15.000   Max.   :33.00   Max.   :29.00   Max.   :20.000   Max.   :17.000   Max.   :8.000

      AY              HR              AR             B365H           B365D           B365A
 Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   : 1.040   Min.   : 2.500   Min.   : 1.080
 1st Qu.:2.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 1.570   1st Qu.: 3.300   1st Qu.: 2.500
 Median :3.000   Median :0.0000   Median :0.0000   Median : 2.100   Median : 3.500   Median : 3.600
 Mean   :2.718   Mean   :0.1284   Mean   :0.1566   Mean   : 2.928   Mean   : 4.431   Mean   : 5.696
 3rd Qu.:4.000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.: 2.900   3rd Qu.: 4.500   3rd Qu.: 6.000
 Max.   :7.000   Max.   :2.0000   Max.   :3.0000   Max.   :26.000   Max.   :17.000   Max.   :41.000
```

Step 4: Analysing the Data:

- Performed Categorical analysis on LaLiga teams.

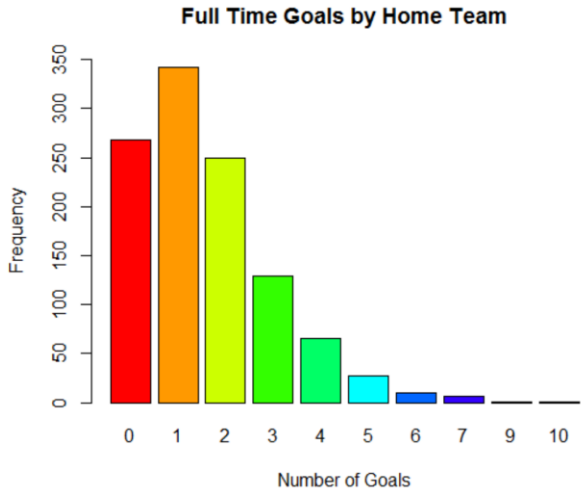
```
> #Graphical Rep of categorical data
> barplot(table(Laliga$HomeTeam),ylim = c(0,60),
+ main = "Matches at Home each team",
+ ylab = "No. of matches", col = rainbow(20),las=2)
> barplot(table(Laliga$AwayTeam),ylim = c(0,60),
+ main = "Matches Away each team",
+ ylab = "No. of matches", col = rainbow(20),las=2)
```

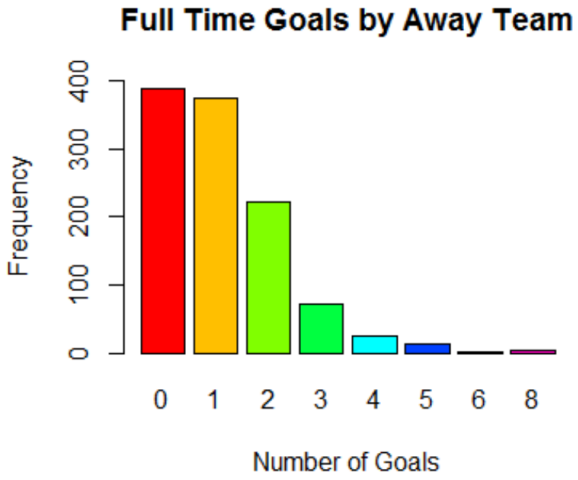




- Performed Numerical analysis on number of goals by LaLiga Teams at home and away from Home.

```
> table(Laliga$FTHG)

  0   1   2   3   4   5   6   7   9  10
268 342 249 129  65  27  10   6   1   1
> barplot(table(Laliga$FTHG),main = "Full Time Goals by Home Team",
+ xlab = "Number of Goals",
+ ylab = "Frequency",ylim =c(0,350),col = rainbow(10),las=2)
> table(Laliga$FTAG)

  0   1   2   3   4   5   6   8
388 374 222  72  24  13   2   3
> barplot(table(Laliga$FTAG),main = "Full Time Goals by Away Team",
+ xlab = "Number of Goals",
+ ylab = "Frequency",ylim =c(0,400),col = rainbow(8))
```





- Real Madrid and FC Barcelona will be the teams on which analysis will be carried out.

```
> #Real Madrid and Barcelona Data
> RealMadrid <- subset(Laliga, Laliga$HomeTeam=="Real Madrid"
+ | Laliga$AwayTeam=="Real Madrid")
> Barcelona <- subset(Laliga, Laliga$HomeTeam=="Barcelona"
+ | Laliga$AwayTeam=="Barcelona")
> #Categorical data table for Real Madrid
> table(RealMadrid$HomeTeam)

    Almeria  Ath Bilbao  Ath Madrid   Barcelona       Betis       Celta     Cordoba       Eibar
          2           3           3           3           2           3           1           2
      Elche     Espanol      Getafe     Granada   La Coruna  Las Palmas     Levante      Malaga
          2           3           3           3           1           1           3           3
    Osasuna Real Madrid     Sevilla    Sociedad    Sp Gijon    Valencia  Valladolid   Vallecano
          1          56           3           2           1           3           1           2
 Villarreal
          3
> table(RealMadrid$HomeTeam)/length(RealMadrid$HomeTeam)

    Almeria  Ath Bilbao  Ath Madrid   Barcelona       Betis       Celta     Cordoba       Eibar
0.018181818 0.027272727 0.027272727 0.027272727 0.018181818 0.027272727 0.009090909 0.018181818
      Elche     Espanol      Getafe     Granada   La Coruna  Las Palmas     Levante      Malaga
0.018181818 0.027272727 0.027272727 0.027272727 0.009090909 0.009090909 0.027272727 0.027272727
    Osasuna Real Madrid     Sevilla    Sociedad    Sp Gijon    Valencia  Valladolid   Vallecano
0.009090909 0.509090909 0.027272727 0.018181818 0.009090909 0.027272727 0.009090909 0.018181818
 Villarreal
0.027272727
```

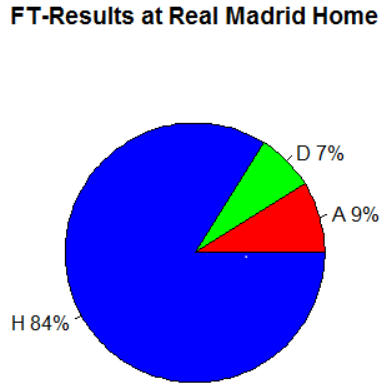
Filter

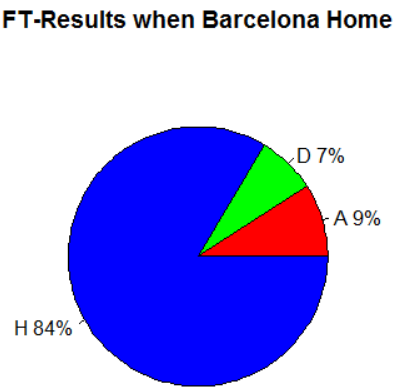
	Month	HomeTeam	AwayTeam	FTHG
6	Aug	Real Madrid	Betis	
20	Aug	Granada	Real Madrid	
27	Aug	Real Madrid	Ath Bilbao	
34	Sep	Villarreal	Real Madrid	
48	Sep	Real Madrid	Getafe	
55	Sep	Elche	Real Madrid	
63	Sep	Real Madrid	Ath Madrid	
77	Oct	Levante	Real Madrid	
83	Oct	Real Madrid	Malaga	
92	Oct	Barcelona	Real Madrid	

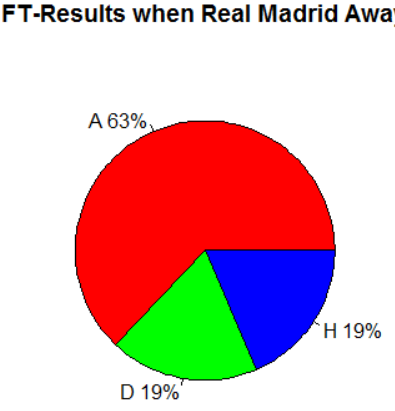
Filter

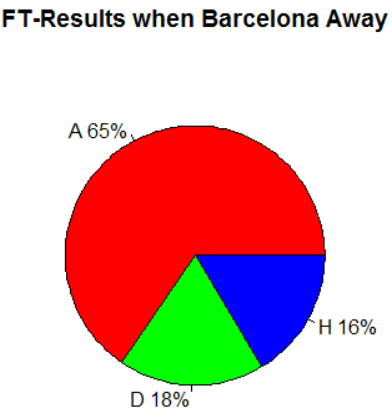
	Month	HomeTeam	AwayTeam	FTHG	FTA
4	Aug	Barcelona	Levante	7	
19	Aug	Malaga	Barcelona	0	
30	Sep	Valencia	Barcelona	2	
32	Sep	Barcelona	Sevilla	3	
45	Sep	Vallecano	Barcelona	0	
52	Sep	Barcelona	Sociedad	4	
62	Sep	Almeria	Barcelona	0	
75	Oct	Barcelona	Valladolid	4	
82	Oct	Osasuna	Barcelona	0	
92	Oct	Barcelona	Real Madrid	2	

```
> #Pie charts
> par(mfrow=c(2,2))
> HomeResults.RM <- table(RealMadrid$FTR[RealMadrid$HomeTeam=="Real Madrid"])
> AwayResults.RM <- table(RealMadrid$FTR[RealMadrid$AwayTeam=="Real Madrid"])
> HomeResults.FCB <- table(Barcelona$FTR[Barcelona$HomeTeam=="Barcelona"])
> AwayResults.FCB <- table(Barcelona$FTR[Barcelona$AwayTeam=="Barcelona"])
> result.labels <- names(table(RealMadrid$FTR))
> results.percents <- round(HomeResults.RM/sum(HomeResults.RM)*100)
> labl <- paste(result.labels,results.percents)
> labl <- paste(labl, "%", sep = "")
> pie(table(RealMadrid$FTR[RealMadrid$HomeTeam=="Real Madrid"]),
+ labels = labl, col = rainbow(3), main = "FT-Results at Real Madrid Home")
> results.percents <- round(AwayResults.RM/sum(AwayResults.RM)*100)
> labl <- paste(result.labels,results.percents)
> labl <- paste(labl, "%", sep = "")
> pie(AwayResults.RM,
+ labels = labl, col = rainbow(3), main = "FT-Results when Real Madrid Away")
> results.percents <- round(HomeResults.FCB/sum(HomeResults.FCB)*100)
> labl <- paste(result.labels,results.percents)
> labl <- paste(labl, "%", sep = "")
> pie(HomeResults.FCB,labels = labl, col = rainbow(3), main = "FT-Results when Barcelona Home")
> results.percents <- round(AwayResults.FCB/sum(AwayResults.FCB)*100)
> labl <- paste(result.labels,results.percents)
> labl <- paste(labl, "%", sep = "")
> pie(AwayResults.FCB,labels = labl, col = rainbow(3), main = "FT-Results when Barcelona Away")
> par(mfrow=c(1,1))
```



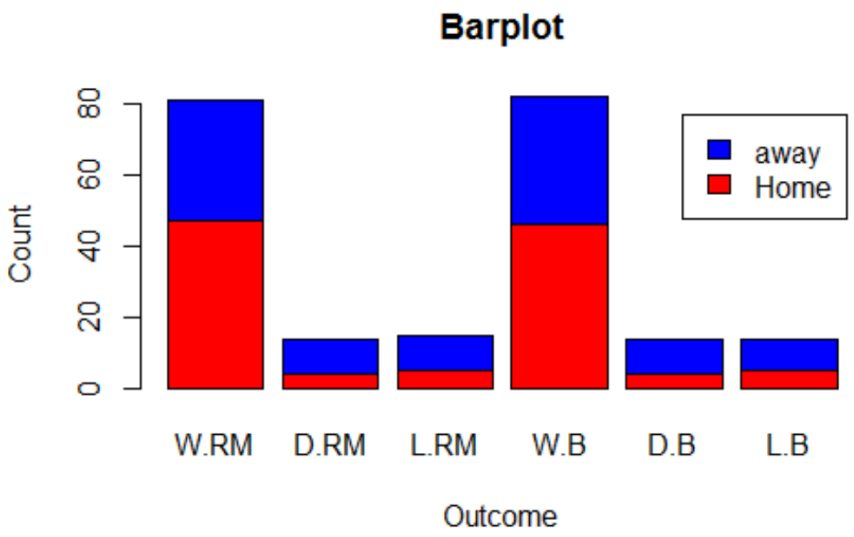


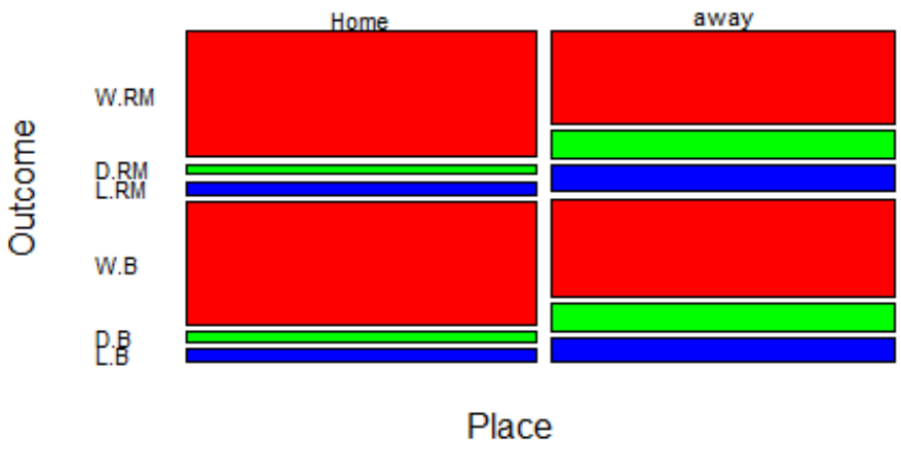




- Creating Mosaic plots about number of wins by Real Madrid and FC Barcelona.

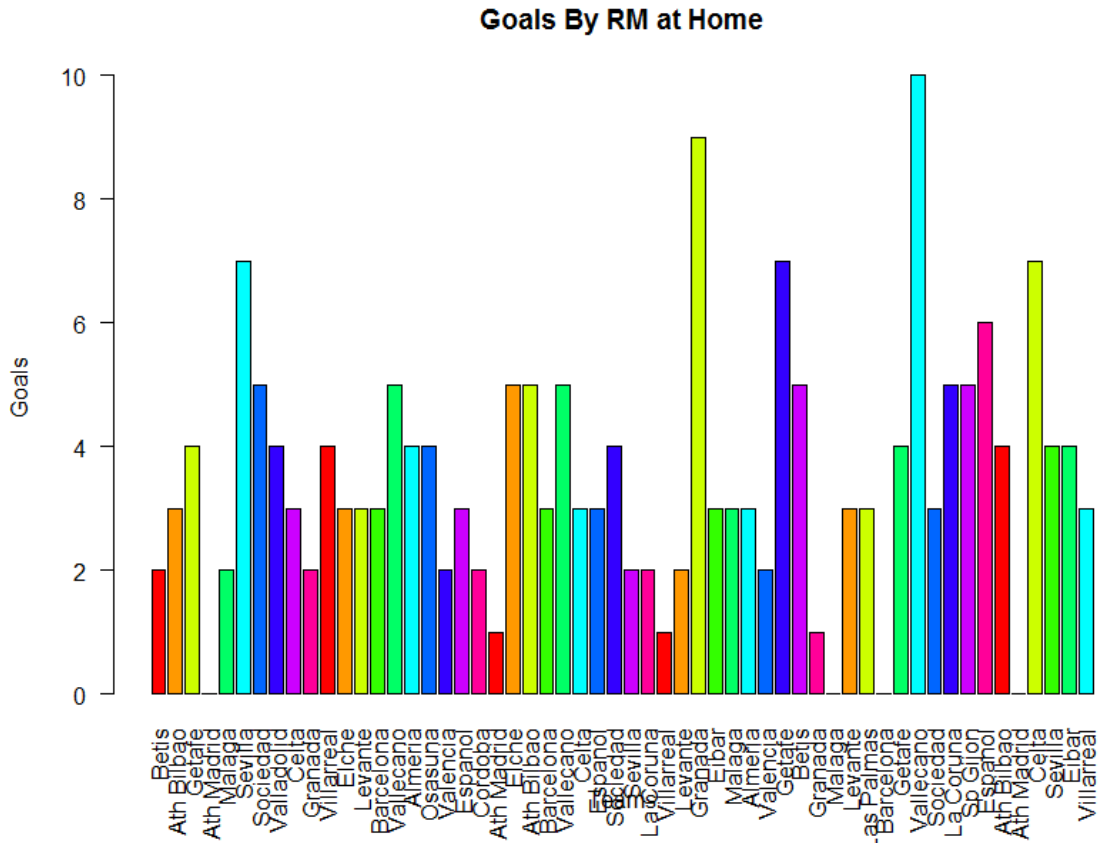
```
> #Categorical Data Matrix and Mosaic plots
> HomeResults.RM

 A  D  H
 5  4 47
> AwayResults.RM

 A  D  H
34 10 10
> HomeResults.FCB

 A  D  H
 5  4 46
> AwayResults.FCB

 A  D  H
36 10  9
> Result <- matrix(c(47,34,4,10,5,10,46,36,4,10,5,9), nrow = 2)
> Result
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]   47    4    5   46    4    5
[2,]   34   10   10   36   10    9
> rownames(Result) <- c("Home","Away")
> colnames(Result) <- c("W.RM","D.RM","L.RM","W.B","D.B","L.B")
> Result
     W.RM D.RM L.RM W.B D.B L.B
Home   47    4    5  46   4   5
Away   34   10   10  36  10   9
> tmp  <- c("Home","away")
> tmp1 <- c("W.RM","D.RM","L.RM","W.B","D.B","L.B")
> Result
       Outcome
Place   W.RM D.RM L.RM W.B D.B L.B
  Home    47    4    5  46   4   5
  away    34   10   10  36  10   9
> barplot(Result,main="Barplot",col = c("red","blue"),legend.text = TRUE,
+ ylab = "Count",xlab = "Outcome")
> mosaicplot(Result,main="Mosaic plot",color = rainbow(3),las = 1)
```



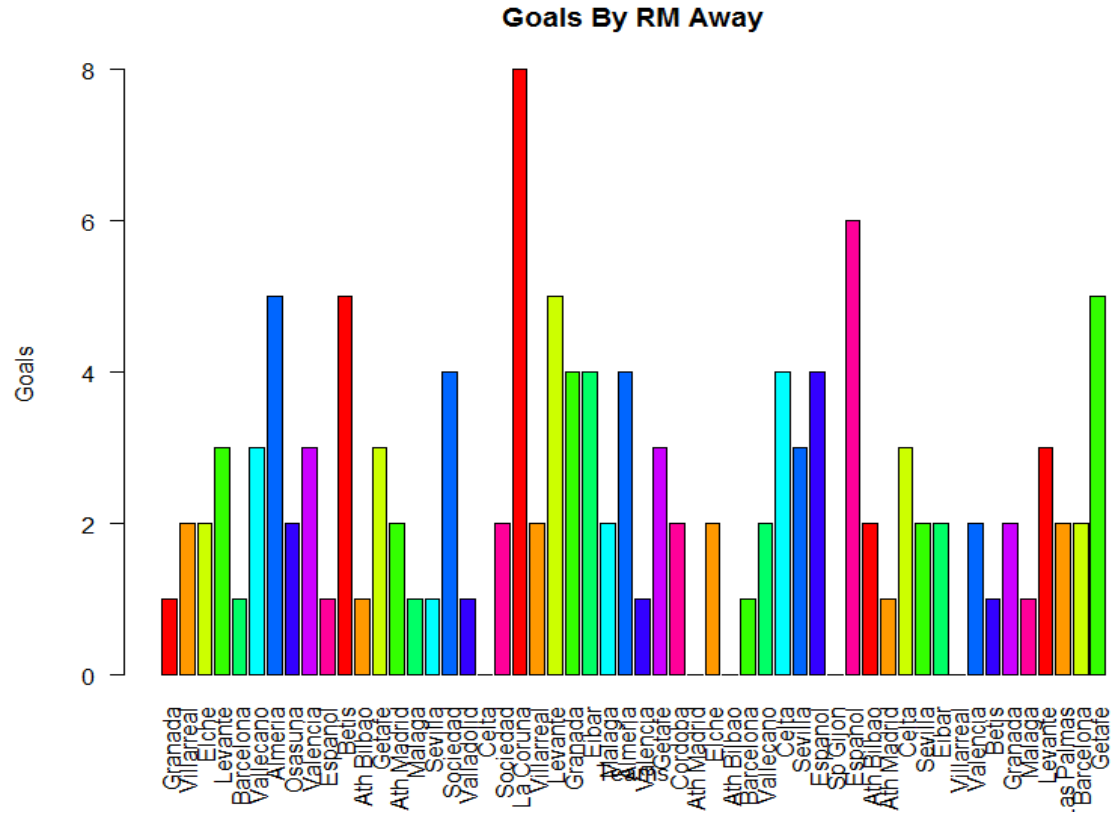


- Comparing the distribution of goals score by Real Madrid and Barcelona.

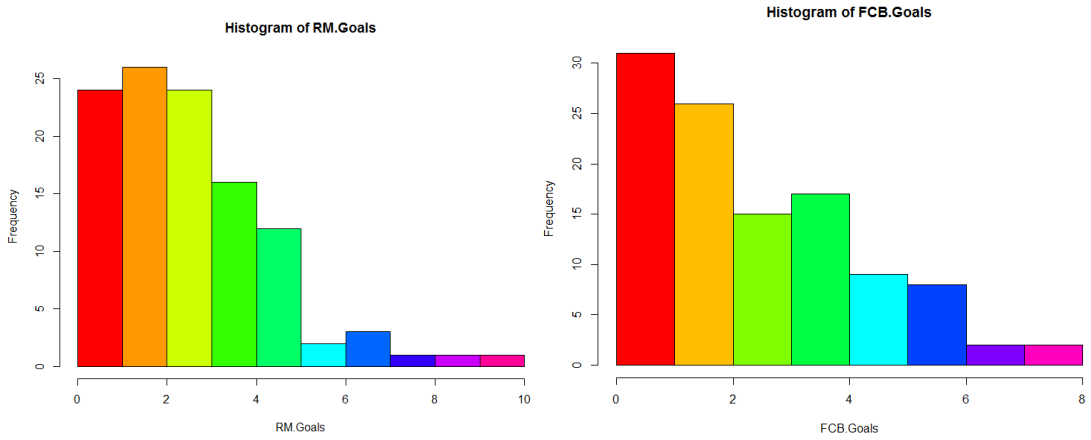
```
> #Plots of Numerical Data - Goal Data
> RM.Home.Goals <- RealMadrid$FTHG[RealMadrid$HomeTeam=="Real Madrid"]
> RM.Away.Goals <- RealMadrid$FTAG[RealMadrid$AwayTeam=="Real Madrid"]
> RM.Goals <- c(RM.Home.Goals,RM.Away.Goals)
> RM.Goals2 <- matrix(RM.Home.Goals,RM.Away.Goals)
> FCB.H.Goals <- Barcelona$FTHG[Barcelona$HomeTeam=="Barcelona"]
> FCB.A.Goals <- Barcelona$FTAG[Barcelona$AwayTeam=="Barcelona"]
> FCB.Goals <- c(FCB.H.Goals,FCB.A.Goals)
> FCB.Goals
  [1] 7 3 4 4 2 1 4 2 4 3 2 6 4 7 3 3 2 2 1 3 2 6 3 0 5 5 5 3 3 5 0 6 2 4 2 6 2 2 1 4 2 5 3 3 4 2 4 4 6
 [51] 6 2 6 1 1 1 3 4 2 0 3 4 0 5 0 1 4 1 0 4 1 0 3 0 1 5 0 2 1 2 1 0 0 4 6 5 3 2 1 2 2 8 1 1 2 1 1 2 4
[101] 0 2 2 3 2 5 4 2 0 8
> mean(RM.Goals)
[1] 2.936364
> mean(FCB.Goals)
[1] 2.781818
> table(RM.Goals)
RM.Goals
 0  1  2  3  4  5  6  7  8  9 10
 9 15 26 24 16 12  2  3  1  1  1
> table(FCB.Goals)
FCB.Goals
 0  1  2  3  4  5  6  7  8
13 18 26 15 17  9  8  2  2
> fivenum(RM.Goals)
[1]  0  2  3  4 10
> fivenum(FCB.Goals)
[1] 0 1 2 4 8
> summary(RM.Goals)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   2.000   3.000   2.936   4.000  10.000
> summary(FCB.Goals)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   2.000   2.782   4.000   8.000
> Opp.names <- RealMadrid$AwayTeam[RealMadrid$HomeTeam=="Real Madrid"]
> Opp.names <- Opp.names[Opp.names!="RealMadrid"]
> barplot(RM.Home.Goals,xlab = "Teams",main = "Goals By RM at Home",
+ col = rainbow(10),names.arg = Opp.names,
+ ylab="Goals", las=2)
```




```
> Opp.names <- RealMadrid$HomeTeam[RealMadrid$AwayTeam=="Real Madrid"]
> barplot(RM.Away.Goals,xlab = "Teams",main = "Goals By RM Away",
+ col = rainbow(10),names.arg = Opp.names,
+ ylab="Goals", las=2)
```



```
> hist(RM.Goals, col = rainbow(10))
> hist(FCB.Goals,col = rainbow(8))
> boxplot(RM.Goals, col = "Green",horizontal = TRUE)
> boxplot(FCB.Goals,col="Green",horizontal = TRUE)
```

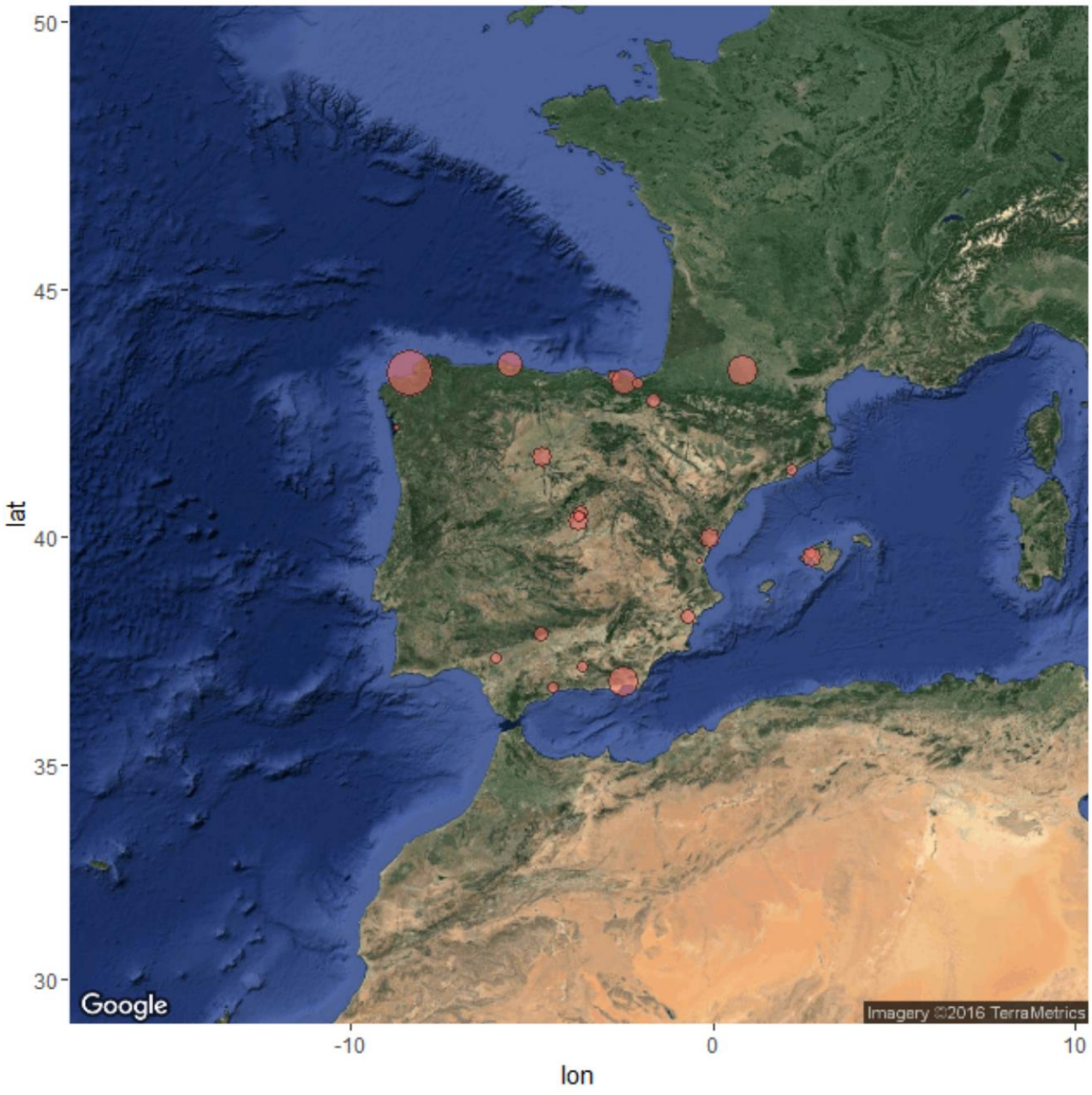


- From the above Histogram we can see that Real Madrid has less number of 0 goals compared to Barcelona which means Real Madrid has better record of scoring at least one goals than Barcelona.
- We can also see the geographical distribution of goal data using GGMap function.

```
> #Geographical representation of data
> Opp.names
 [1] Granada    Villarreal Elche      Levante    Barcelona  Vallecano  Almeria    Osasuna    Valencia
[10] Espanol    Betis      Ath Bilbao Getafe     Ath Madrid Malaga     Sevilla    Sociedad   Valladolid
[19] Celta      Sociedad   La Coruna  Villarreal Levante    Granada    Eibar      Malaga     Almeria
[28] Valencia   Getafe     Cordoba    Ath Madrid Elche      Ath Bilbao Barcelona  Vallecano  Celta
[37] Sevilla    Espanol    Sp Gijon   Espanol    Ath Bilbao Ath Madrid Celta      Sevilla    Eibar
[46] Villarreal Valencia   Betis      Granada    Malaga     Levante    Las Palmas Barcelona  Getafe
25 Levels: Almeria Ath Bilbao Ath Madrid Barcelona Betis Celta Cordoba Eibar Elche Espanol ... Villarreal
> Teamnames <- levels(Opp.names)
> Teamnames <- Teamnames[-(18)]
> Teamnames
 [1] "Almeria"    "Ath Bilbao" "Ath Madrid" "Barcelona"  "Betis"      "Celta"      "Cordoba"
 [8] "Eibar"      "Elche"      "Espanol"    "Getafe"     "Granada"    "La Coruna"  "Las Palmas"
[15] "Levante"    "Malaga"     "Osasuna"    "Sevilla"    "Sociedad"   "Sp Gijon"   "Valencia"
[22] "Valladolid" "Vallecano"  "Villarreal"
> freq <- c(1,2)
> for(i in 1:24) freq[i] <- RealMadrid$FTAG[RealMadrid$HomeTeam==Teamnames[i]]
There were 18 warnings (use warnings() to see them)
> freq
 [1] 5 1 2 1 5 0 2 4 2 1 3 1 8 2 3 1 2 1 4 0 3 1 3 2
> freq <- freq[1:24]

> for(i in 1:24) freq[i] <- RealMadrid$FTAG[RealMadrid$HomeTeam==Teamnames[i]]
There were 18 warnings (use warnings() to see them)
> freq
 [1] 5 1 2 1 5 0 2 4 2 1 3 1 8 2 3 1 2 1 4 0 3 1 3 2
> freq <- freq[1:24]
> for (i in 1:24){
+ freq[i] <- freq[i]+1
+ }
> freq
 [1] 6 2 3 2 6 1 3 5 3 2 4 2 9 3 4 2 3 2 5 1 4 2 4 3
> length(freq)
[1] 24
> length(Teamnames)
[1] 24
> library(ggmap)
```

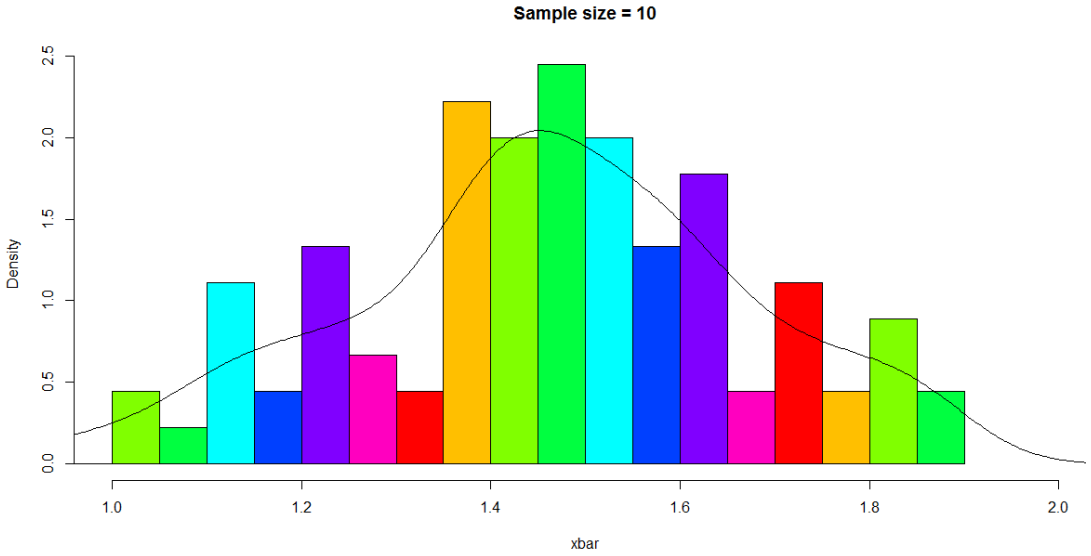
```
> Teamnames[6] <- "Vigo"
> Teamnames[10] <- "RCD Espanyol"
> Teamnames[14] <- "Las Palmas"
> Teamnames[17] <- "Pamplona"
> Teamnames[19] <- "Anoeta"
> Teamnames[23] <- "Madrid"
> Teamnames
 [1] "Almeria"      "Ath Bilbao"   "Ath Madrid"   "Barcelona"    "Betis"        "Vigo"
 [7] "Cordoba"      "Eibar"        "Elche"        "RCD Espanyol" "Getafe"       "Granada"
[13] "La Coruna"    "Las Palmas"   "Levante"      "Malaga"       "Pamplona"     "Sevilla"
[19] "Anoeta"       "Sp Gijon"     "Valencia"     "Valladolid"   "Madrid"       "Villarreal"
> location <- geocode(as.character(Teamnames))
```

```
> ggmap(map) +
+ geom_point(data = location, aes(x = location$lon, y = location$lat, fill = "red", alpha = 0.8), size = freq,
  shape = 21) +
+ guides(fill=FALSE, alpha=FALSE, size=FALSE)
Warning message:
Removed 1 rows containing missing values (geom_point).
```

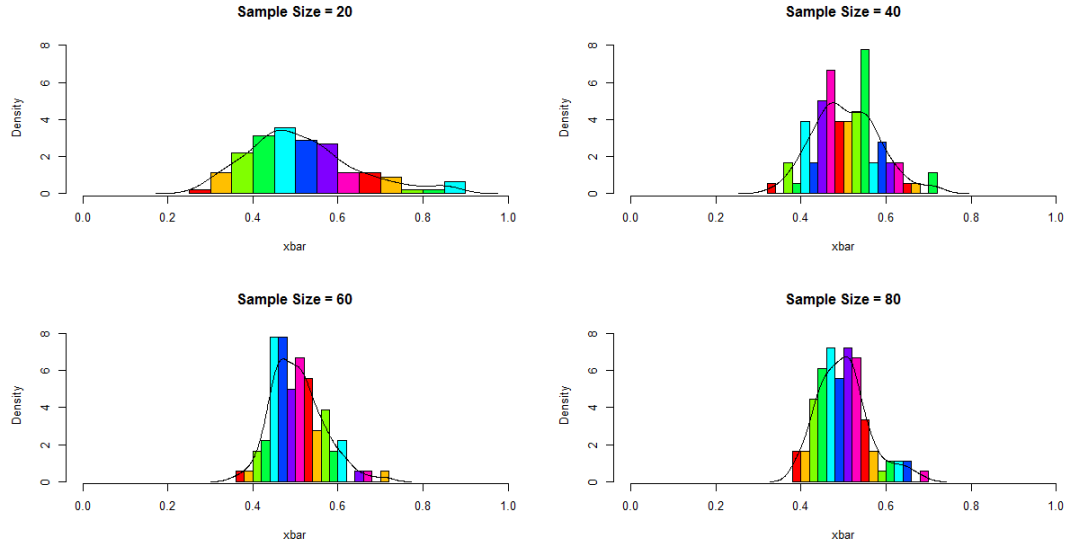


- Random sampling of BET 365 data to see the distribution of betting odds of Real Madrid in both Home and Away matches.

```
> #Part 3
> #Sample Means
> RM.b.Home <- RealMadrid$B365H[RealMadrid$HomeTeam=="Real Madrid"]
> RM.b.Away <- RealMadrid$B365A[RealMadrid$AwayTeam=="Real Madrid"]
> B.RM <- c(RM.b.Home,RM.b.Away)
> mbet<-mean(B.RM)
> mbet
[1] 1.431909
> sigma<-sd(B.RM)
> sigma
[1] 0.6519561
> samples<-90
> sample.size<-10
> xbar<-numeric(samples)
> for(i in 1:samples){
+ xbar[i]<- mean(rnorm(sample.size, mean= mbet, sd<-sigma))
+ }
> B.RM
  [1] 1.17 1.17 1.13 1.62 1.20 1.18 1.14 1.13 1.10 1.05 1.33 1.07 1.05 2.25 1.04 1.05 1.05 1.14 1.33 1.06
 [21] 1.67 1.08 1.20 2.38 1.05 1.09 1.08 1.18 1.29 1.08 1.20 1.06 1.08 1.06 1.14 1.06 1.40 1.13 1.10 1.06
 [41] 1.11 1.08 1.09 2.50 1.10 1.10 1.10 1.11 1.05 1.10 1.29 1.73 1.30 1.33 1.17 1.22 1.33 1.57 1.29 1.33
 [61] 4.33 1.25 1.29 1.22 1.57 1.33 1.29 1.62 1.29 2.05 1.25 1.53 1.57 1.20 1.40 1.36 1.25 1.62 1.17 1.20
 [81] 1.18 1.29 1.17 1.57 1.22 1.14 2.45 1.18 1.44 4.50 1.20 1.57 2.05 1.50 1.22 1.33 1.62 2.30 1.91 1.91
[101] 1.40 1.65 1.53 1.30 1.25 1.73 1.44 1.33 5.50 1.29
```



❖ The graph follows Normal distribution. So Central Limit theorem can be applied. The following graph shows the distribution of means of sample sizes of 20, 40, 60, 80 and as the sample size increase we can see that the distribution is narrower.



```
Sample Size =  20  Mean =  0.506735  SD =  0.1125792
Sample Size =  40  Mean =  0.520515  SD =  0.08374967
Sample Size =  60  Mean =  0.4958534  SD =  0.07038503
Sample Size =  80  Mean =  0.5052637  SD =  0.05806121
```

- Below are various sampling techniques that I Performed on Bet365 data of the Real Madrid.

```
> #Part 4
> #Random Sampling
> library(sampling)
> head(RealMadrid)
   Month    HomeTeam    AwayTeam FTHG FTAG FTR HTHG HTAG HTR HS AS HST AST HF AF HC AC HY AY HR AR B365H
6    Aug Real Madrid       Betis    2    1   H    1    1   D 20 11   9   4 11 20  5  7  1  2  0  0  1.17
20   Aug     Granada Real Madrid    0    1   A    0    1   A  8 21   3   8 14 10  5  8  4  2  0  0  7.50
27   Aug Real Madrid  Ath Bilbao    3    1   H    2    0   H 20 12   7   2 15 13  9  6  1  2  0  0  1.17
34   Sep  Villarreal Real Madrid    2    2   D    1    1   D 19 20   9   6 14 13 10  3  2  3  0  0  5.50
48   Sep Real Madrid      Getafe    4    1   H    2    1   H 29  7  11   4 13 15  9  5  1  3  0  1  1.13
55   Sep       Elche Real Madrid    1    2   A    0    0   D 11 12   3   5 23  9  4  5  8  2  1  0 10.00
   B365D B365A
6    7.0 17.00
20   5.5  1.33
27   7.0 16.00
34   4.2  1.57
48   8.5 19.00
55   5.5  1.29
> #Simple random sample of size 40 with replacement
> s<- srswr(40, nrow(RealMadrid))
> s
  [1] 1 1 1 0 0 0 0 0 0 1 3 1 0 1 0 0 0 0 4 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0 0 0 1 1 0 2 1 0 1 1 0
 [51] 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 2 1 2 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 2 0 0
[101] 0 0 0 0 0 0 0 0 1 2
> s[s!= 0]
 [1] 1 1 1 1 3 1 1 4 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 2 1 1 1 2 1 2
> rows<- (1:nrow(RealMadrid))[s !=0]
> rows<- rep(rows, s[s != 0])
```

```
> rows
 [1]   1   2   3  10  11  11  11  12  14  19  19  19  19  21  27  29  38  42  43  45  45  46  48  49  51
[26]  59  71  75  75  76  77  77  82  89  90  98  98 109 110 110
> #The data of the selected sample and frequency is shown below
> sample.1<-RealMadrid[rows, ]
> head(sample.1)
      Month    HomeTeam    AwayTeam FTHG FTAG FTR HTHG HTAG HTR HS AS HST AST HF AF HC AC HY AY HR AR
6       Aug Real Madrid       Betis    2    1   H    1    1   D 20 11   9   4 11 20  5  7  1  2  0  0
20      Aug     Granada Real Madrid    0    1   A    0    1   A  8 21   3   8 14 10  5  8  4  2  0  0
27      Aug Real Madrid  Ath Bilbao    3    1   H    2    0   H 20 12   7   2 15 13  9  6  1  2  0  0
92      Oct   Barcelona Real Madrid    2    1   H    1    0   H 12 10   5   6 19 16  4  3  2  5  0  0
104     Oct Real Madrid     Sevilla    7    3   H    3    2   H 18 16  10   6 17 15  2  1  3  2  0  1
104.1   Oct Real Madrid     Sevilla    7    3   H    3    2   H 18 16  10   6 17 15  2  1  3  2  0  1
      B365H B365D B365A
6      1.17   7.0 17.00
20     7.50   5.5  1.33
27     1.17   7.0 16.00
92     1.75   4.0  4.33
104    1.18   7.0 13.00
104.1  1.18   7.0 13.00
```

```
> table(sample.1$HS[RealMadrid$HomeTeam=="Real Madrid"])

 6  8 11 12 14 15 16 18 19 20 25 31
 1  1  3  3  1  1  1  2  1  4  1  1
> table(sample.1$AS[RealMadrid$AwayTeam=="Real Madrid"])

 5  6  8 10 11 13 14 15 16 18 21 23 27
 1  1  1  2  1  1  1  4  1  2  2  2  1
> ##################
> s<- srswr(40, nrow(RealMadrid))
> sample.2<-RealMadrid[s != 0, ]
> head(sample.2)
    Month    HomeTeam    AwayTeam FTHG FTAG FTR HTHG HTAG HTR HS AS HST AST HF AF HC AC HY AY HR AR B365H
20    Aug     Granada Real Madrid    0    1   A    0    1   A  8 21   3   8 14 10  5  8  4  2  0  0  7.50
27    Aug Real Madrid  Ath Bilbao    3    1   H    2    0   H 20 12   7   2 15 13  9  6  1  2  0  0  1.17
34    Sep  Villarreal Real Madrid    2    2   D    1    1   D 19 20   9   6 14 13 10  3  2  3  0  0  5.50
63    Sep Real Madrid  Ath Madrid    0    1   A    0    1   A 20 13   5   4 21 18  2  7  4  4  0  0  1.62
77    Oct     Levante Real Madrid    2    3   A    0    0   D 13 27   5  10 18 13  3  8  1  3  0  0  9.00
146   Dec Real Madrid  Valladolid    4    0   H    2    0   H 26  9  12   3  6  8 10  1  1  1  0  0  1.13
    B365D B365A
20   5.50  1.33
27   7.00 16.00
34   4.20  1.57
63   3.75  5.50
77   5.00  1.33
146  8.50 17.00
```

```
> table(sample.2$HS[RealMadrid$HomeTeam=="Real Madrid"])

 8  9 11 12 13 18 19 24 26
 1  1  4  2  4  2  2  1  1
> table(sample.2$AS[RealMadrid$AwayTeam=="Real Madrid"])

 6  8  9 10 11 12 13 15 18 21 24
 2  1  3  1  1  1  3  1  2  1  1
```

```
> #Systematic sampling
> N<-110
> n<-30
> k<-ceiling(N/n)
> k
[1] 4
> #random sample from first group
> r<- sample(k,1)
> r
[1] 4
> #Select every kth item
> s <- seq(r, by=k, length=n)
> sample.3<-RealMadrid[s, ]
```

```
> table(sample.3$HS[RealMadrid$HomeTeam=="Real Madrid"])

 6  7 12 13 17 18 19 20 23 30
 1  1  1  1  1  1  3  2  1  1
> table(sample.3$AS[RealMadrid$AwayTeam=="Real Madrid"])

 8  9 12 13 14 16 17 18 19 21 24 27
 1  1  1  1  1  1  2  2  1  1  1  1
> #Unequal Probablities
> pik<- inclusionprobabilities(RealMadrid$HS[RealMadrid$HomeTeam=="Real Madrid"], 30)
> length(pik)
[1] 56
> sum(pik)
[1] 30
> s<- UPsystematic(pik)
> sample.4<-RealMadrid[s != 0, ]
> head(sample.4)
   Month    HomeTeam    AwayTeam FTHG FTAG FTR HTHG HTAG HTR HS AS HST AST HF AF HC AC HY AY HR AR B365H
20   Aug     Granada Real Madrid    0    1   A    0    1   A  8 21   3   8 14 10  5  8  4  2  0  0  7.50
27   Aug Real Madrid  Ath Bilbao    3    1   H    2    0   H 20 12   7   2 15 13  9  6  1  2  0  0  1.17
48   Sep Real Madrid      Getafe    4    1   H    2    1   H 29  7  11   4 13 15  9  5  1  3  0  1  1.13
63   Sep Real Madrid  Ath Madrid    0    1   A    0    1   A 20 13   5   4 21 18  2  7  4  4  0  0  1.62
83   Oct Real Madrid      Malaga    2    0   H    0    0   D 22  3  15   1  5 12 10  1  1  5  0  0  1.20
92   Oct   Barcelona Real Madrid    2    1   H    1    0   H 12 10   5   6 19 16  4  3  2  5  0  0  1.75
   B365D B365A
20  5.50  1.33
27  7.00 16.00
48  8.50 19.00
63  3.75  5.50
83  7.00 11.00
92  4.00  4.33
> table(sample.4$HS[RealMadrid$HomeTeam=="Real Madrid"])

 4  7  8  9 10 11 12 13 14 15 17 18 20 22 25 26 29
 1  2  1  1  1  3  1  2  2  1  2  2  4  2  1  1  2
```

- Calculating the number of goals scored within the confidence intervals.

```
> #Part 5
> #Confidence Intervals for Population mean.
> table(RM.Goals)
RM.Goals
 0  1  2  3  4  5  6  7  8  9 10
 9 15 26 24 16 12  2  3  1  1  1
> pop.sd<-sd(RM.Goals)
> pop.mean<-mean(RM.Goals)
> pop.mean
[1] 2.936364
> conf<- c(80,90)
> typeof(pop.mean)
[1] "double"
> pop.mean - 2*pop.sd
[1] -0.9262045
> pop.mean+qnorm(1-i/2)*pop.sd
[1] NaN
Warning message:
In qnorm(1 - i/2) : NaNs produced
> alpha<-1-conf/100
> alpha
[1] 0.2 0.1
> qnorm(alpha/2)
[1] -1.281552 -1.644854
> qnorm(1-alpha/2)
[1] 1.281552 1.644854
```

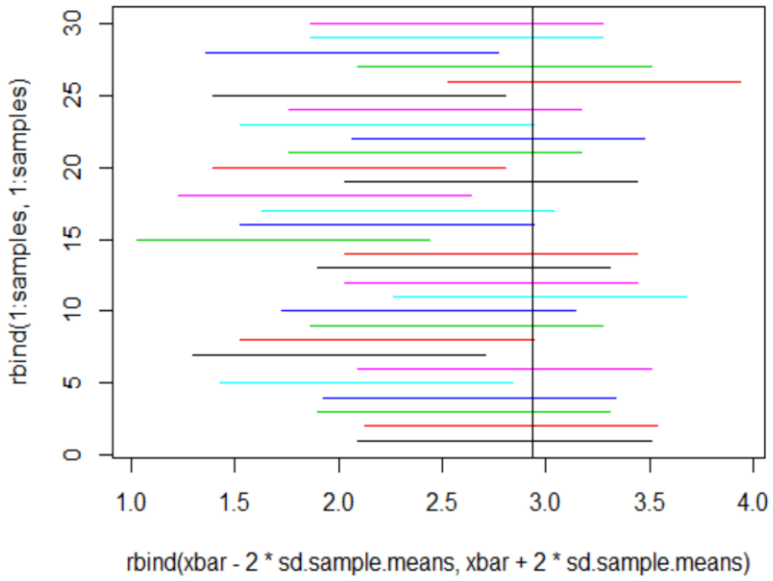
```
> for (i in alpha) {
+ str <- sprintf("%2d%% Confidence Level (alpha = %.2f), z: %.2f , %.2f",
+ 100*(1-i), i,
+ pop.mean- qt(1-i/2,df=n-1)*pop.sd,
+ pop.mean+ qt(1-i/2,df=n-1)*pop.sd)
+ cat(str,"\n")
+ }
80% Confidence Level (alpha = 0.20), z: 0.40 , 5.47
90% Confidence Level (alpha = 0.10), z: -0.35 , 6.22
> ###Precision
> for (i in alpha) {
+ str <- sprintf("%2d%% Confidence Level (alpha = %.2f), Precision: %.2f",
+ 100*(1-i), i,
+ 2* qt(1-i/2,df=n-1)*pop.sd)
+ cat(str,"\n")
+ }
80% Confidence Level (alpha = 0.20), Precision: 5.07
90% Confidence Level (alpha = 0.10), Precision: 6.56
```

❖ We can see that the precision increased with increase in confidence levels.

```
> #Sample means
> x <- rnorm(1000, mean = pop.mean, sd = pop.sd)
> x <- as.integer(x)

> samples <- 30
> sample.size <- 30
> sd.sample.means <- pop.sd/sqrt(sample.size)
> sd.sample.means
[1] 0.3526026
> xbar <- numeric(samples)
> for (i in 1: samples) {
+ sample.data.1 <- sample(x, size=sample.size)
+ xbar[i] <- mean(sample.data.1)
+ str <- sprintf("%2d: xbar = %.2f, CI = %.2f - %.2f",
+ i, xbar[i], xbar[i] - 2*sd.sample.means,
+ xbar[i] + 2*sd.sample.means)
+ cat(str,"\n")
+ }

 1: xbar = 2.80, CI = 2.09 - 3.51
 2: xbar = 2.83, CI = 2.13 - 3.54
 3: xbar = 2.60, CI = 1.89 - 3.31
 4: xbar = 2.63, CI = 1.93 - 3.34
 5: xbar = 2.13, CI = 1.43 - 2.84
 6: xbar = 2.80, CI = 2.09 - 3.51
 7: xbar = 2.00, CI = 1.29 - 2.71
 8: xbar = 2.23, CI = 1.53 - 2.94
 9: xbar = 2.57, CI = 1.86 - 3.27
10: xbar = 2.43, CI = 1.73 - 3.14
11: xbar = 2.97, CI = 2.26 - 3.67
12: xbar = 2.73, CI = 2.03 - 3.44
13: xbar = 2.60, CI = 1.89 - 3.31
14: xbar = 2.73, CI = 2.03 - 3.44
15: xbar = 1.73, CI = 1.03 - 2.44
16: xbar = 2.23, CI = 1.53 - 2.94
17: xbar = 2.33, CI = 1.63 - 3.04
18: xbar = 1.93, CI = 1.23 - 2.64
19: xbar = 2.73, CI = 2.03 - 3.44
20: xbar = 2.10, CI = 1.39 - 2.81
21: xbar = 2.47, CI = 1.76 - 3.17
22: xbar = 2.77, CI = 2.06 - 3.47
23: xbar = 2.23, CI = 1.53 - 2.94
24: xbar = 2.47, CI = 1.76 - 3.17
25: xbar = 2.10, CI = 1.39 - 2.81
26: xbar = 3.23, CI = 2.53 - 3.94
27: xbar = 2.80, CI = 2.09 - 3.51
28: xbar = 2.07, CI = 1.36 - 2.77
29: xbar = 2.57, CI = 1.86 - 3.27
30: xbar = 2.57, CI = 1.86 - 3.27
```

- Now lets map the Matplot for confidence level of 90.

```
> for(i in alpha){
+ matplot(rbind(xbar - 2*sd.sample.means, xbar + 2*sd.sample.means),
+ rbind(1:samples, 1:samples), type="l", lty=1)
+ abline(v = pop.mean)
+ }
```



❖ In the above plot, 3 confidence intervals are not having a range with population mean in them. It is a confidence level of 90.

Discussion on what I learned:

- Real Madrid and FC Barcelona performed similarly in "Home" games.
- FC Barcelona scored more goals in Home matches than Real Madrid.
- When two teams tie with points by the end of the season, winner will be determined by the net number of goals scored against each other and other teams. So in this case, Barcelona would win the league even if Real Madrid levelled the points.
- FC Barcelona has more number of "Away" wins compared to Real Madrid which is the main reason why they won the league.
- Real Madrid performed well against the teams that come from southern part of Spain. This might be because of the weather conditions that they must endure during their trip towards north.