CS544 Final Term Project

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:

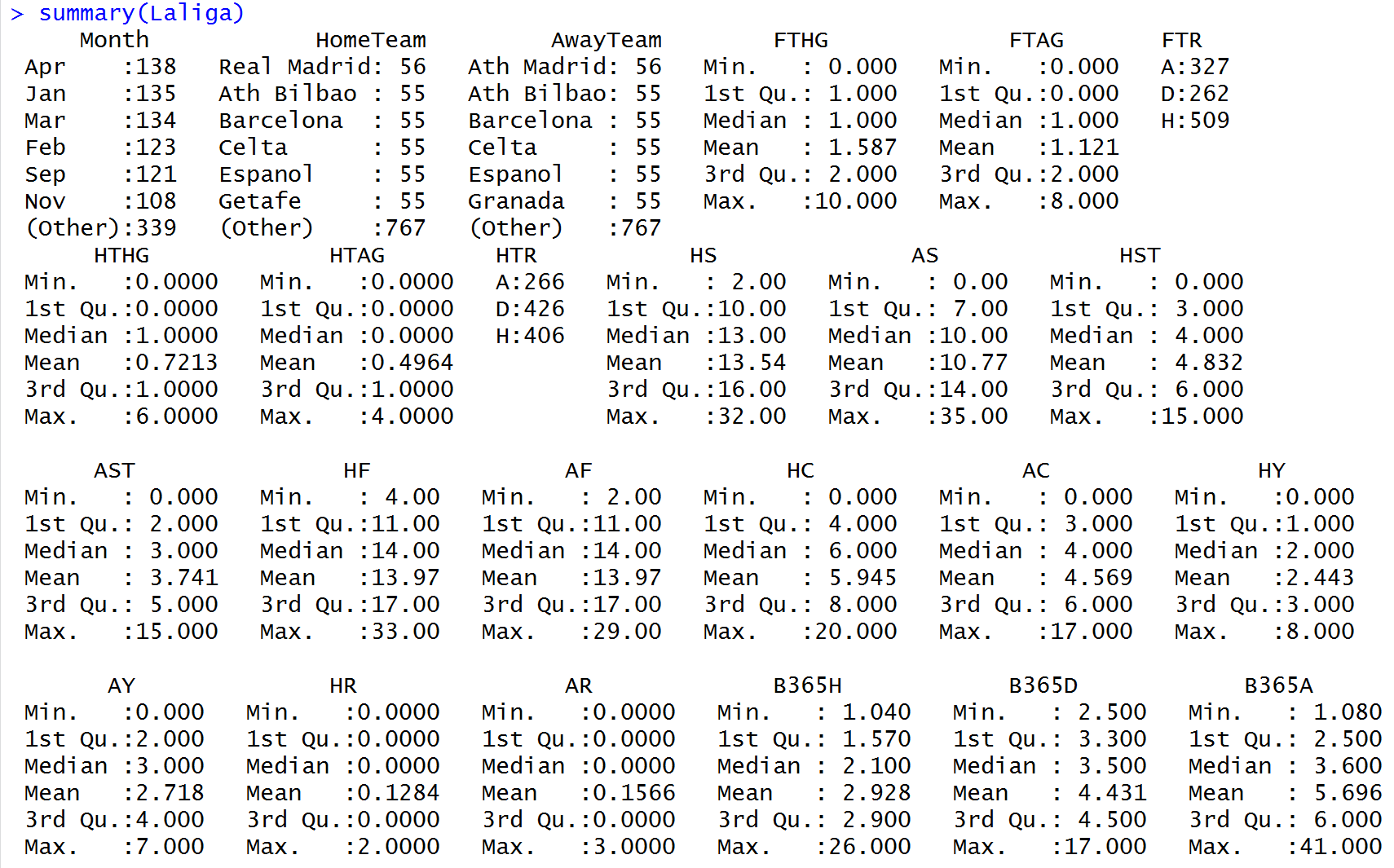
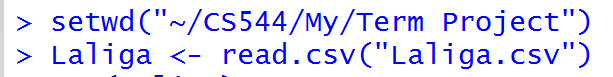
The three seasons data are merged into one excel file and unnecessary attributes were removed. Because of anomalies in the date attribute of matches, I needed to change them to month name using the following steps.

* 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.

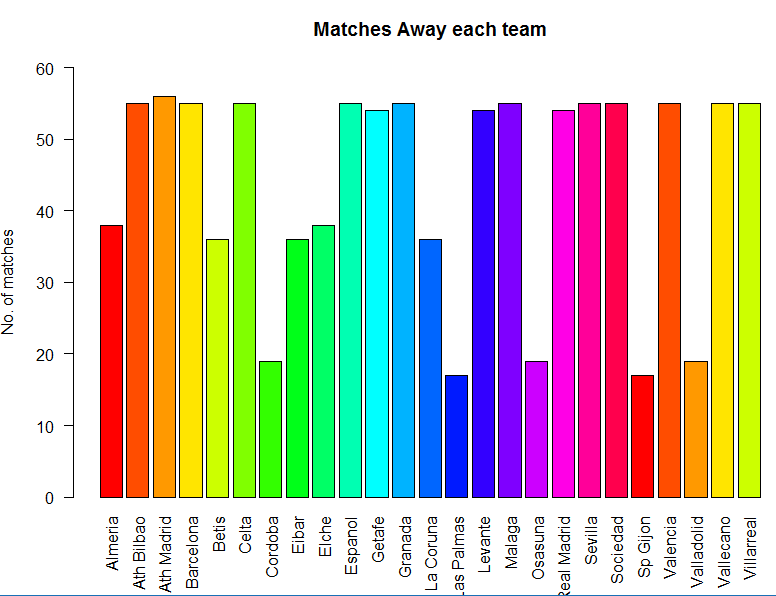
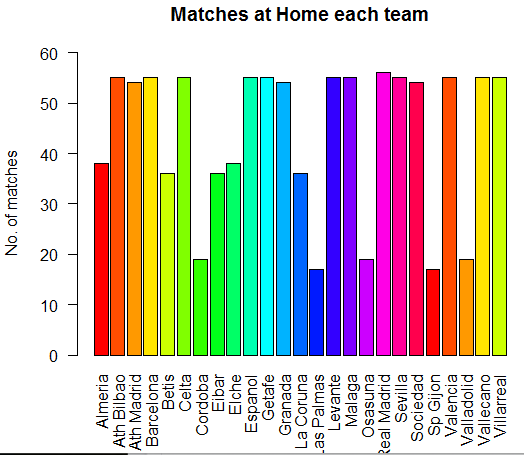
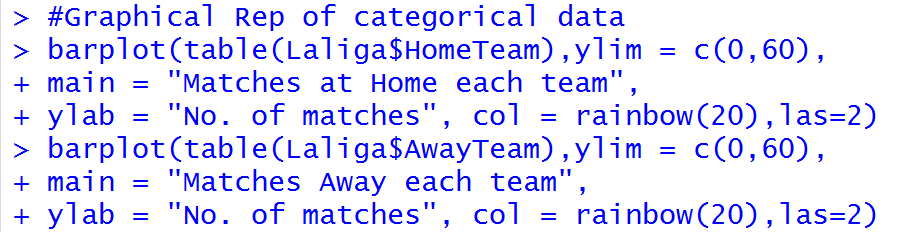


# Step 4: Analysing the Data:

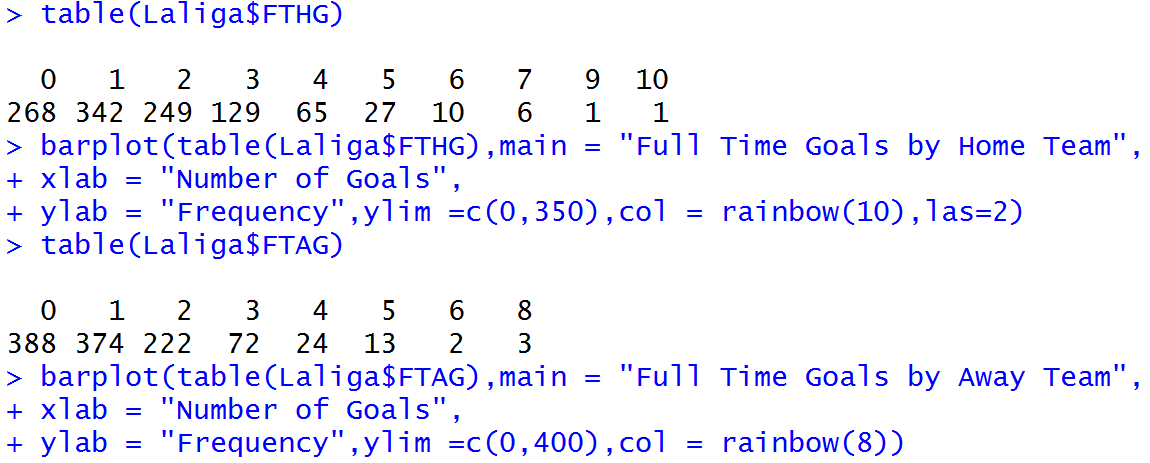
## Part 1: Do the Analysis as in Module 3 for categorical and numerical data. Show appropriate plots for your data.

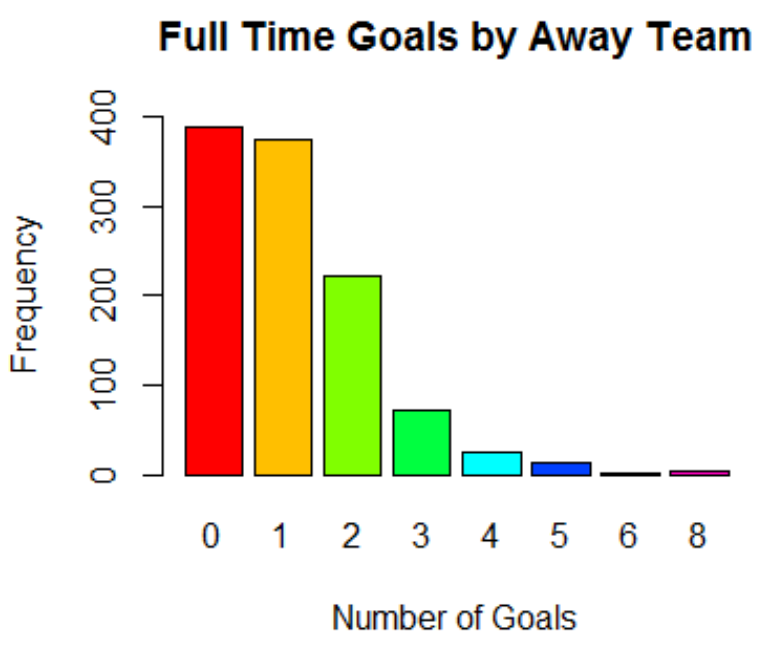
Ans:

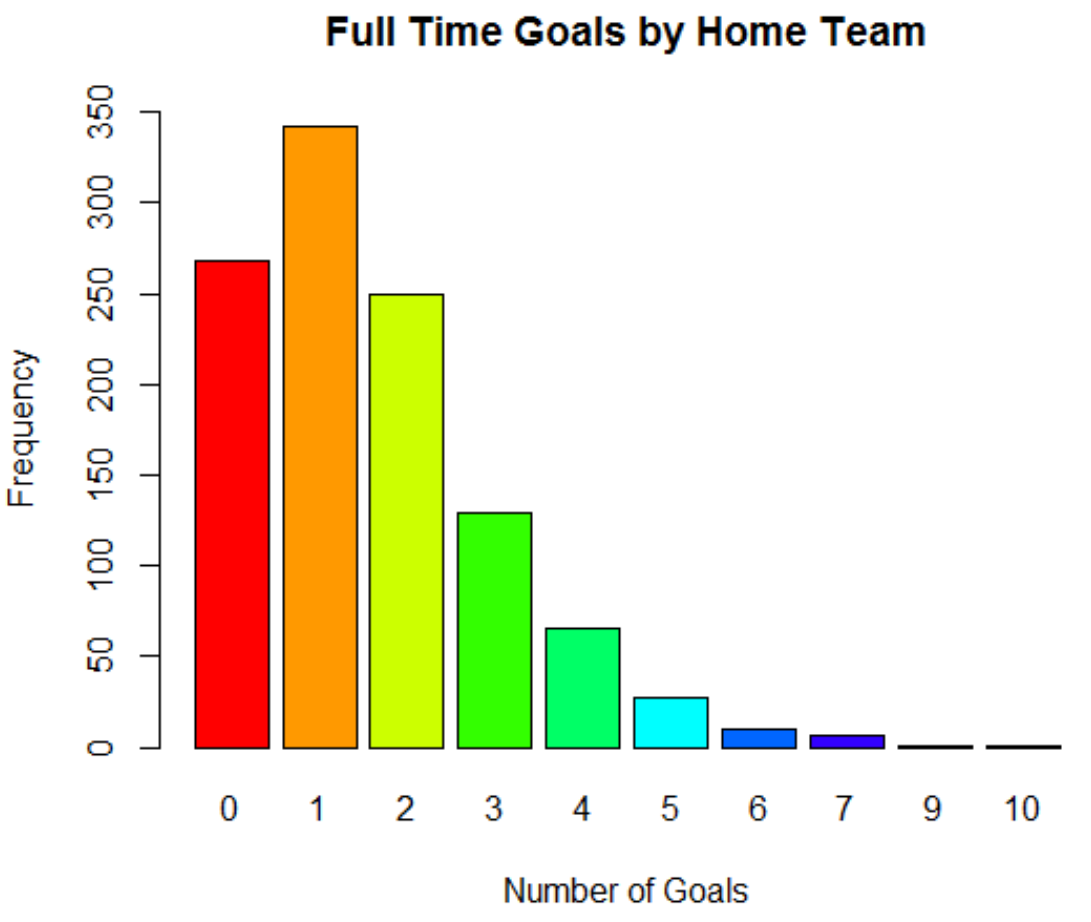
Performed Categorical analysis on LaLiga teams.



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

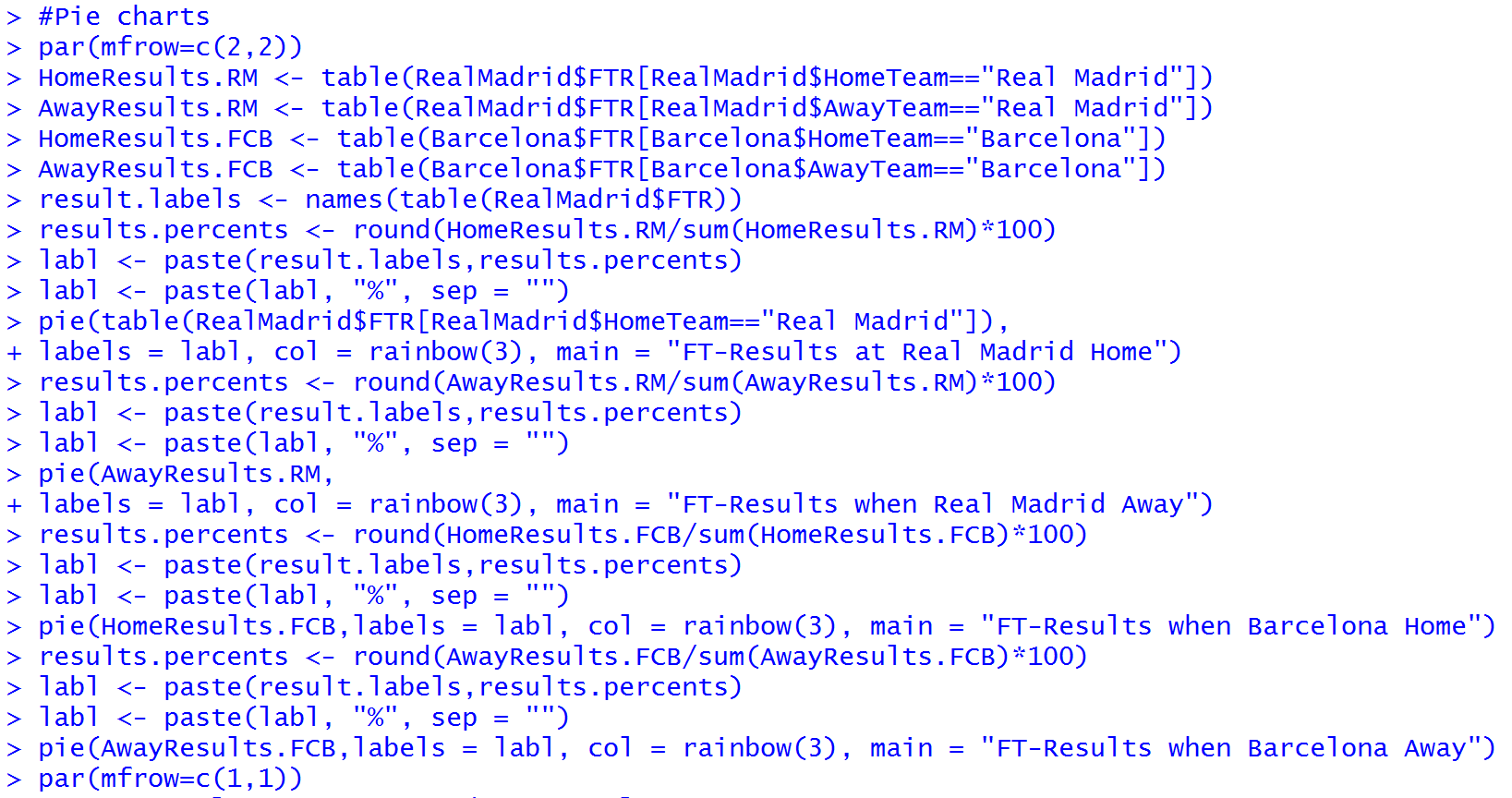


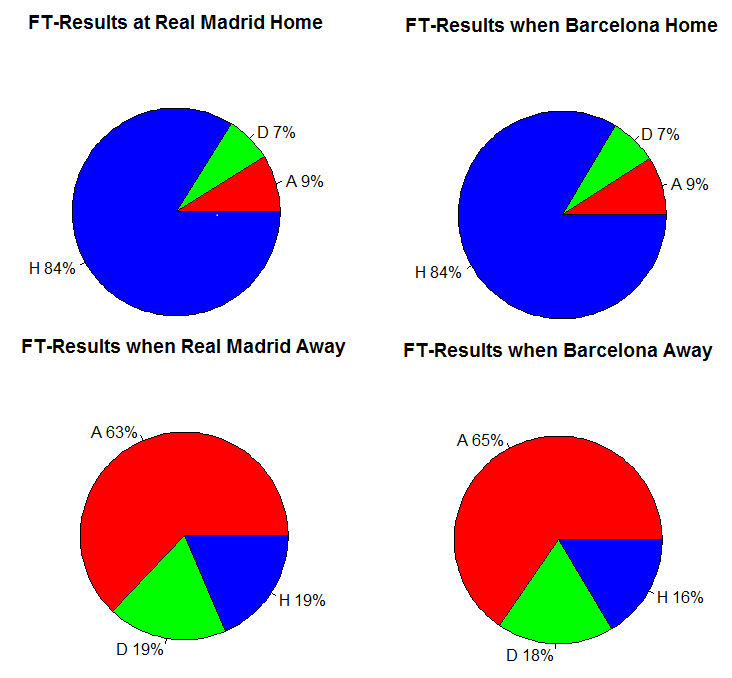




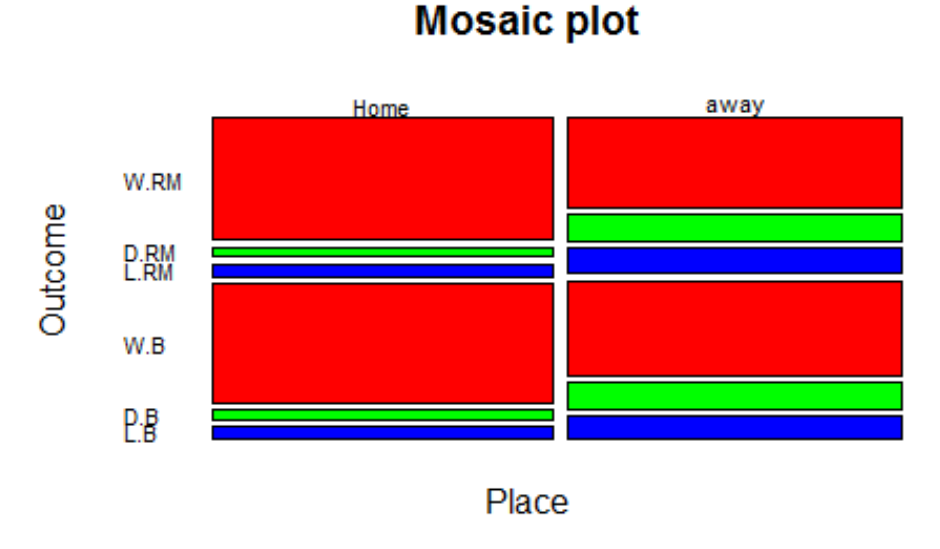
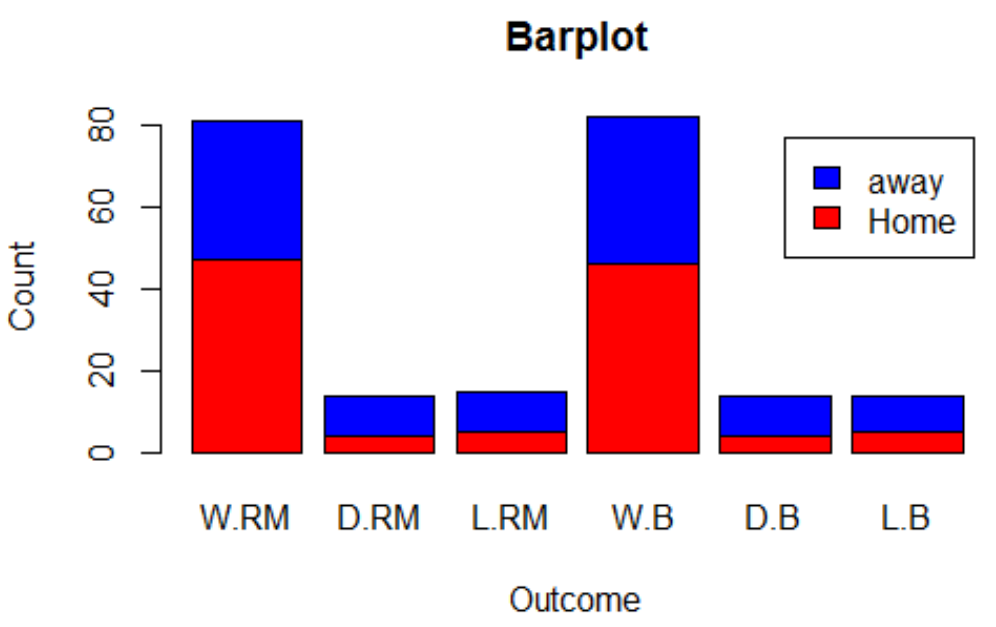
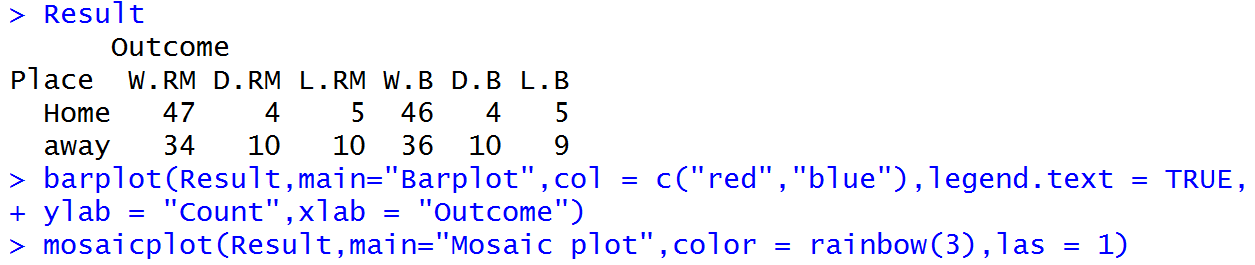
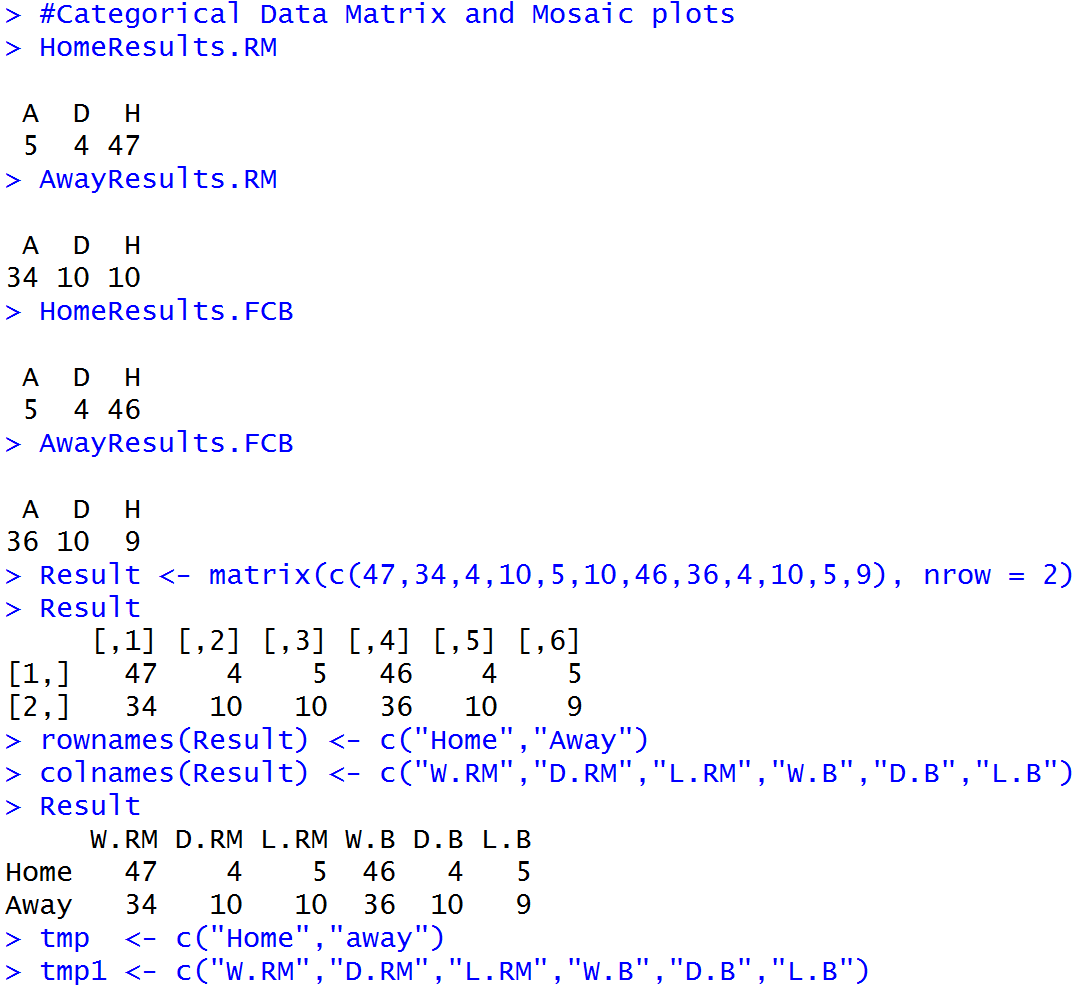
For the purpose of other types of analysis, I am taking data of Real Madrid and Barcelone





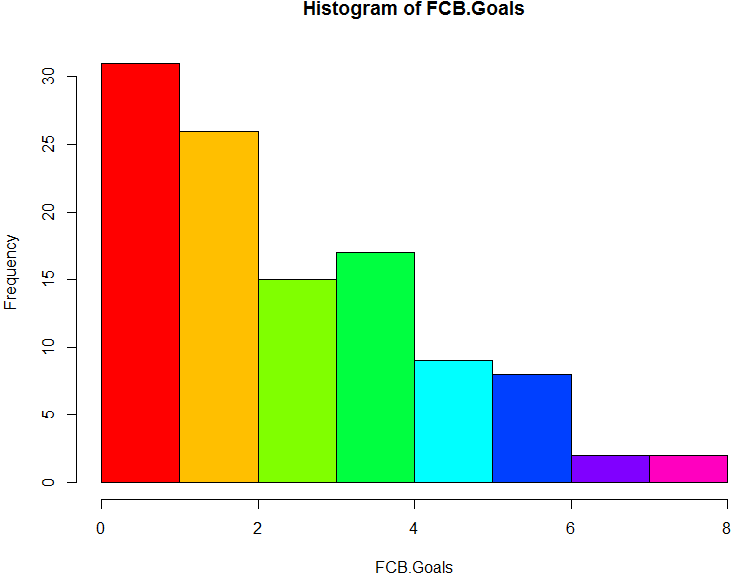
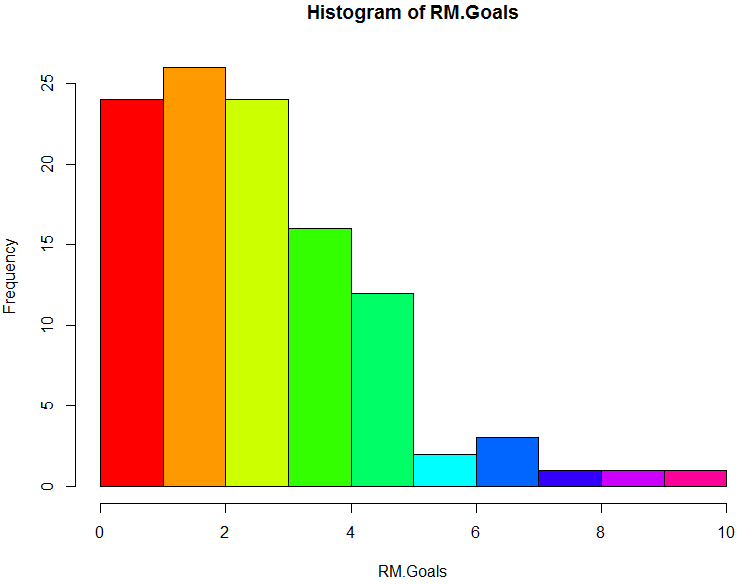
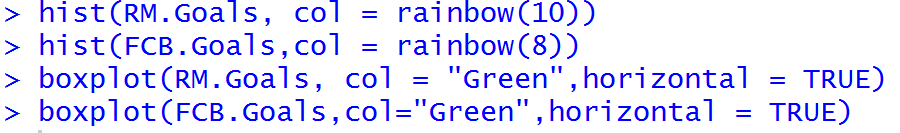
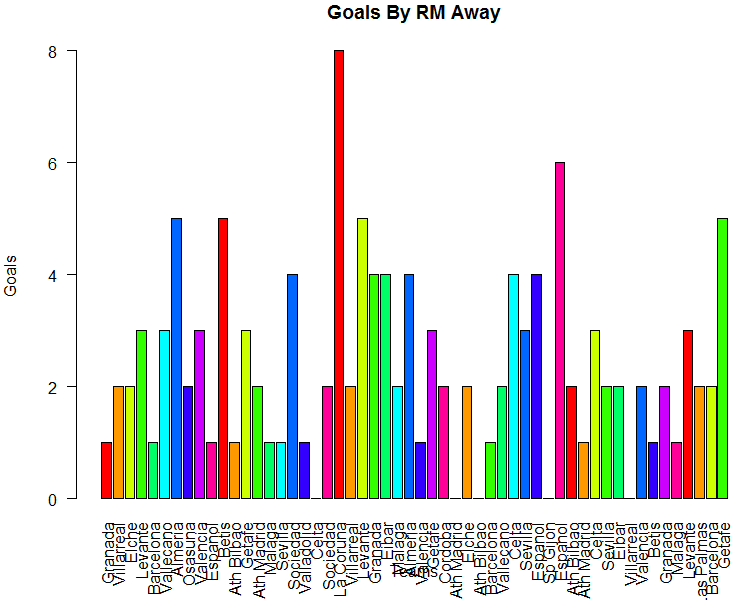
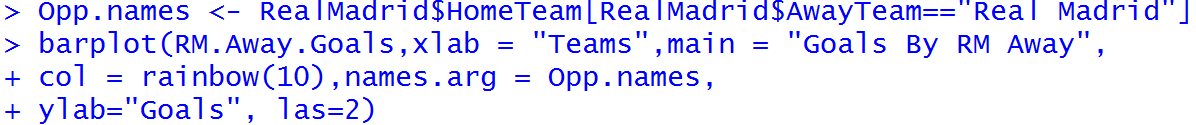
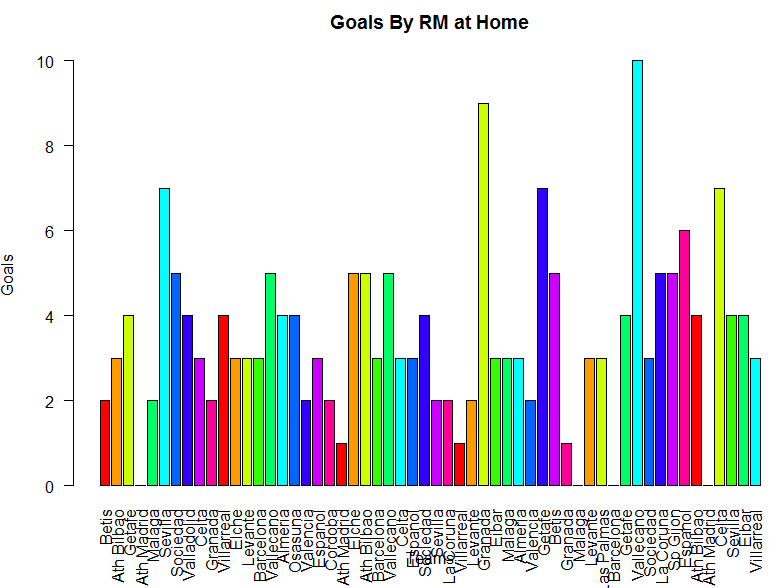
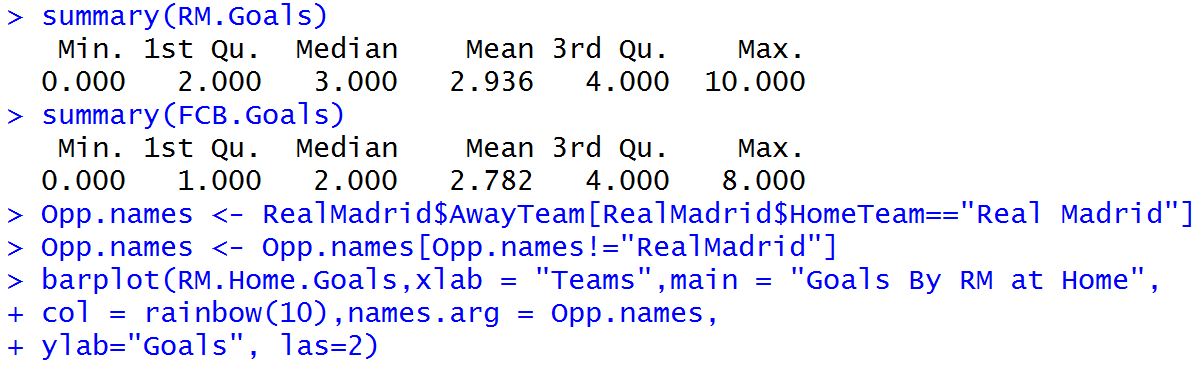
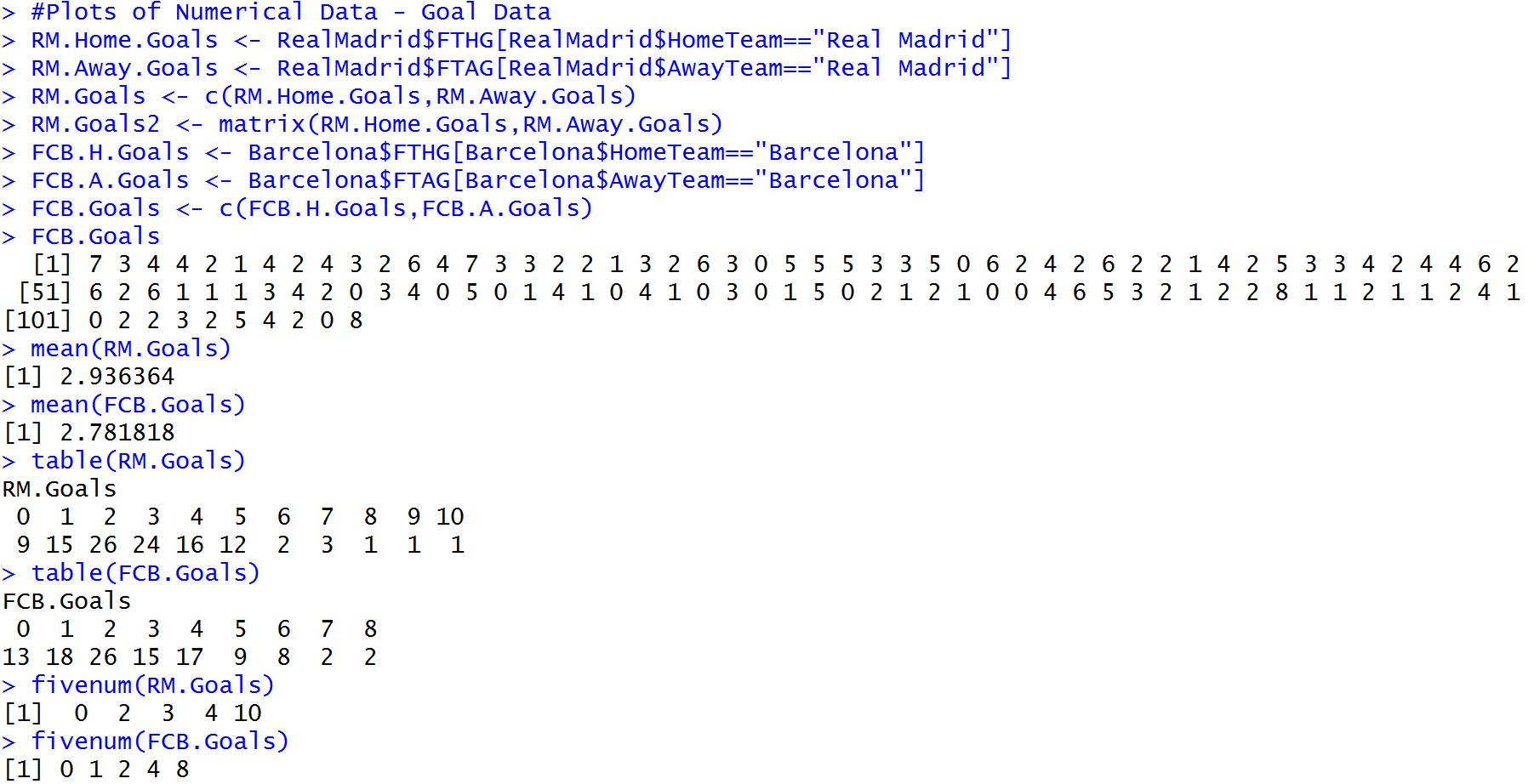


Now I am preparing data matrix of categorical data and Mosaic plots form that data.



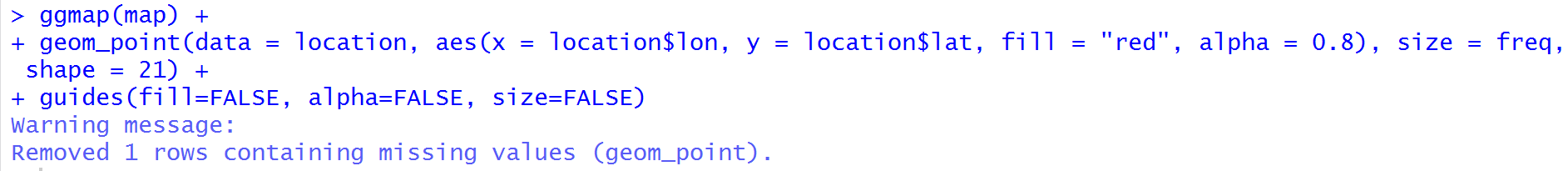
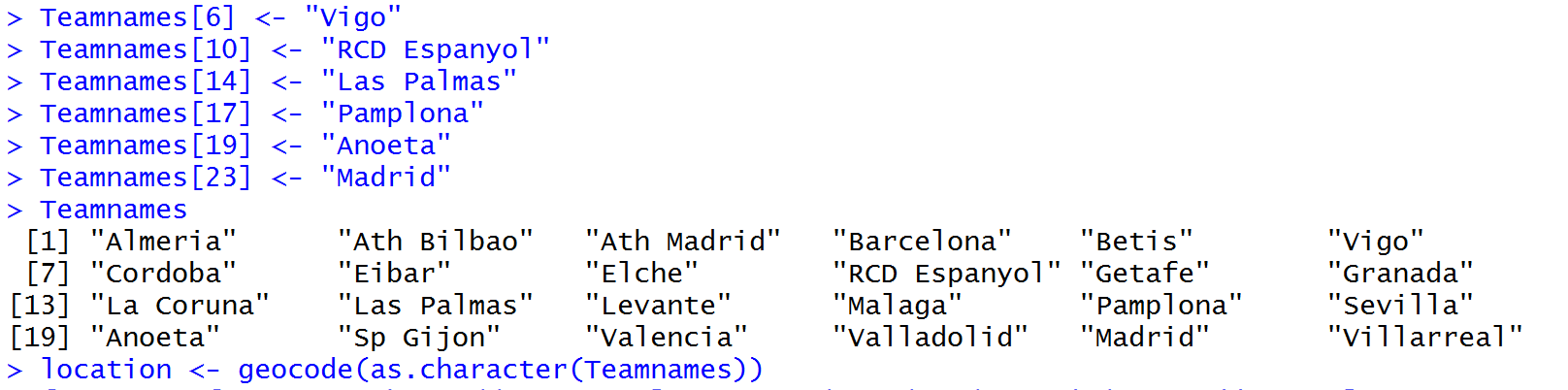
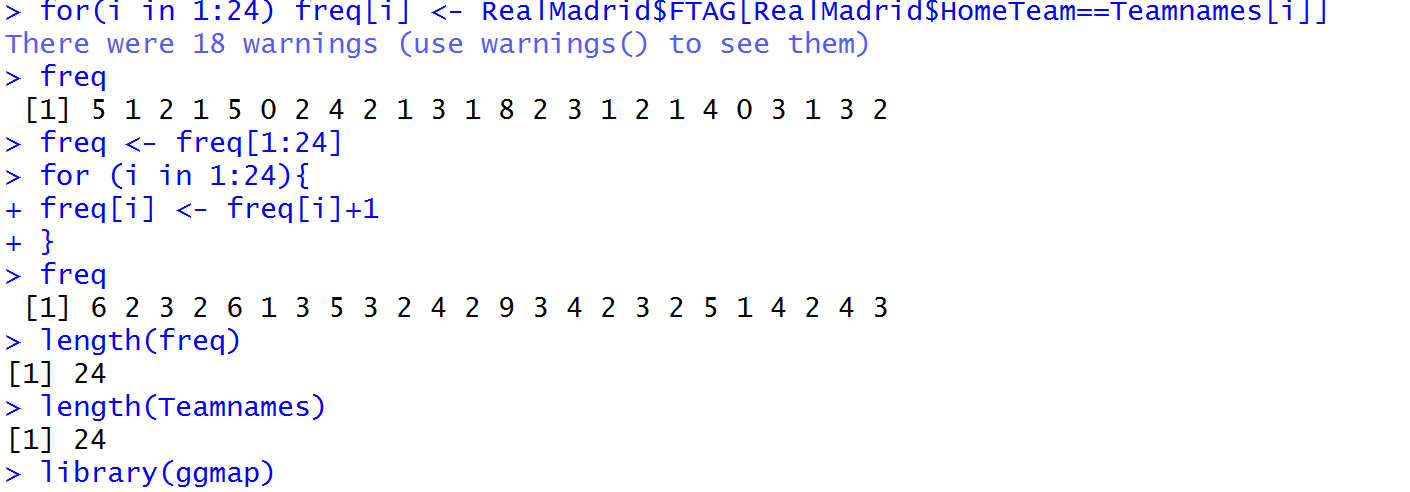
## Part 2: Pick one variable with numerical data and examine the distribution of the data.

I am selecting the Goals data of Real Madrid and Barcelona and comparing the distribution of goals.



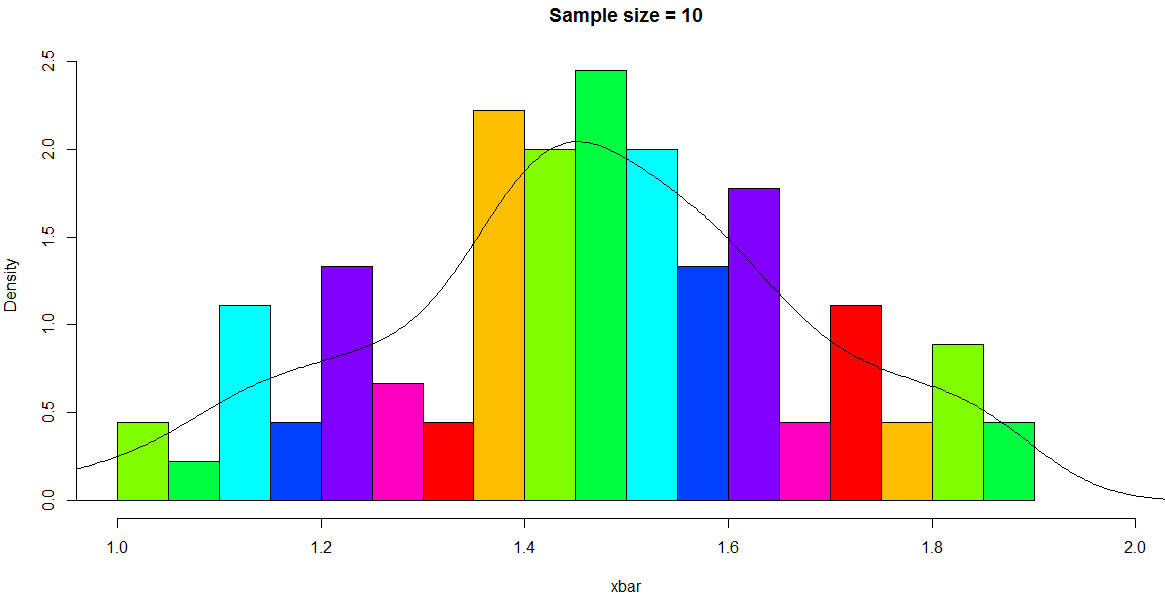
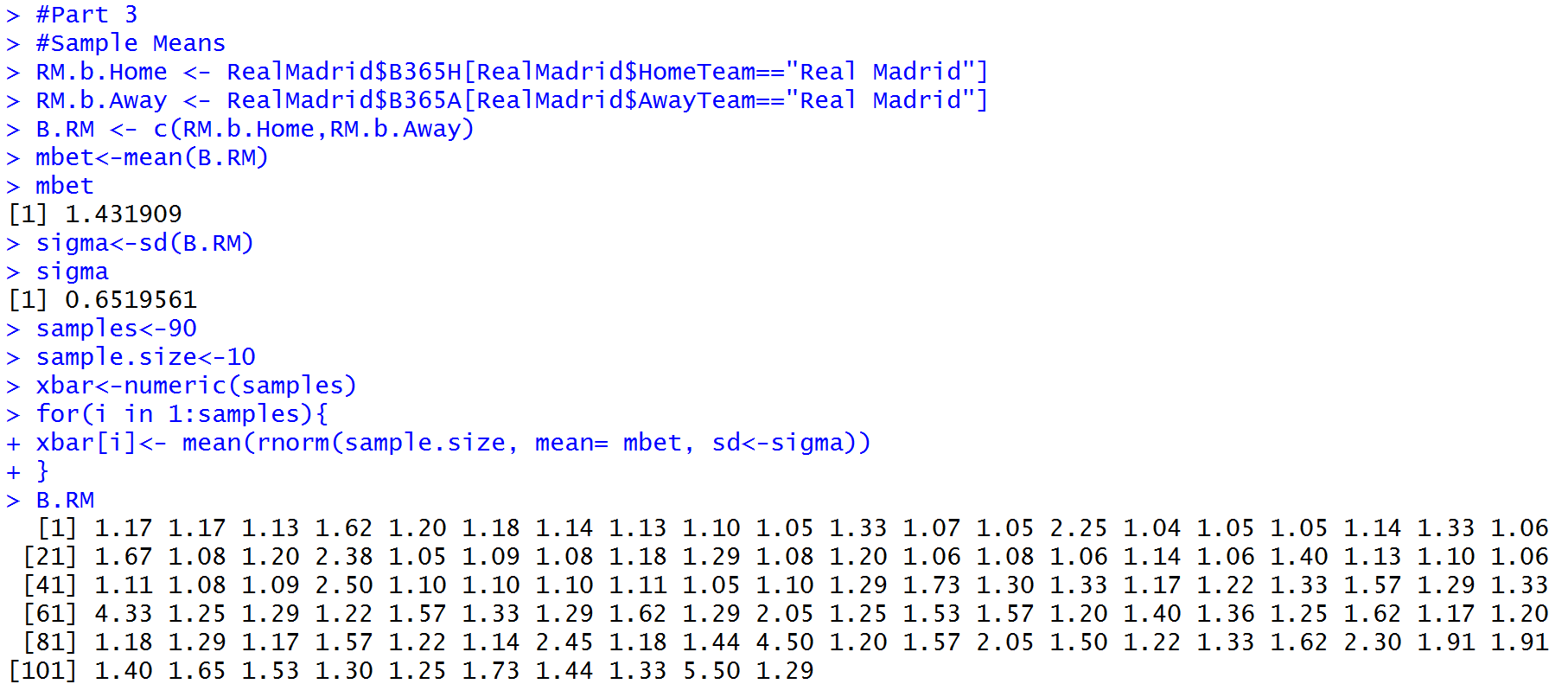
From the above Histogram we can see that Real Madrid has less number of 0 goals compared to Barcelona which means Real Madrid has better record of scoring at least one goals than Barcelona.

We can also see the geographical distribution of numerical data using GGMap function.

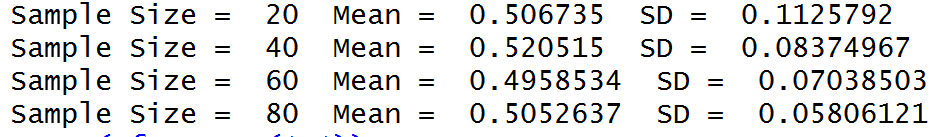
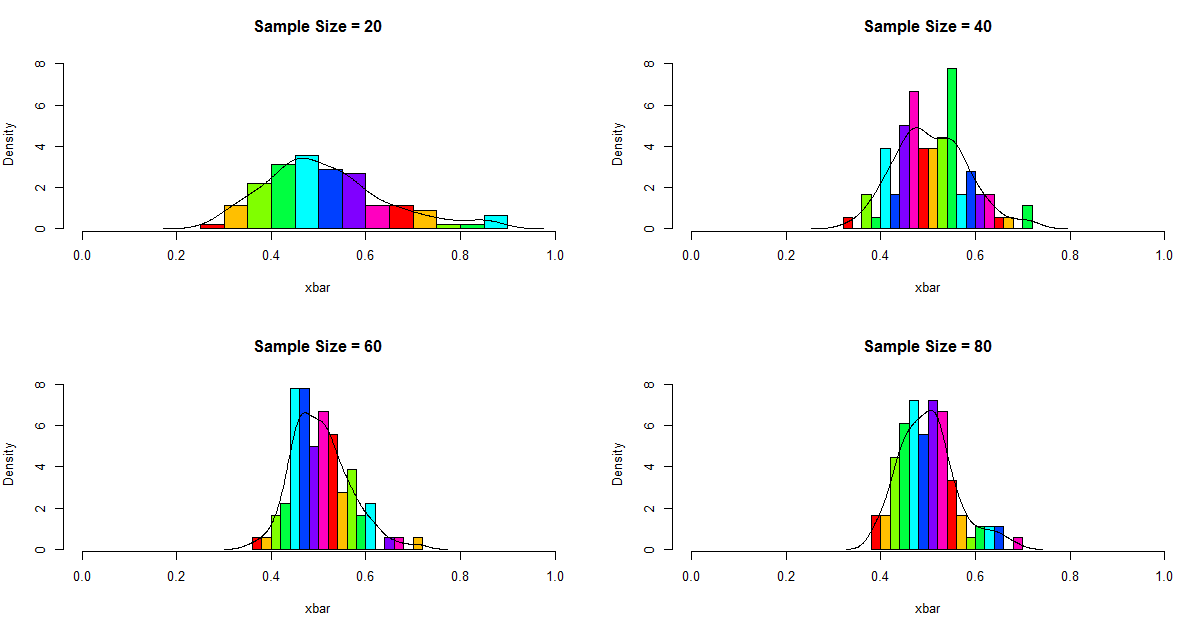


## Part 3: Draw various random Samples of the data and show the applicability of the Central Limit Theorem for this variable.

For the purpose of checking the distribution I selected the B365 odds of Real Madrid both Home and Away.

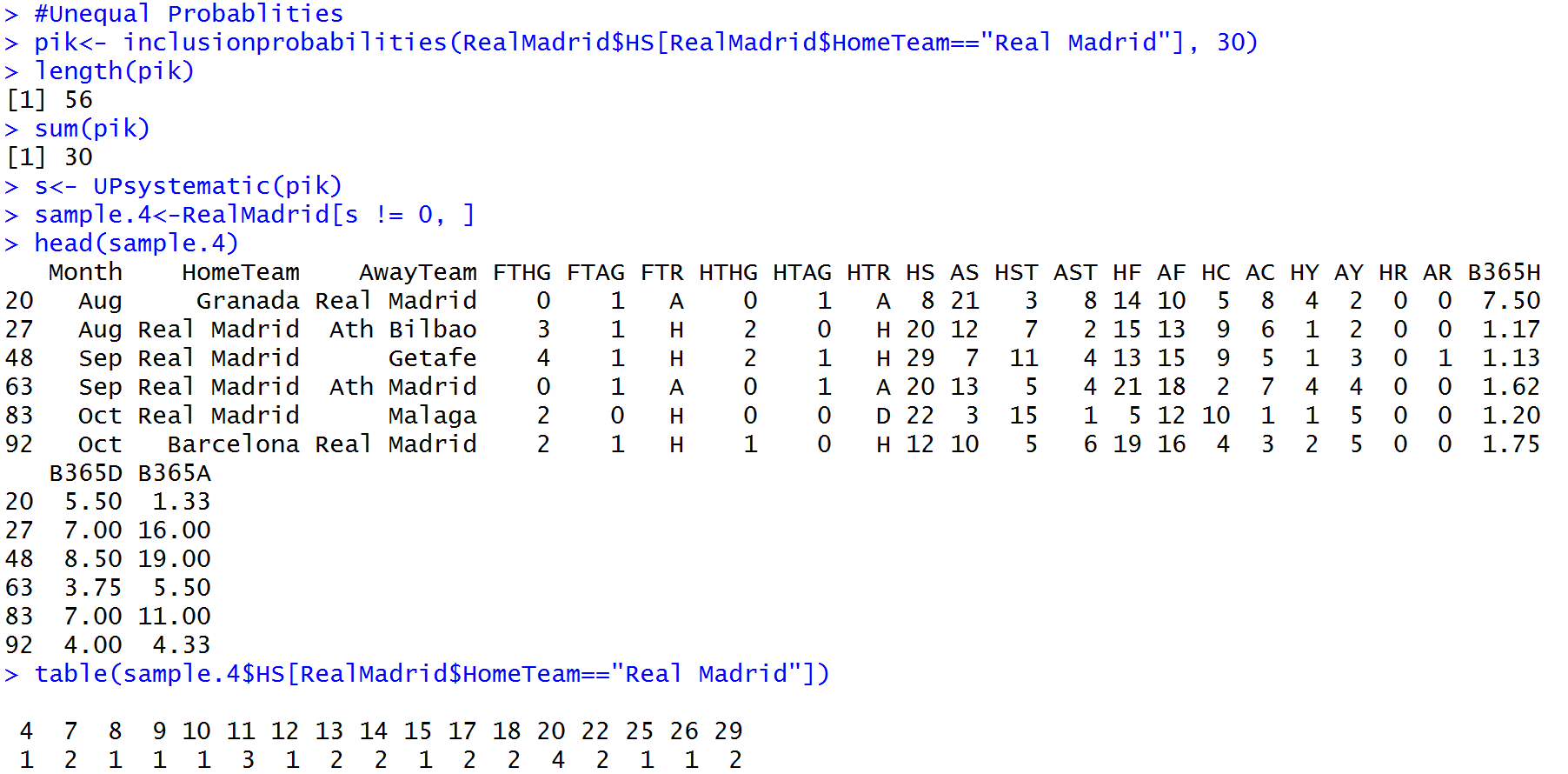
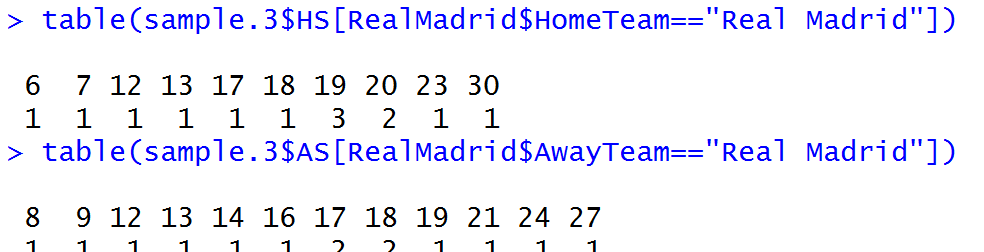
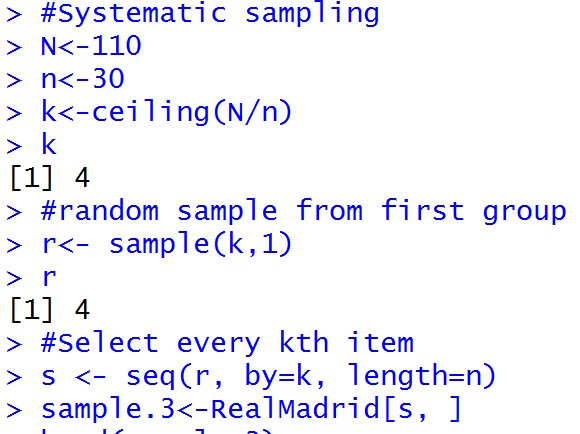
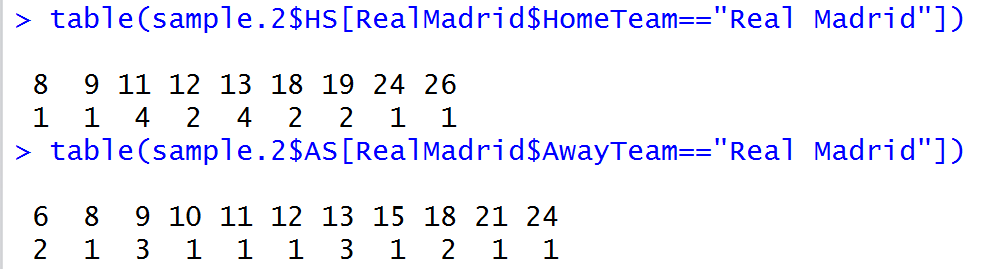
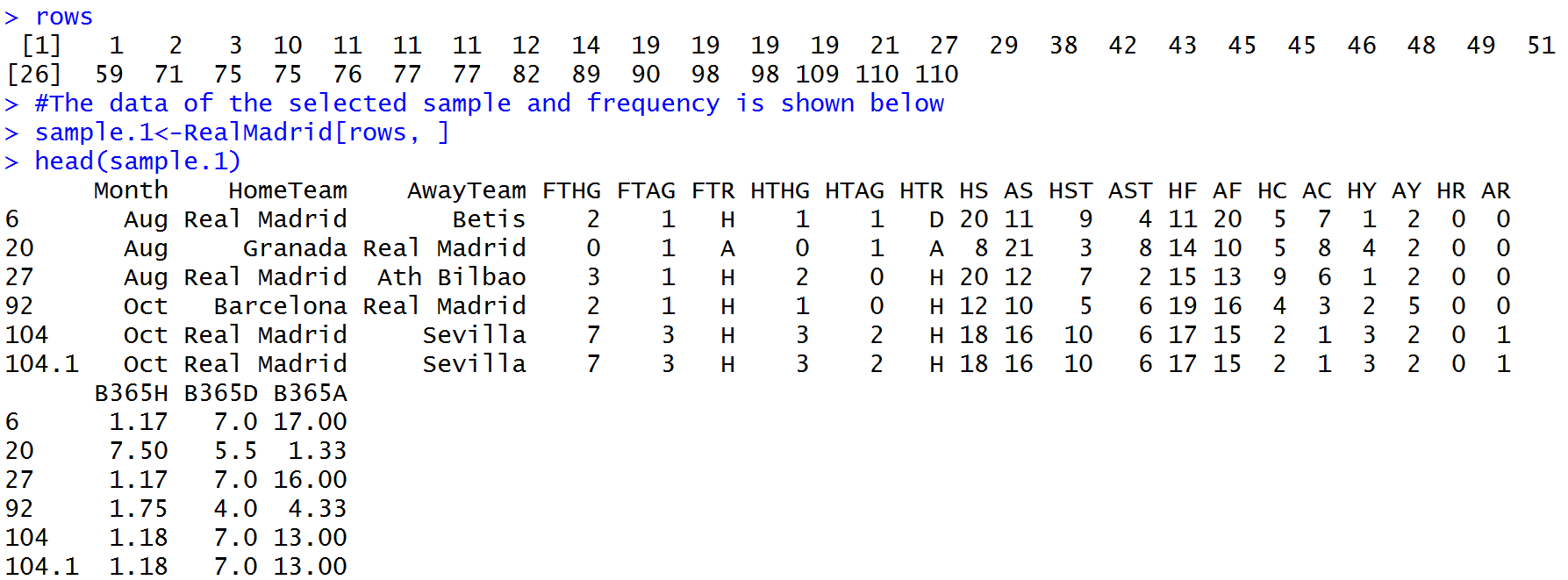
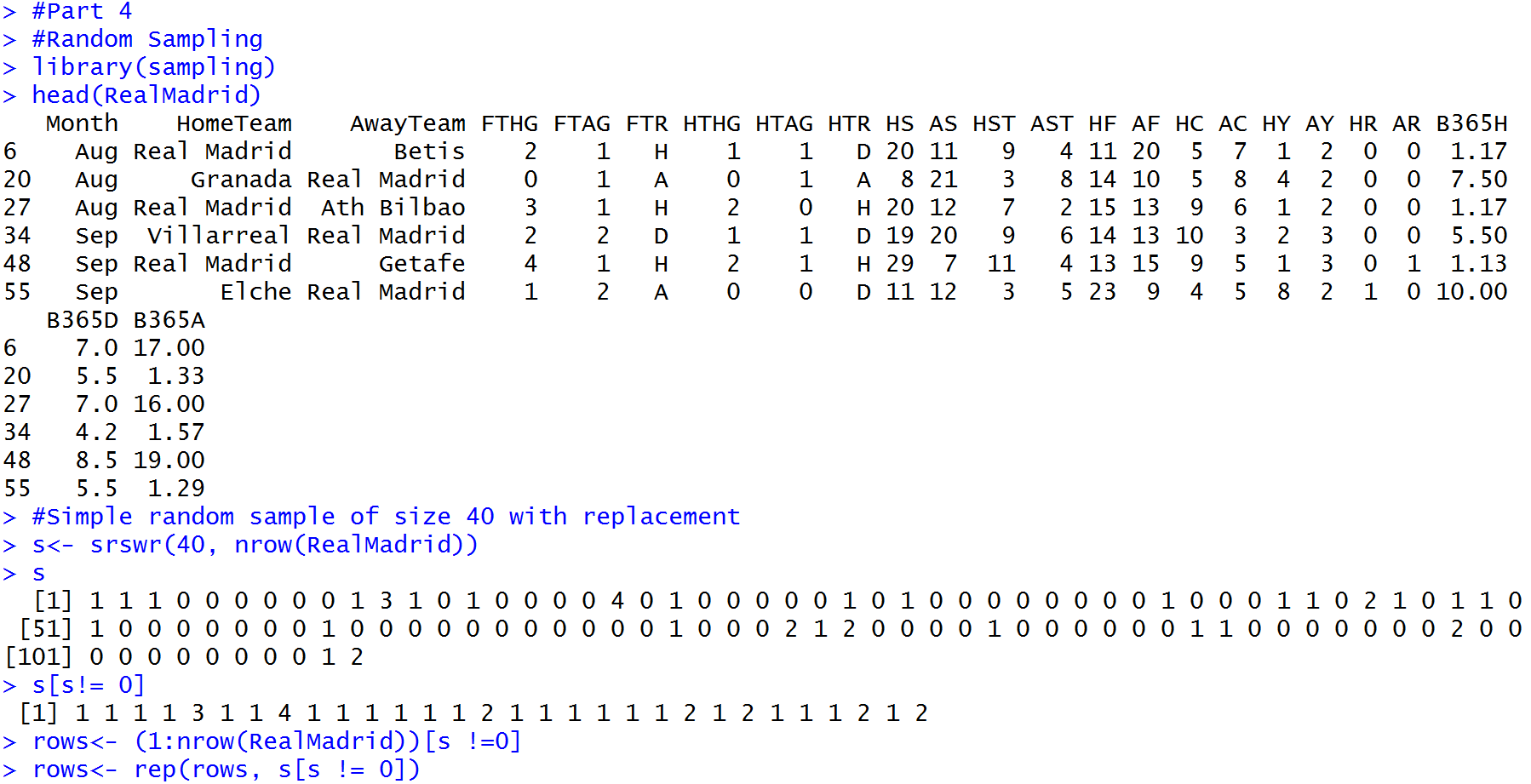


The graph follows Normal distribution. So Central Limit theorem can be applied. The following graph shows the distribution of means of sample sizes of 20, 40, 60, 80 and as the sample size increase we can see that the distribution is narrower.

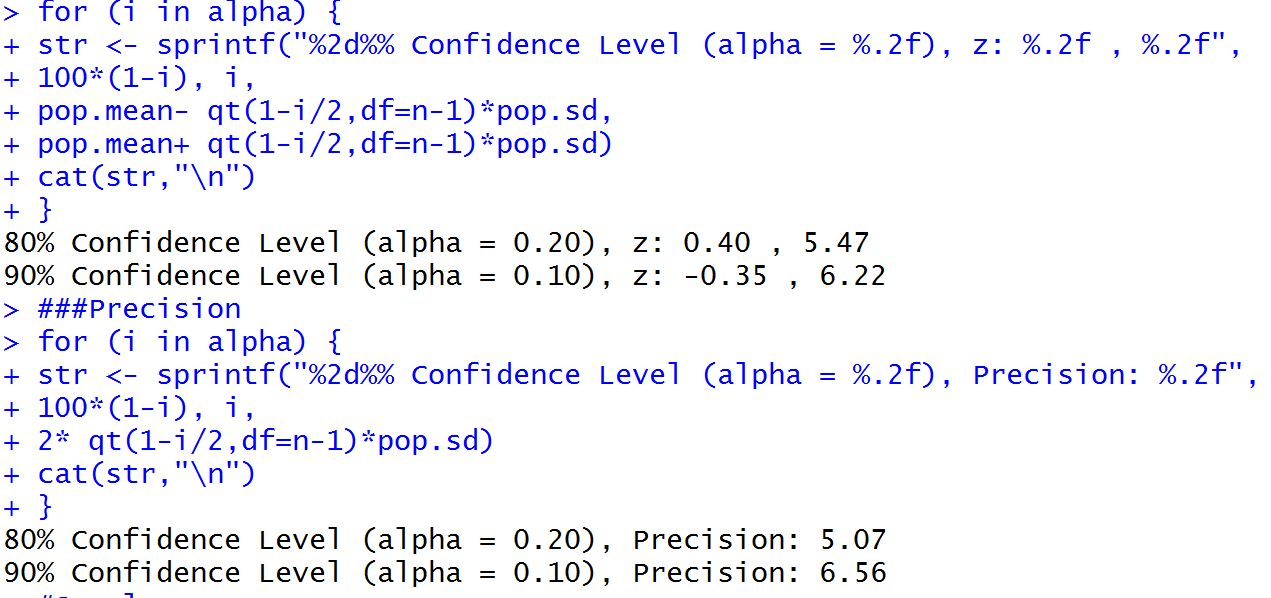
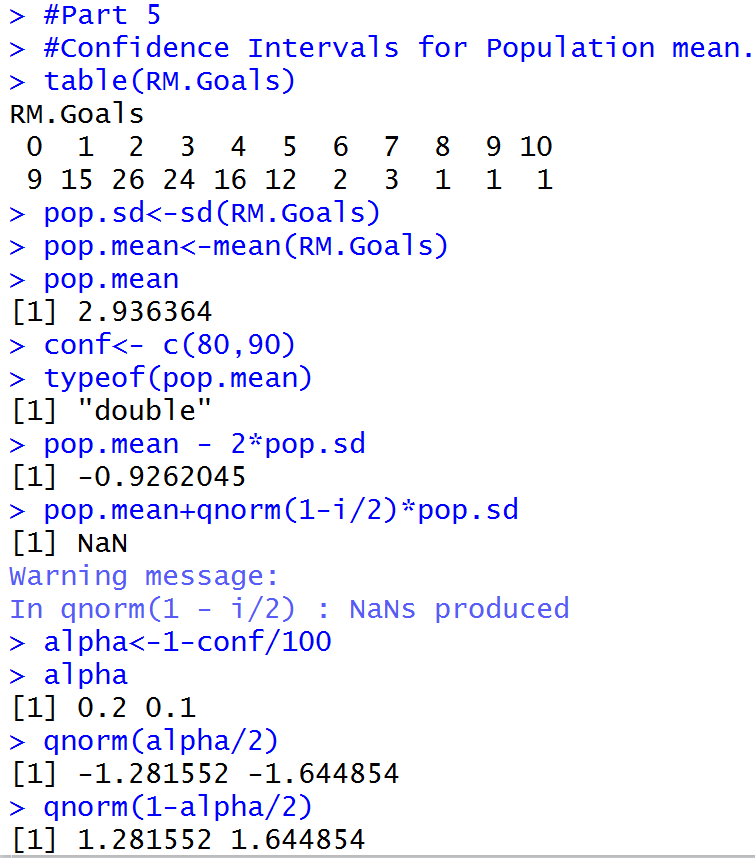


## Part 4: Sampling Methods:

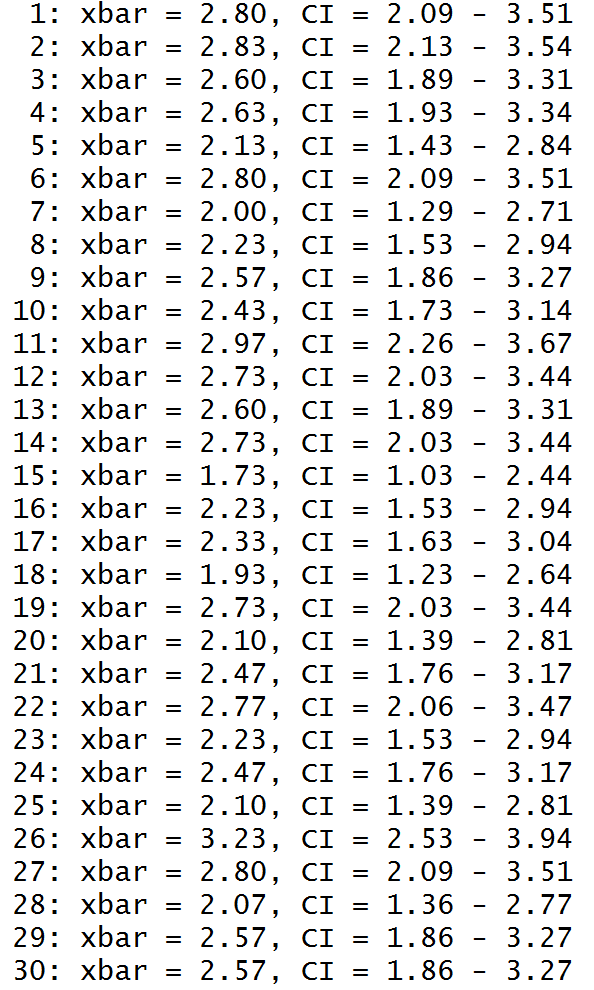
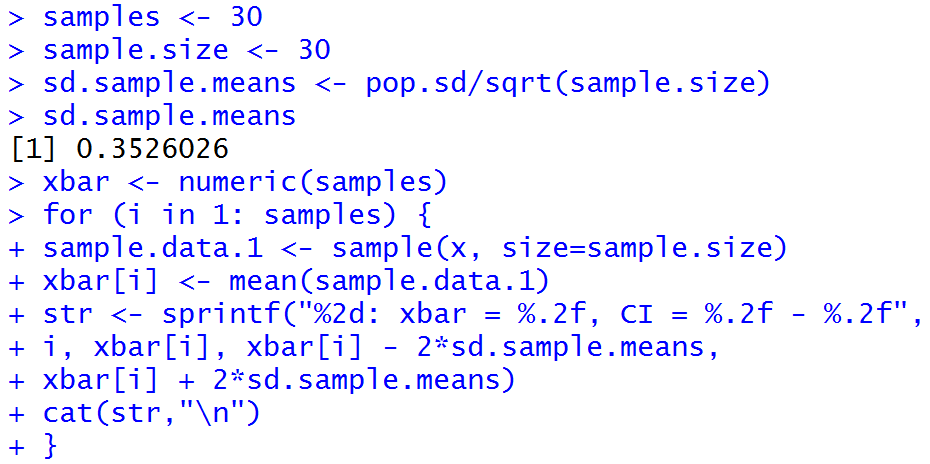
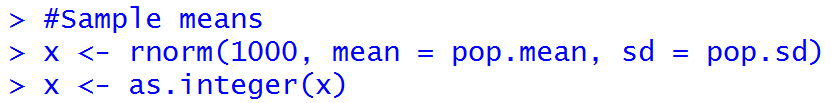
Below are various sampling techniques that I Performed on Bet365 data of the Real Madrid. I performed sampling techniques like Random Sampling with and without replacement, Systematic sampling and sampling of unequal probabilities.



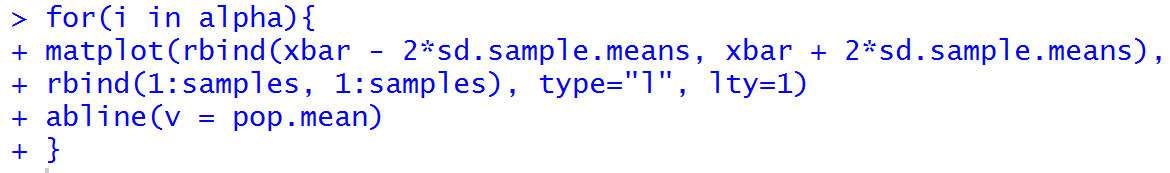
## Part 5: Confidence intervals of means of population and samples at confidence levels of 80 and 90.

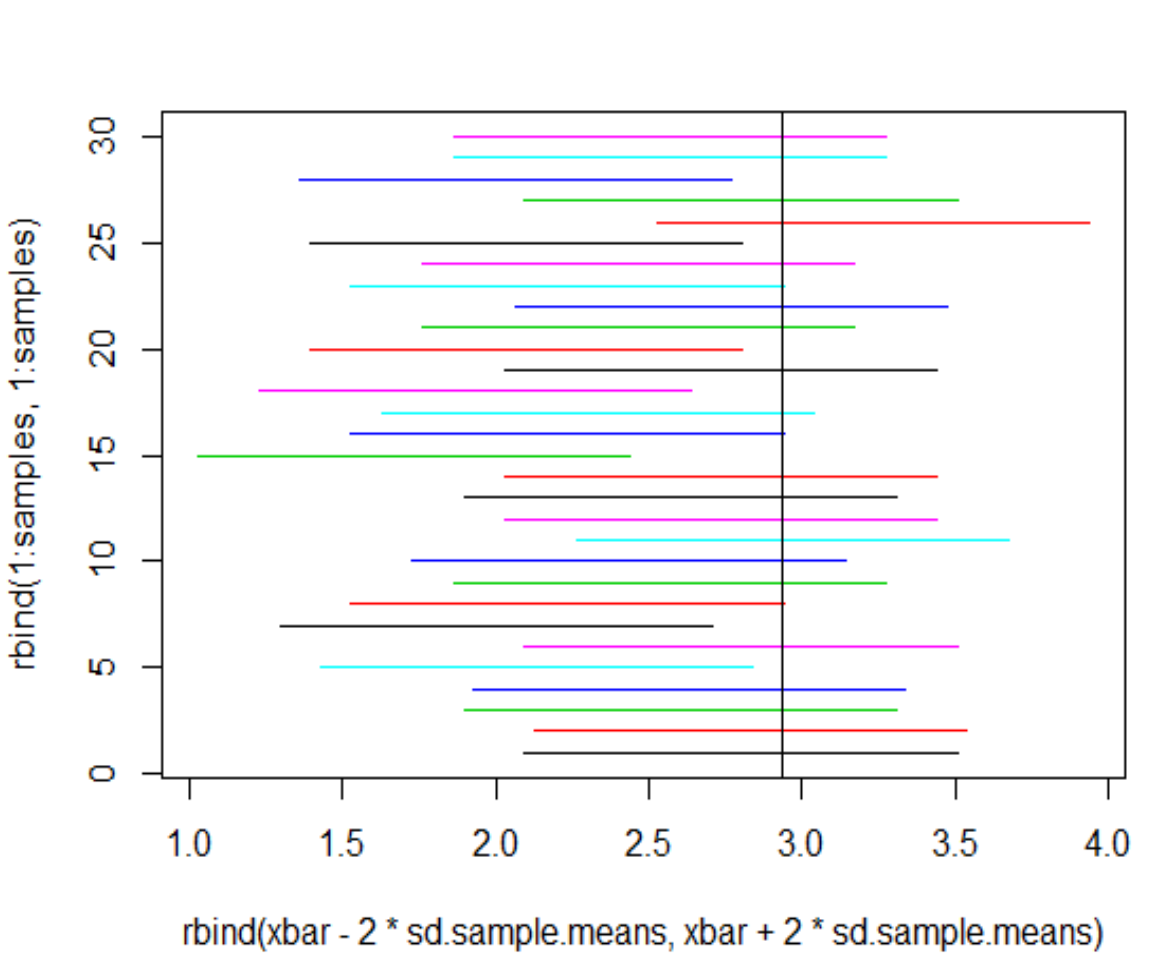


We can see that the precision increased with increase in confidence levels.



Now lets map the Matplot for confidence level of 90.





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:

* I learn the preprocessing of data like fixing the date attributes.
* I learned collecting the data from various sources and I understood different types of data.
* I understood various analytical techniques that can be performed on data and the power of Rstudio in data analysis.