INTERNATIONAL FOOTBALL MATCHES

Sarthak Gupta

sgupta2@albany.edu

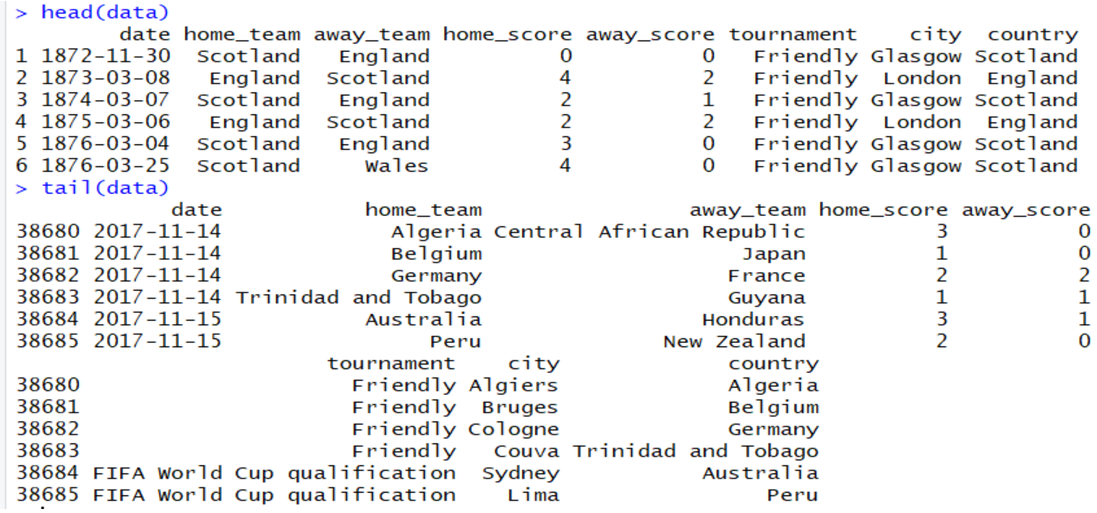
ABSTRACT

We have many international matches played across the world and we have lot of data related to these matches. The motivation behind the analysis being the dataset of the football matches that took place between 1872 to 2017. The analysis involves calculating the home goals scored by teams in different tournaments and calculating the likelihood of the dependency of home score in different tournaments. The result of a Bayesian analysis retains the uncertainty of the estimated parameters, which is very useful in decision analysis. The analysis involves the use of graph, histogram and tables to analyses the data and study the different columns available in the datasets.

DATASETS USED

- The data contains description about the matches held in between 1872-2017.The description involves information like the date of the match, the home and away goal score of the team and the tournament in which the matches were held.
- The data is gathered from several sources including but not limited to Wikipedia, fifa.com, rsssf.com and individual football associations' websites.
- The column involved in the data are:

1. Date
2. Home_Team
3. Away_Team
4. Home_Score
5. Away_Score .
6. Tournament
7. City
8. Country



```
> head(data)
        date home_team away_team home_score away_score tournament    city  country
1 1872-11-30  Scotland   England          0          0   Friendly Glasgow Scotland
2 1873-03-08   England  Scotland          4          2   Friendly  London  England
3 1874-03-07  Scotland   England          2          1   Friendly Glasgow Scotland
4 1875-03-06   England  Scotland          2          2   Friendly  London  England
5 1876-03-04  Scotland   England          3          0   Friendly Glasgow Scotland
6 1876-03-25  Scotland     Wales          4          0   Friendly Glasgow Scotland
> tail(data)
            date           home_team                away_team home_score away_score
38680 2017-11-14             Algeria Central African Republic          3          0
38681 2017-11-14             Belgium                    Japan          1          0
38682 2017-11-14             Germany                   France          2          2
38683 2017-11-14 Trinidad and Tobago                   Guyana          1          1
38684 2017-11-15           Australia                 Honduras          3          1
38685 2017-11-15                Peru              New Zealand          2          0
                     tournament    city             country
38680                  Friendly Algiers             Algeria
38681                  Friendly  Bruges             Belgium
38682                  Friendly Cologne             Germany
38683                  Friendly   Couva Trinidad and Tobago
38684 FIFA World Cup qualification  Sydney           Australia
38685 FIFA World Cup qualification    Lima                Peru
```

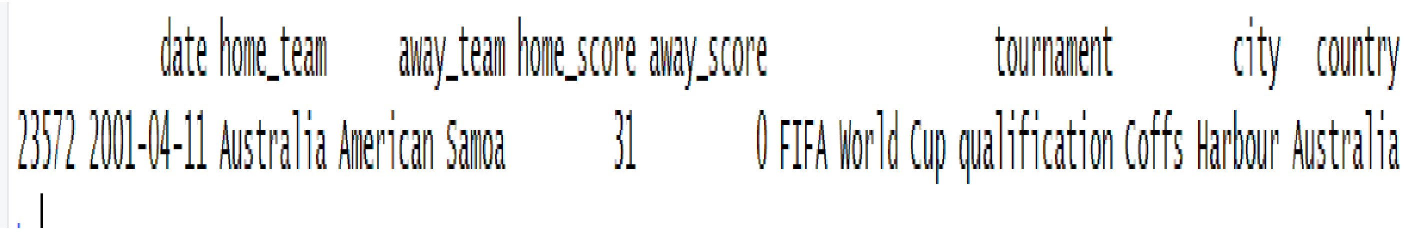
INSTALLATION OF RStudio

- Go to www.rstudio.com and click on the "Download RStudio" button.
- Click on "Download RStudio Desktop."
- Click on the version recommended for your system, or the latest Mac version, save the .dmg file on your computer, double-click it to open, and then drag and drop it to your applications folder.

PACKAGES INSTALLED

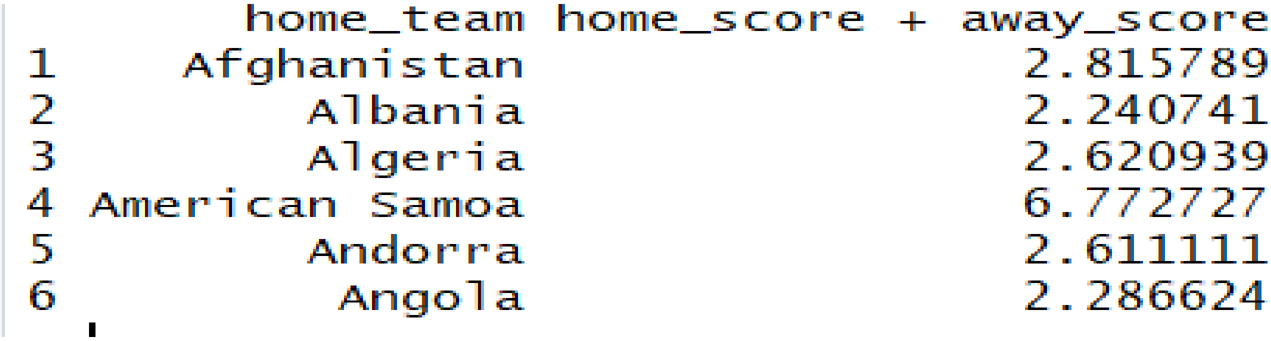
- Ggplot2
- Stringr
- XML
- Lubridate
- Plyr
- Useful
- Arm

ANALYSIS OF THE FOOTBALL



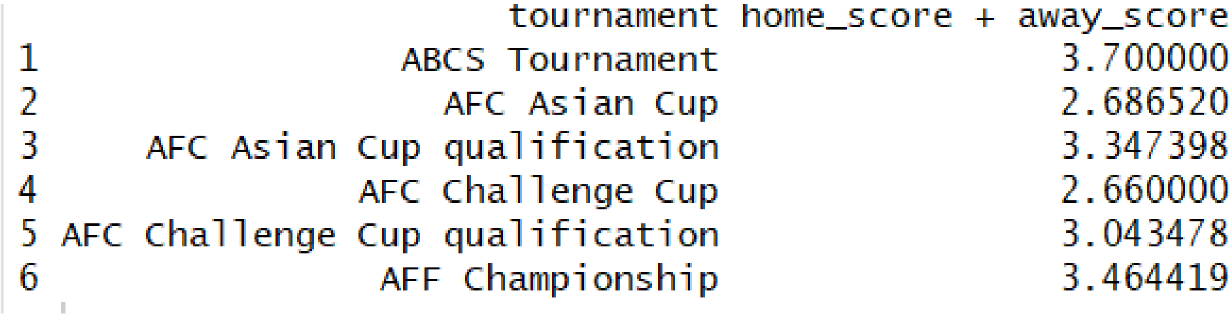
```
            date home_team      away_team home_score away_score                   tournament          city   country
23572 2001-04-11 Australia American Samoa         31          0 FIFA World Cup qualification Coffs Harbour Australia
```

- The above image shows us the team which has scored home scores.



```
       home_team home_score + away_score
1    Afghanistan                2.815789
2        Albania                2.240741
3        Algeria                2.620939
4 American Samoa                6.772727
5        Andorra                2.611111
6         Angola                2.286624
```

- The above image provides us the average scores scored by different home team.



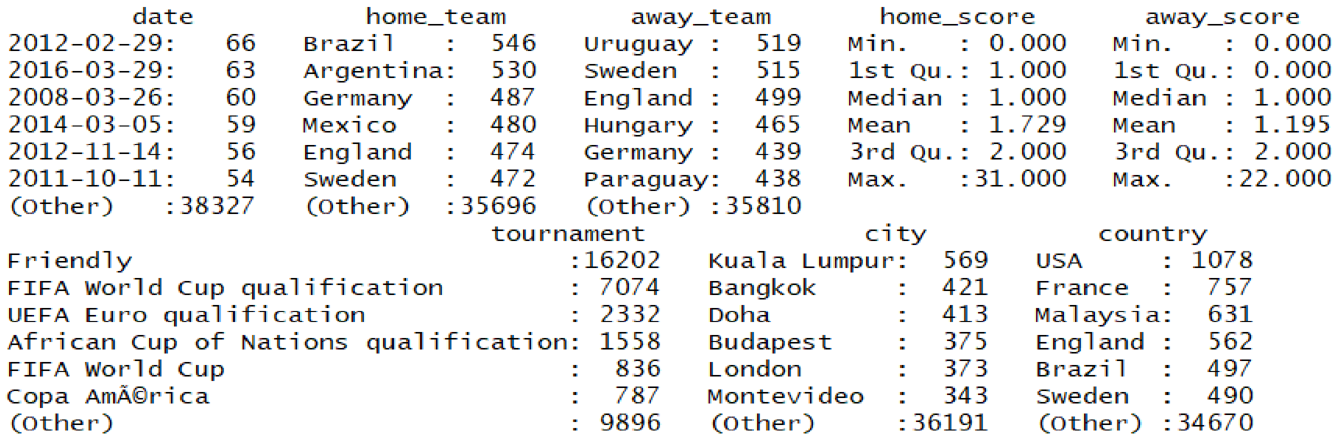
```
                     tournament home_score + away_score
1               ABCS Tournament                3.700000
2                 AFC Asian Cup                2.686520
3   AFC Asian Cup qualification                3.347398
4            AFC Challenge Cup                2.660000
5 AFC Challenge Cup qualification              3.043478
6              AFF Championship                3.464419
```

- The above image provides us the average scores scored by teams in different tournaments.

- `cor(new_data$home_score,new_data$away_score)`

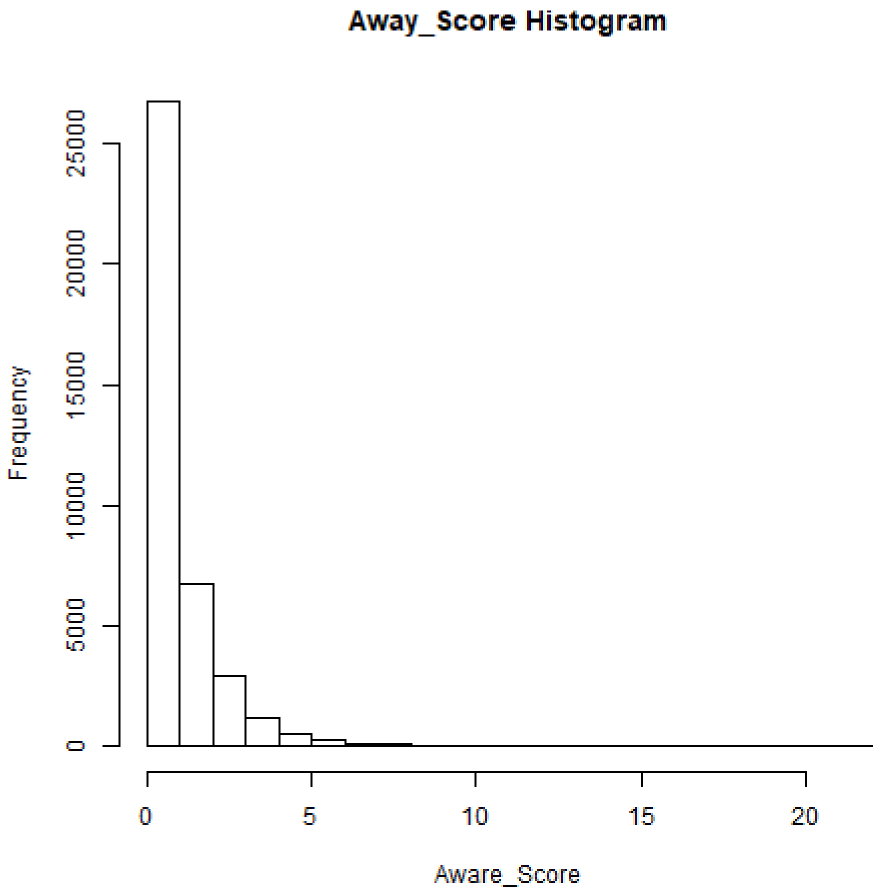
  -0.131301

- This negative sign shows that the home and away score are opposites to each other means that when the teams scores at ho me then they do not score that well at away city. That means that the home team are advantage at scoring at home.

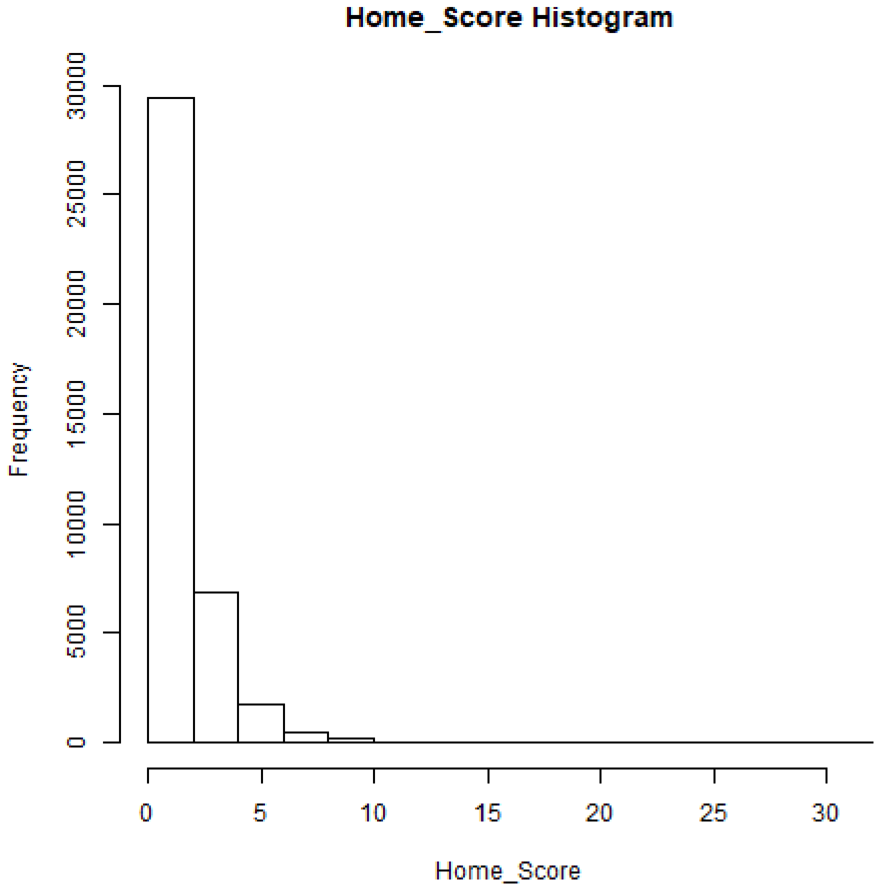


```
          date              home_team           away_team          home_score         away_score    
 2012-02-29:    66   Brazil    :   546   Uruguay :   519   Min.    : 0.000   Min.    : 0.000  
 2016-03-29:    63   Argentina:   530   Sweden  :   515   1st Qu.: 1.000   1st Qu.: 0.000  
 2008-03-26:    60   Germany   :   487   England :   499   Median : 1.000   Median : 1.000  
 2014-03-05:    59   Mexico    :   480   Hungary :   465   Mean   : 1.729   Mean   : 1.195  
 2012-11-14:    56   England   :   474   Germany :   439   3rd Qu.: 2.000   3rd Qu.: 2.000  
 2011-10-11:    54   Sweden    :   472   Paraguay:   438   Max.   :31.000   Max.   :22.000  
 (Other)    :38327   (Other)  :35696   (Other) :35810                                        
                                   tournament                city               country     
 Friendly                               :16202   Kuala Lumpur:   569   USA     : 1078  
 FIFA World Cup qualification           : 7074   Bangkok     :   421   France  :   757  
 UEFA Euro qualification                : 2332   Doha        :   413   Malaysia:   631  
 African Cup of Nations qualification: 1558   Budapest    :   375   England :   562  
 FIFA World Cup                         :  836   London      :   373   Brazil  :   497  
 Copa AmÃ©rica                          :  787   Montevideo  :   343   Sweden  :   490  
 (Other)                                : 9896   (Other)     :36191   (Other) :34670  
```

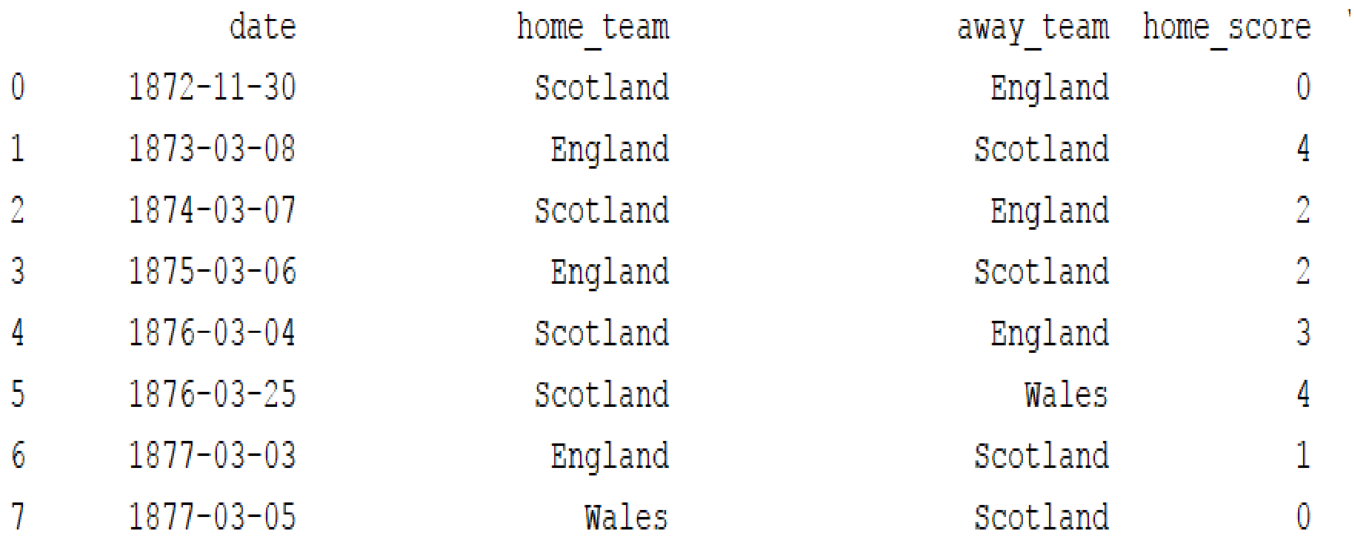
- The above shows us the summary of the data which calculate summary statistics, including the mean, standard deviation, range, and percentiles.



- The above histogram gives us the frequency of the away scores across all the tournaments. As you can see that the maximum number of away goals that has been scored are around 0 to 3. It means that the maximum number of teams scores goals ranging from 0-3.

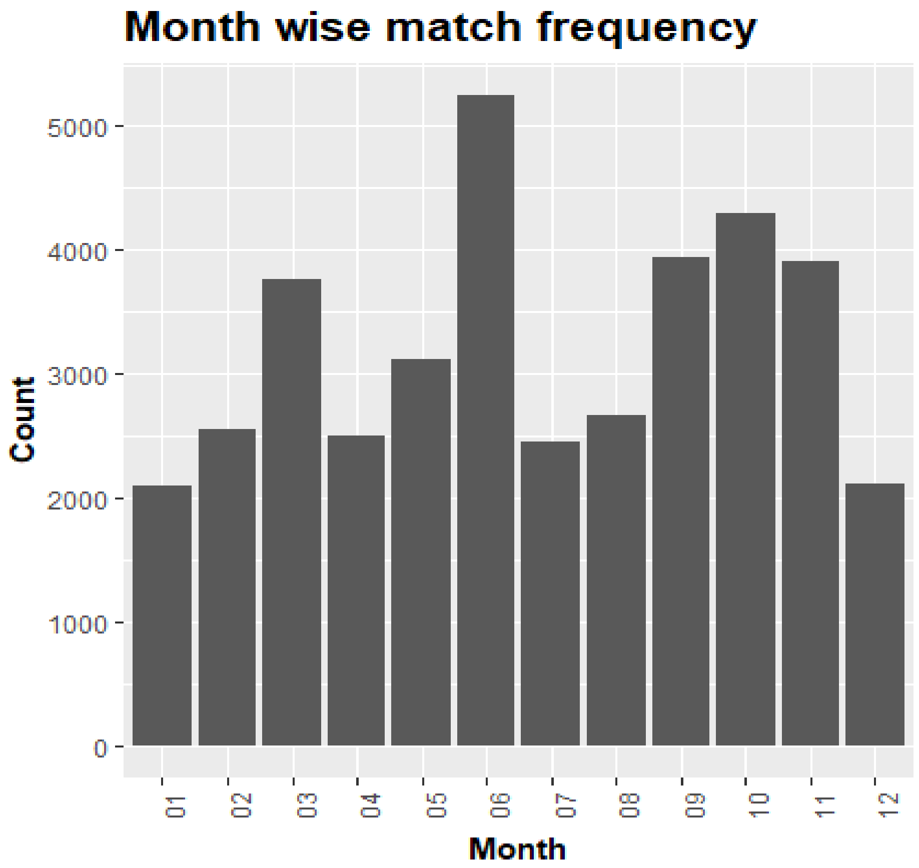


- The above histogram gives us the frequency of the home scores across all the tournaments. As you can see that the maximum number of home goals that has been scored are around 0 to 3. It means that the maximum number of teams scores goals ranging from 0-3.

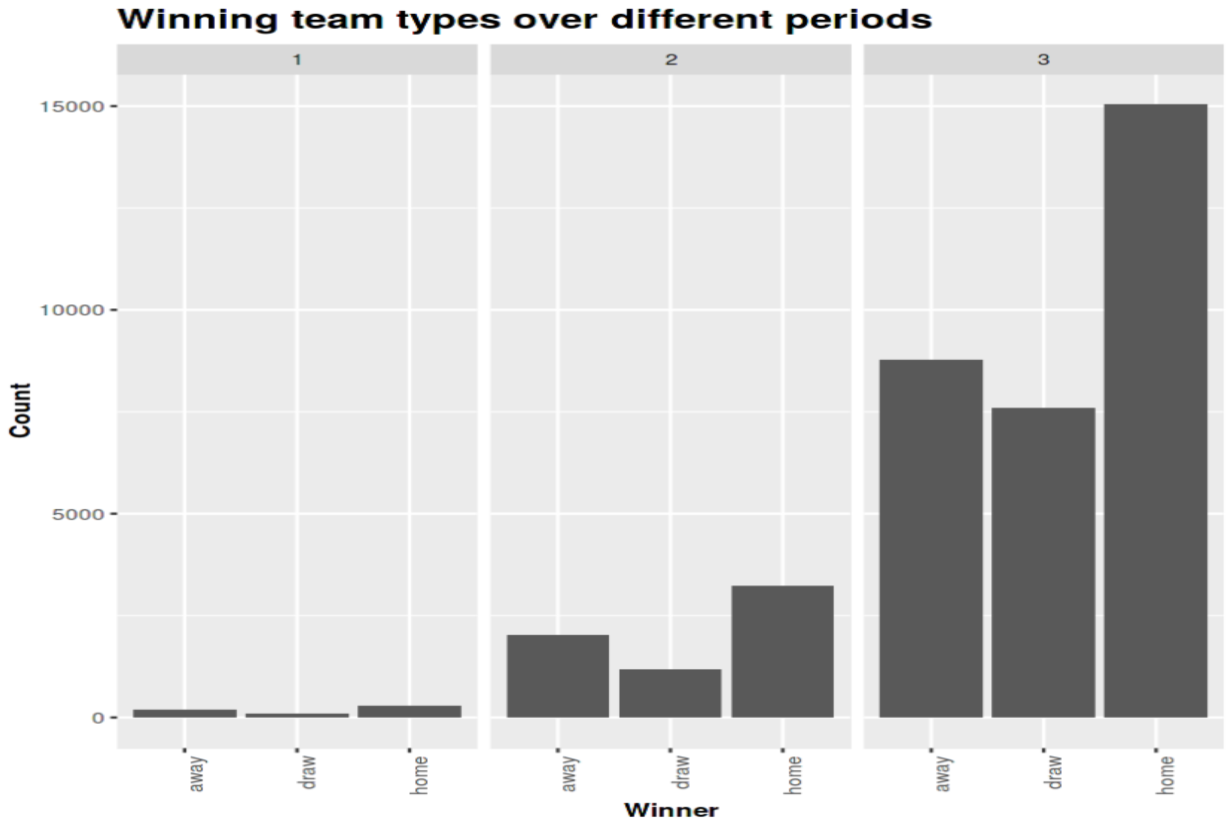


```
         date   home_team   away_team  home_score
0  1872-11-30    Scotland     England           0
1  1873-03-08     England    Scotland           4
2  1874-03-07    Scotland     England           2
3  1875-03-06     England    Scotland           2
4  1876-03-04    Scotland     England           3
5  1876-03-25    Scotland       Wales           4
6  1877-03-03     England    Scotland           1
7  1877-03-05       Wales    Scotland           0
```

- The above shows us the total home scores scored by teams in different year.

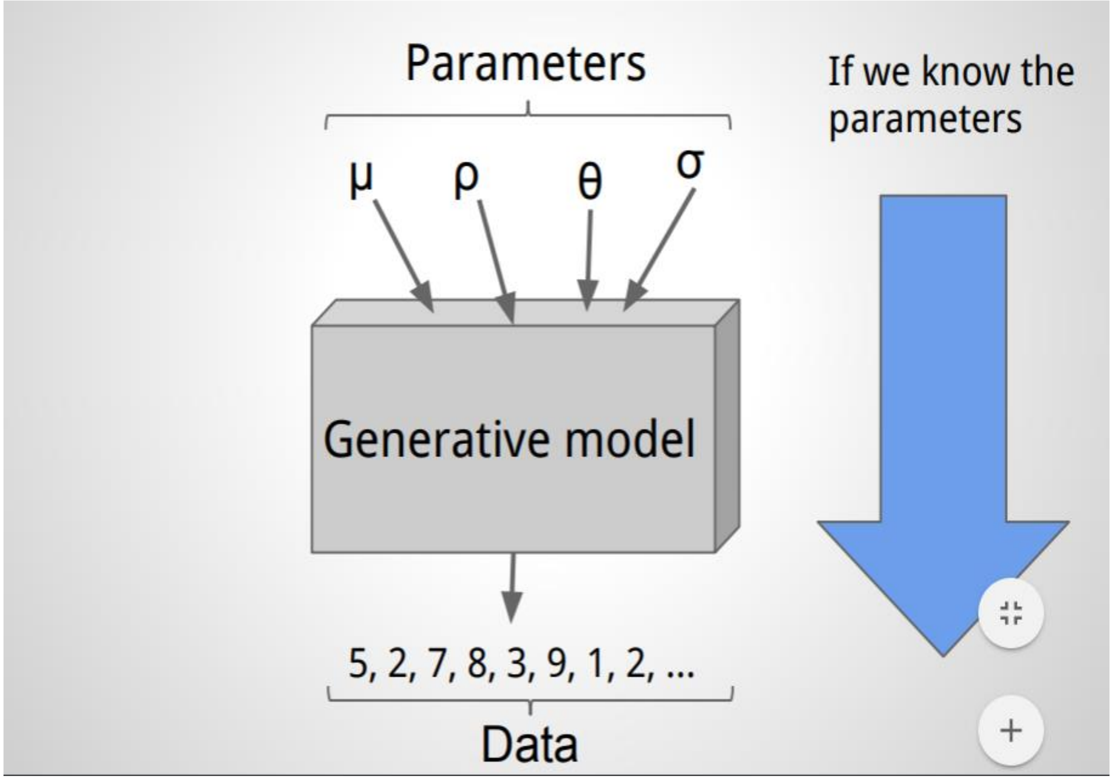


- The above graph gives us the count of matches played in different months.



- The graph shows us the frequency of teams winning over different periods. You can see that the maximum goals are scored in the third-time period i.e. last 50 years. The time periods are divided in 3-time periods like first 50, second 50 and the last 50 years.

BAYESIAN DATA ANALYSIS

- It is when you use probability to represent uncertainty in all parts of a statistical model.
- A flexible extension of maximum likelihood.
- Potentially the most information-efficient method to fit a statistical model.
- 

Parameters

If we know the parameters

Generative model

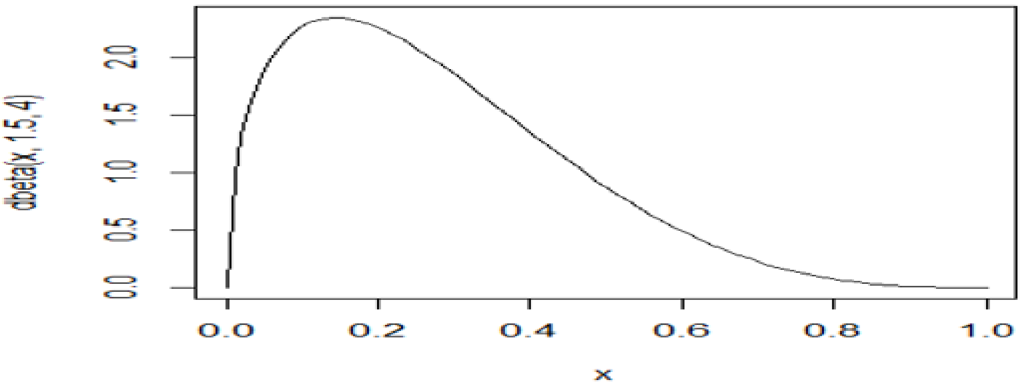
5, 2, 7, 8, 3, 9, 1, 2, ...

Data

Problem Statement:

To find the likelihood of the teams scoring home goals and away scores in different tournaments.

- Step 1-
- **Specifying the Prior for a proportion**
- For this I took the three quantiles as 50%,99,99%,0.001%
- Using (LearnBayes) package in R we will get the prior proportion and then plot the prior density graph.



The above graph shows us the value of around 0.1-0.2 having the maximum.

➢ Step 2-

- **Calculating the Likelihood Function**

➢ Step 3-

- **Calculating the Posterior Distribution**

- **Using the function glm ()**

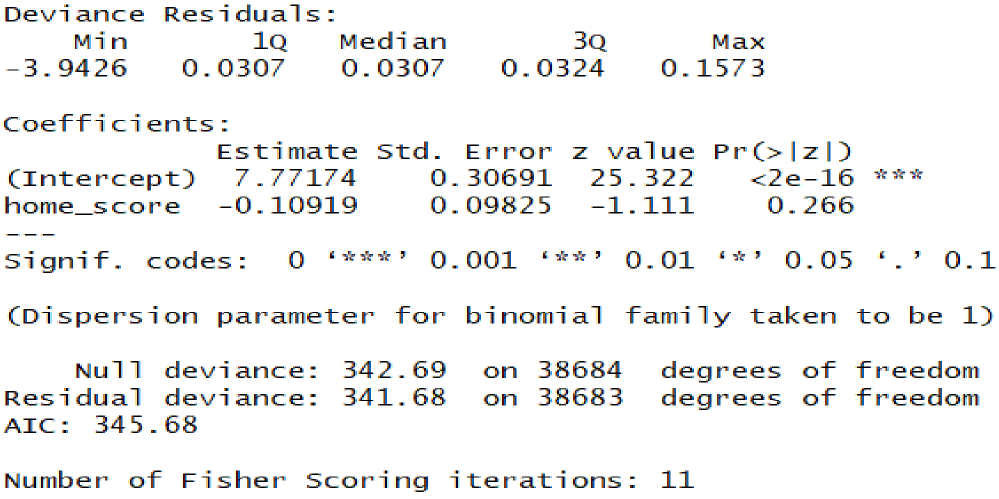
- Bayesian logistic regression. In the arm (Applied Regression and Multilevel modeling) package

- I Replaces glm (), estimates are more numerically and computationally stable I

- Student-t prior distributions for regression coefficients.

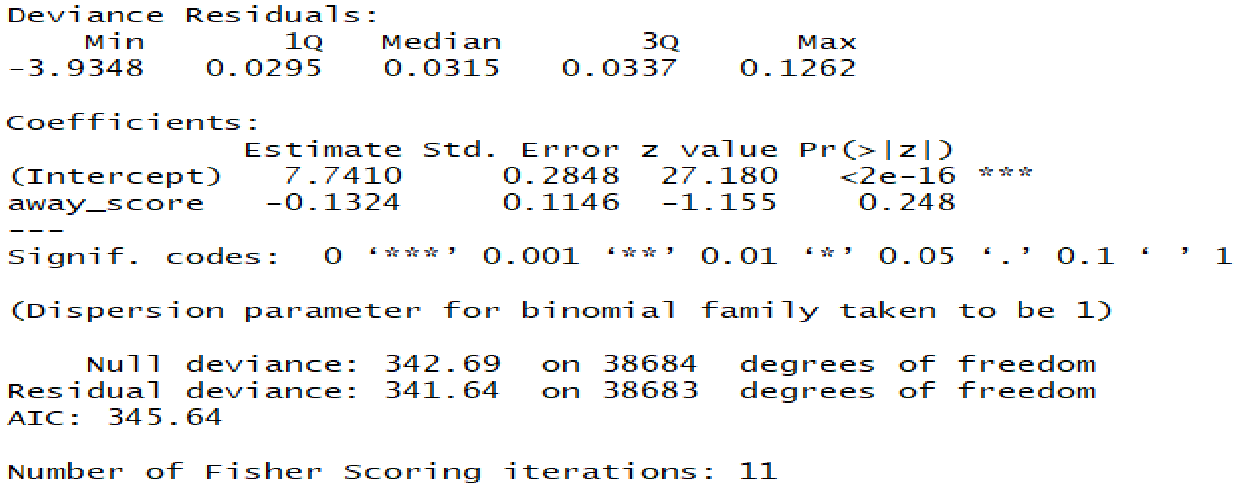
- We went inside glm.fit to augment the iteratively weighted least squares step

- *bayesglm(formula = tournament ~ away_score, family = "binomial", data = new_data)*



```
Deviance Residuals:
    Min       1Q    Median       3Q      Max
-3.9426   0.0307   0.0307   0.0324   0.1573

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  7.77174    0.30691  25.322   <2e-16 ***
home_score  -0.10919    0.09825  -1.111    0.266
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 342.69  on 38684  degrees of freedom
Residual deviance: 341.68  on 38683  degrees of freedom
AIC: 345.68

Number of Fisher Scoring iterations: 11
```

- The above shows us the values of home score versus tournaments using Bayesian generalized linear models.
- The value that can be read from the above image is estimated standard deviation, error estimation value.
- We use Student-t prior distributions for the coefficients. The prior distribution for the constant term is set so it applies to the value when all predictors are set to their mean values.



```
Deviance Residuals:
    Min       1Q    Median       3Q      Max
-3.9348   0.0295   0.0315   0.0337   0.1262

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   7.7410     0.2848  27.180   <2e-16 ***
away_score   -0.1324     0.1146  -1.155    0.248
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 342.69  on 38684  degrees of freedom
Residual deviance: 341.64  on 38683  degrees of freedom
AIC: 345.64

Number of Fisher Scoring iterations: 11
```

RESULTS

	Prior	Likelihood	Posterior
Mean	**0.27272727**	**0.030733838**	**0.03074398**
Mode	**0.14285714**	**0.03070957**	**0.030719722**
Standard Deviation	**0.1746852**	**0.00877489**	**0.008775897**

- The tables show the prior, likelihood and posterior probabilities for the given datasets which explains the dependency of home scores and away score in different tournaments.

CONCLUSION

The research on dataset gave me very notable observations like we calculated home scores, away score and dependency of tournaments and teams on the goals. Also, the correlation between the various features in the dataset and the data being very user specific gives us variations and allow us in using different models.

FUTURE WORK

As my research work is limited only to tournaments and the scores, we would like to extend it further in taking the team scores individually and predicting the scores between two teams. Also, the approach can be used on different dataset and that might give us more interesting observations.

ACKNOWLEDGMENTS

My thanks to Professor Knuth for giving me this opportunity to play with large datasets and also use Bayesian approaches.

REFERENCES

1. Gelman, Andrew, et al. 2014. *Bayesian Data Analysis*. 3rd Edition. Boca Raton, FL: CRC Press.
2. Enjoyable **historical context**: McGrayne, Sharon Bertsch. 2012. *The Theory That Would Not Die*. New Haven: Yale University Press.
3. https://www.rdocumentation.org/packages/arm/versions/1.9-3/topics/bayesglm
4. http://evolution.gs.washington.edu/gs560/2011/lecture7.pdf