JOINT PROJECT REPORT

Soccer Match Score Prediction Model

Stochastic Processes – MC303

Batch: MC G2

Sreya Majumder

2K19/MC/127

sreyamajumder_2k19mc127@dtu.ac.in

Tanishka Singh

2K19/MC/129

tanishkasingh_2k19mc129@dtu.ac.in

Highlights of the Project:

1. We have proved using various techniques that Premier League goal scoring fits the characteristics of a Poisson process. Our first result was that a Poisson distribution can be used to predict the number of matches with each number of goals scored.
2. Additionally, we found that the time between each individual goal in a season can be described by an exponential distribution. We also have evidence that the goal scoring time positions after being standardized are uniformly distributed.
3. We also used the data to predict what would happen in 2018-19. We got each team's goal scoring rate at home and away from home by doing Poisson regression, and then performed simulations using those rate parameters. Different team metrics like how many points each team got and what place each team finished were being kept track of from the simulations.
4. We extracted data for all the possible combinations of teams that played the EPL in last season and stored the results in an excel. Using PowerBi, we developed a web-report that will concisely help us view the results.
5. **Contributions:**
6. Sreya – I was responsible for data importing and tidying, proving the relationship of the time between goals and exponential distribution. I also verified the correlation among scoring time and uniform distribution and also helped in the prediction model. Tanishka – I demonstrated the relationship between goal scoring and the poisson process. I was responsible for making the model used to predict 2018-2019 seasons result from prior data. The final PowerBi web-report that helped us view the results concisely was also made by me.

Table of Content

Abstract

The English Premier League is well-known for being not only one of the most-watched football leagues in the world but also one of the toughest competitions to predict. The purpose of this project was to look at goal scoring data of the English Premier League and use statistical modeling to predict Premier League match results. This project attempts to determine whether goal scoring in the Premier League can be modeled by a Poisson process, specifically, the relationships between the number of goals and the Poisson distribution. It turns out to be that the Poisson process performs a great job of describing Premier League goal scoring. In addition to this, various models of predicting a Premier League season's results will be observed, with a large number of simulations being involved in each method.

Introduction

Often referred to as the king of sport, football can be played almost anywhere, from grass fields to indoor gyms, streets, parks, or beaches, due to the simplicity in its principal rules and essential equipment. Europe is known to be the birthplace of modern football. The Old Continent is home to numerous top-level professional football leagues, and the English Premier League (EPL) is certainly the best of them all.

The EPL was founded in 1992, and over the last three decades, we have witnessed numerous memorable matches and countless outstanding performances by clubs and their players. The EPL is currently a competition of twenty English football clubs. At the end of each season, the bottom three teams get relegated to the second-highest division of English football, in exchange for three promoted teams. A Premier League season usually takes place from mid-August to mid-May. Each team gets to play every other team twice, once at home and once on the road, hence there are a total of thirty-eight fixtures in a season for each team.

The most important aspect of the game of football is scoring goals. The rule is very simple: in order to win, you must score more than your opponent. In the Premier League, each match happens within the span of 90 minutes (plus stoppage time), and the match consists of two 45-minutes halves. Each team can get one of these three results after each match: a win, a draw, or a loss. If there’s a draw, the two clubs receive a point apiece, and for non-drawing matches, the winner is rewarded with three points and the losing team gets punished with zero points. Thus, the club with the most points at the end of the year will have their hands on the exquisite EPL trophy, and the total points also determines the fates of teams in the relegation zone. This makes every single match so critical, as losing one tiny single point could end up costing a team’s chance of winning a title or remaining in the top tier soccer league in England.

Data Importing and Tidying

Software Used

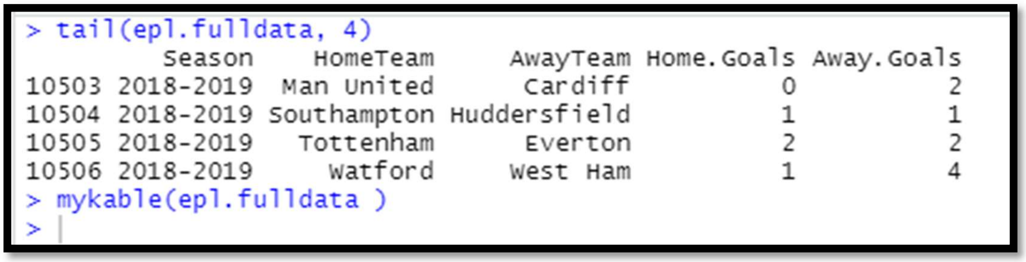
The software that has been used in the entire project is R. Suitable packages were added. The packages used are: dplyr, knitr, tidyverse, kableExtra, mosaic, readxl, surveillance, knitcitations.

Methodology

The ultimate data file for this project will simply consist of match final scores of all Premier League games from its inaugural competition, the 1992-93 season, to the last fixture of the last completed 38-matchweek contest, 2018-19 season. The 5 main attributes of this dataset are Season, Home Team, Away Team and the number of goals scored by each team.

In addition, since we are interested in investigating the relationship between scoring time and the time between goals, data on these two topics for Manchester United, a Premier League club, are collected and stored in a file.

We write a function that will help us in formatting the table when needed.We also write an importing function. This function takes in the starting year of the season, reads in the data file from its file path, and performs some table transformations including creating a new column to represent the seasons, renaming and selecting variables that are needed in further steps. Then we combine all the data from every year into one big data table. Here’s a quick glimpse at the last 4 rows of our data table.



```
> tail(epl.fulldata, 4)
          Season    HomeTeam     AwayTeam Home.Goals Away.Goals
10503 2018-2019  Man United      Cardiff          0          2
10504 2018-2019 Southampton Huddersfield          1          1
10505 2018-2019   Tottenham      Everton          2          2
10506 2018-2019     Watford     West Ham          1          4
> mykable(epl.fulldata )
> 
```

Goal Scoring and the Poisson Process

The Poisson Process

The Poisson process is a randomly determined process used to model the occurrence (or arrival) of phenomena over a continuous interval, which in most cases represents time. There are several characteristics of the Poisson process that can be observed, including, the number of events happening in a given time period; the time between those events; and when (at what point of time) the events occur. Playing a huge role in the Poisson process is the Poisson distribution, which deals with the number of occurrences of an event in a fixed period of time, with a rate of occurrence parameter λ. Another key distribution in this process is the exponential distribution, which has a strong connection with the Poisson distribution, as if the number of occurrences per interval of time are illustrated by Poisson, then the description of the length of time between occurrences are provided by the exponential distribution. If Poisson events take place on average at the rate of λ per unit of time, then the sequence of time between events (or interarrival times) are independent and identically distributed exponential random variables, having mean $\beta={}^{1}/_{\lambda}$. Furthermore, there’s a relationship between Poisson and another famous probability distribution - the continuous uniform distribution. If a Poisson process contains a finite number of events in a given time interval, then the unordered times, or locations, or positions, or points of time at which those events happen are uniformly distributed on that interval.

We suspect that goal scoring in soccer can be modeled by a Poisson process. According to the characteristics described above, if goal scoring for a club happens at a certain rate in a given time period, then a Poisson distribution can be used to model the number of goals scored. Additionally, the waiting time (usually in minutes) between successive instances of goal can be described using an exponential distribution. Moreover, the positions of time, better known as “minute marks”, in a game at which scoring events transpire may be uniformly distributed. We’re going to answer these questions in this research. With that goal, we’re now moving on to the modeling and analysis phases of this report.

The Poisson distribution, named for French mathematician Siméon Denis Poisson, is a discrete probability distribution that expresses the number of occurrences of an event over a given period of time. A Poisson random variable can represent many instances in our daily lives such as the number of phone calls coming into the Math Workshop requesting a tutor in a week, the number of misprints in a newspaper, or the number of cars arriving at a fast-food drive-through in an hour. The probability function of a Poisson random variable X with parameter λ is given by:

$$p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!};\ x = 0, 1, 2, \ldots \text{ and } \lambda > 0$$

where X represents the number of occurrences of an event in a given unit time period, and λ is the constant rate of occurrence per time period.

The mean and variance of our Poisson Random Variable X, denoted by μ_X and σ_X^2 respectively, are

$$\mu_X = \lambda$$

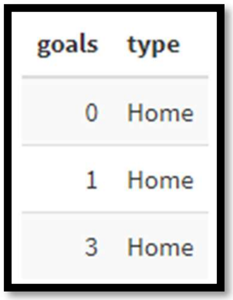
and

$$\sigma_X^2 = \lambda$$

We wanted to use this idea to model the goal scoring rate for soccer clubs during the Premier League era.

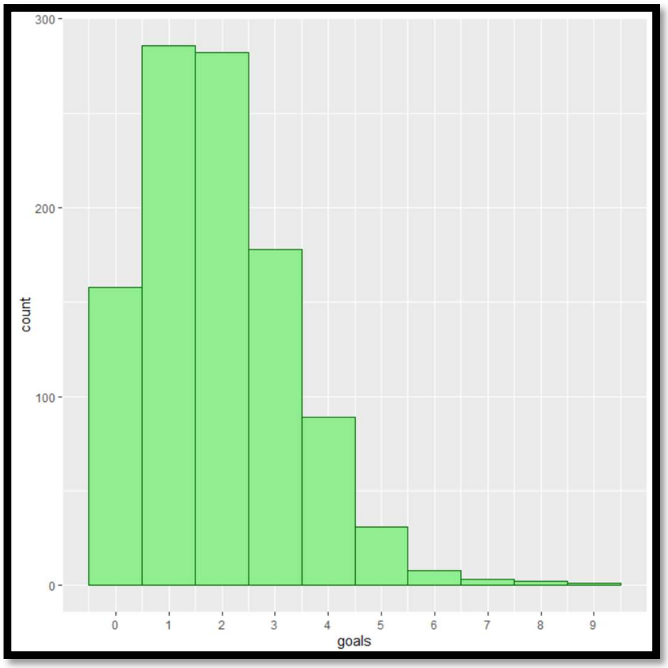
Poisson and the Number of Goals Scored

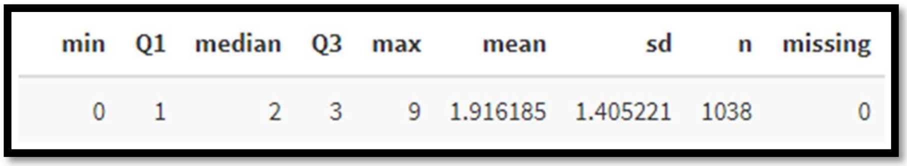
We use Manchester United (Man Utd), the most successful English club of all time, as our case of inspection. The question here is “Does Man Utd’s number of goals scored follow a Poisson distribution?” To answer this question, we first create a table of Man Utd’s goals. The table will consist of 2 columns: 1 for how many goals they scored in a match, and 1 for where the match took place - home or away. Therefore, each row of the table represents the number of goals scored in a match and the type of goal.



goals	type
0	Home
1	Home
3	Home

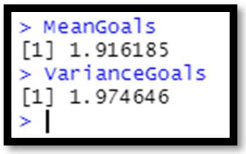
We now analyze with a simple histogram and some summary statistics of Man Utd's goals.





min	Q1	median	Q3	max	mean	sd	n	missing
0	1	2	3	9	1.916185	1.405221	1038	0

We now take a look at the mean and variance (the square of the standard deviation) of Man Utd's scoring rate. We are hoping to see these two values to be equal, since we know that the mean and variance of a Poisson random variable are the same.

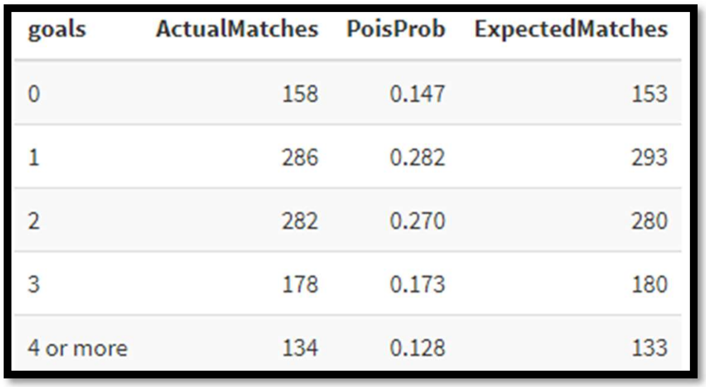


```
> MeanGoals
[1] 1.916185
> VarianceGoals
[1] 1.974646
> 
```

The mean and variance of scoring rate are 1.916 and 1.975 respectively, which are pretty close to each other, and this is exactly what we anticipated.

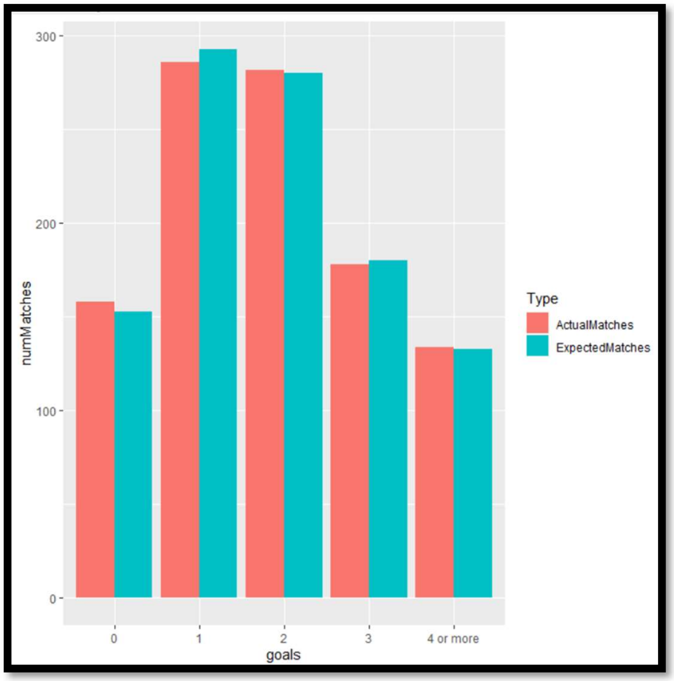
We now create a table that will have the possible values for number of goals, along with the following for each value: number of matches, Poisson probability, and expected number of matches. We first compile the number of goals scored and number of matches having those goal values. Since there are only a small number of matches with 4 goals or more, it'd be a good idea to combine them into a row called "4 or more." This will also help us while conducting a Chi-square Goodness of fit test, as the test is appropriate when the expected value of the number of counts in each level of the variable is at least 5.

We utilize the Poisson probabilities for each goal value to calculate the expected number of matches associated with those goal categories.



goals	ActualMatches	PoisProb	ExpectedMatches
0	158	0.147	153
1	286	0.282	293
2	282	0.270	280
3	178	0.173	180
4 or more	134	0.128	133

Just by looking at the expected number of matches, we can tell that what was predicted by our model is very similar to the actual number of matches. Furthermore, the bar graph below shows that our model does a pretty great job of predicting the number of matches with each number of goals.

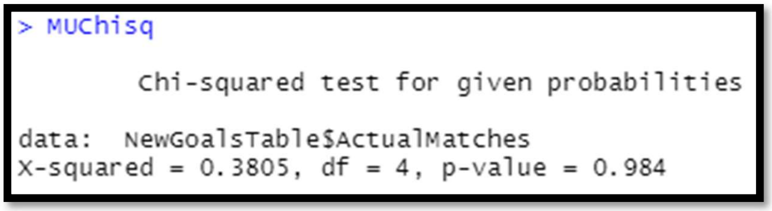


We now conduct a Chi-square Goodness of fit test to confirm that the Man Utd’s actual distribution of goals scored follows a Poisson distribution. The objective of this test is to compare the observed sample distribution with the expected probability distribution (here it’s Poisson). The null and alternative hypotheses for this test, denoted by H_0 and H_A respectively, are –

H_0: the data can be properly modeled by a specified distribution, and
H_A: the specified distribution does not fit the data appropriately.
For our specific case, the hypotheses are –
H_0: The distribution of Man Utd's goals scored follows a Poisson distribution, and
H_A: The distribution of Man Utd's goals scored does not follow a Poisson distribution.
The Chi-square test statistics (χ^2) is defined by the following formula:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed number of observations in category i, and E_i is the expected number of observations in category i. In our case, we have 5 total categories for goals scored: 0, 1, 2, 3, and 4 or more. The p-value for the Chi-square Goodness of fit test can be obtained from the upper tail of a Chi-square distribution of the χ^2 statistics on $k - 1$ degrees of freedom, where k is the number of categories. We use the Chi-square test to find out how well the Poisson distribution fits our data.



```
> MUChisq

        Chi-squared test for given probabilities

data:  NewGoalsTable$ActualMatches
X-squared = 0.3805, df = 4, p-value = 0.984
```

Since we get a very large p-value (0.984), we do not reject the null hypothesis of our Chi-square test. So, we don't have evidence to claim that the data don't fit a Poisson distribution. We now use the idea of the power of a hypothesis test to make a final conclusion. The power of a test is the probability that it will correctly reject a false null hypothesis. It is highly influenced by the test's sample size, the larger the number of observations, the higher the statistical power of a test. Since we have a large sample size of 1038 Man Utd's games here, it is safe to say that our Chi-square goodness of fit test has high power. Hence, we can conclude that there is no significant difference between the data's and expected distribution. Thus, the distribution of Man Utd's goal scoring data is consistent with a Poisson distribution.

Time Between Goals and Exponential Distribution

The Exponential Distribution

The exponential distribution is closely related to the Poisson distribution that was discussed in the previous section. Recall that the Poisson process is used to model some random and sporadically occurring event in which the mean, or rate of occurrence (per time unit) is λ. We used the Poisson distribution to model goal scoring rate per match for Man United, and since we only focused on integer goal values, our Poisson random variable is discrete. We are now interested in modeling the time until the next occurrence of goal, which we can also think of in terms of “time between goals”. If we have a non-negative random variable X that is the time until the next occurrence in a Poisson process, then X follows an exponential distribution with probability density function.

$$f_X(x) = \lambda e^{-\lambda x} = \frac{1}{\beta} e^{-\frac{1}{\beta}x};\ x \geq 0$$

where λ represents the average rate of occurrence and β is the average time between occurrences. The mean and variance of an exponentially distributed random variable X are

$$\mu_X = \frac{1}{\lambda} = \beta$$

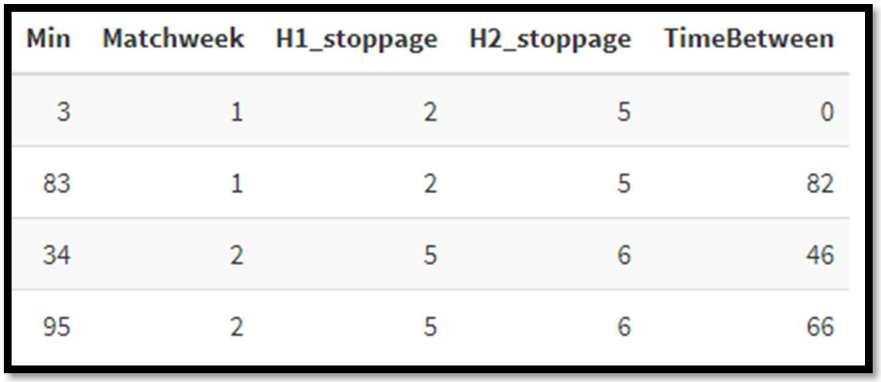
and

$$\sigma_X^2 = \frac{1}{\lambda^2} = \beta^2$$

We’re going to use this idea to model the time between each goal for a Premier League team in a given season.

Time Between Goals

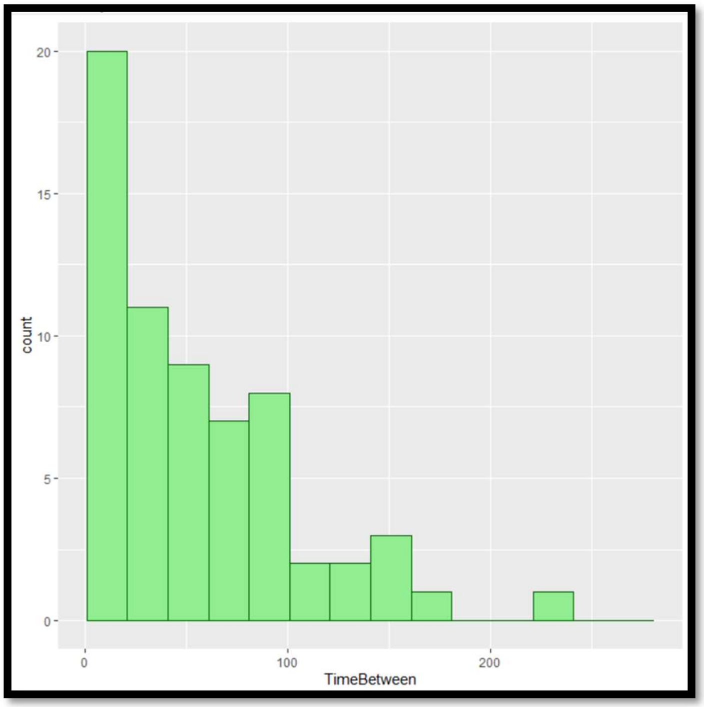
For this analysis, we use the data file that contains 5 columns: Minutes, which is the point of time during a match at which a goal was scored; Matchweek, which is the fixture number of each game; the stoppage time in minutes for both halves of each game, and finally, the time between goals, which takes into account the stoppage time. These 5 variables belong to all Man Utd’s goals during their 2018-19 Premier League campaign. Here’s a quick glance at our data table after we read in the data file.

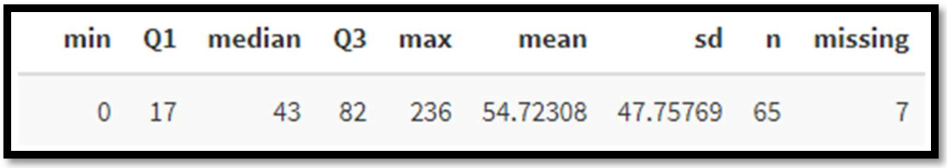


Min	Matchweek	H1_stoppage	H2_stoppage	TimeBetween
3	1	2	5	0
83	1	2	5	82
34	2	5	6	46
95	2	5	6	66

We’re going to take a moment to explain some of the above table’s attributes. In soccer, the term “stoppage time” is used to describe the number of minutes added to the end of each half to help make up for time lost during the course of the half, due to various reasons such as fouls, injuries, players and referee's arguments, or goal celebrations, for all of which the game is not being played. The variable of time between goals is calculated by measuring the minutes between a goal and the goal before it in the season.

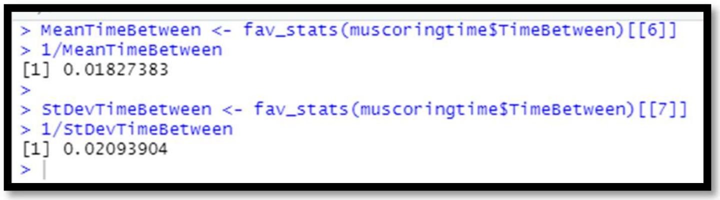
Below are a histogram and some descriptive statistics of Man Utd's time between each goal last season.





min	Q1	median	Q3	max	mean	sd	n	missing
0	17	43	82	236	54.72308	47.75769	65	7

The overall shape of this histogram looks like an exponential distribution curve, which is a continuous, smooth and concave up graph, displaying exponential decay. This is a good sign for us, since we suspected that time between goals for Man Utd is exponentially distributed. Now we're going to compare 1/(mean time between), which is the Poisson rate of occurrence λ, and 1/(standard deviation of time between). We are hoping that these 2 values are equal to each other, since for an exponentially distributed random variable, its variance is equal to its mean squared, and we also know that standard deviation is the square root of variance, so mean and standard deviation should be equal, and so are their reciprocals.



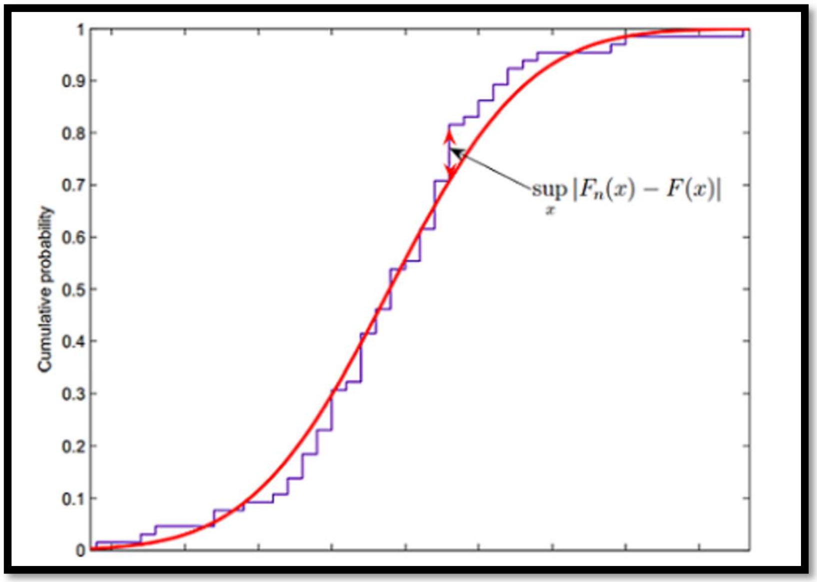
```
> MeanTimeBetween <- fav_stats(muscoringtime$TimeBetween)[[6]]
> 1/MeanTimeBetween
[1] 0.01827383
>
> StDevTimeBetween <- fav_stats(muscoringtime$TimeBetween)[[7]]
> 1/StDevTimeBetween
[1] 0.02093904
>
```

The values are not that far away, which leads us to believe more that our initial suspicion is true.

Now, we would like to know whether the time between goals of Man Utd follows an exponential distribution. To check this, we perform a Kolmogorov-Smirnov (KS) test, which is another Goodness of fit test. This KS test applies to continuous distributions, like our distribution of interest, exponential; whereas the Chi-square test we used in the previous section works best for categorized data, meaning that the data has been counted and divided into categories. The null and alternative hypotheses for this test are – H_0: the data follow a specified distribution (in our case, the minutes between goal follow an exponential distribution), and H_A: the data do not follow the specified distribution.

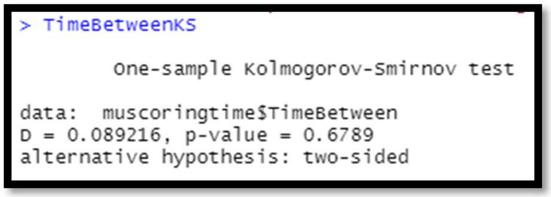
The test statistic for a KS test is defined as:

$$D = \max_x |F_n(x) - F(x)|$$



As shown in the figure above, this D value represents the greatest vertical distance, denoted by max for maximum (or sometimes by sup for supremum) between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution (here it's exponential). The closer (or lower) the D value to zero, the more probable that the data follow the specified distribution. The higher D is to 1, the more probable that they have different distributions.

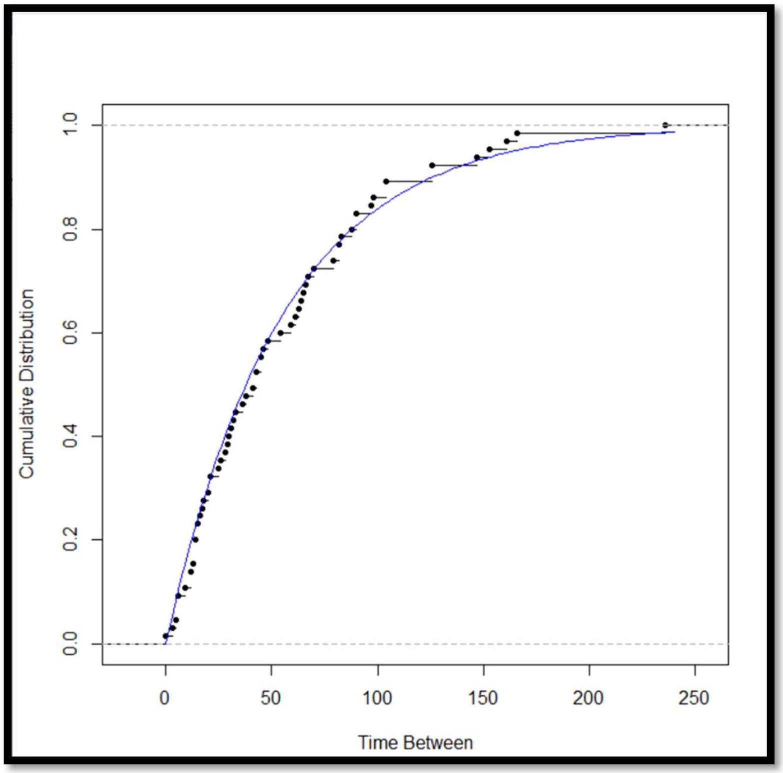
We perform a KS test which takes in the data and the distribution (with its parameter(s) fully specified) we want to compare our data with, and gives us the KS test statistics D and also a p-value of the test. We also create a graph to compare the cumulative distributions of time between goals and the hypothetical distribution - the exponential.



```
> TimeBetweenKS

        One-sample Kolmogorov-Smirnov test

data:  muscoringtime$TimeBetween
D = 0.089216, p-value = 0.6789
alternative hypothesis: two-sided
```



Since we get D = 0.089 and p-value = 0.679, we fail to reject the null hypothesis. Thus, there's not sufficient evidence to support a conclusion that our data are not consistent with the exponential distribution. For this test, since we have a large enough number of data values, the test has high statistical power, and this allows us to say that the time between goals of Man Utd follows an exponential distribution.

Scoring Time and the Uniform Distribution

The Continuous Uniform Distribution

The continuous uniform distribution is a probability distribution with equally likely outcomes, meaning that its probability density is the same at each point in an interval $[A, B]$. The graph of a uniform distribution results in a rectangular shape, hence this is why it is sometimes referred to as the "rectangular distribution."

A continuous random variable X is uniformly distributed on $[A, B]$ if its probability density function is defined by

$$f_X(x) = \frac{1}{B-A};\ A \le x \le B$$

The mean and variance of a uniformly distributed random variable X are

$$\mu_X = \frac{A+B}{2}$$

and

$$\sigma_X^2 = \frac{(B-A)^2}{12}$$

One popular version of the uniform distribution is the standard uniform distribution. The domain for this special case is the unit interval $[0, 1]$. A random variable U is said to have a standard uniform distribution if it has probability density function

$$f(u) = 1;\ 0 \le u \le 1$$

The mean and variances of U on $[0, 1]$ are

$$\mu_U = \frac{1}{2}$$

and

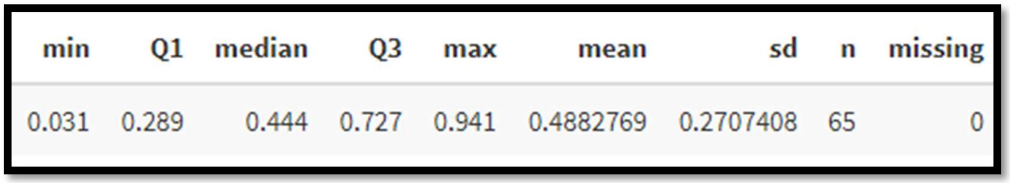
$$\sigma_U^2 = \frac{1}{12}$$

We suspect that the scoring time of Man Utd is uniformly distribution. To that end, let's find out whether this is true.

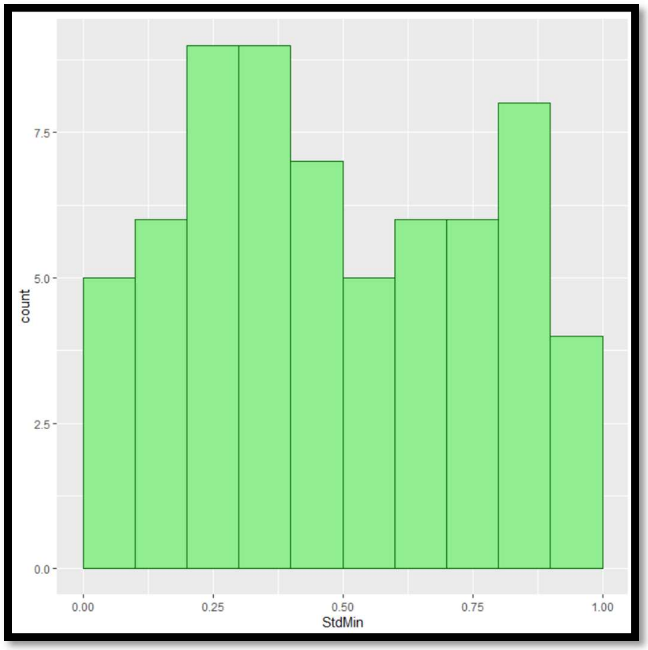
Are the Scoring Time Uniformly Distributed?

We first standardize the minutes by dividing each one of them by the total minutes of their respective game. The reason for standardizing the minutes is because the match total time varies, since we also take into account the stoppage time, which means some matches lasted longer than others.

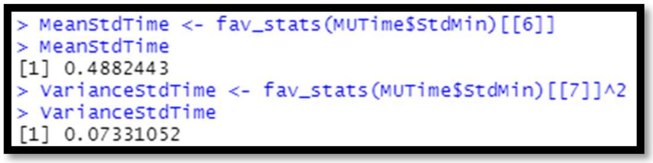
Below are a histogram and some summary statistics of the standardized version of Man Utd's scoring minutes last season. The overall shape of the distribution of standardized minutes is not as rectangular as the usual uniform curve. What we can tell from this distribution is goals tend to occur in the middle minutes of each half, and less goals take place at the beginning and end of each playing period.



min	Q1	median	Q3	max	mean	sd	n	missing
0.031	0.289	0.444	0.727	0.941	0.4882769	0.2707408	65	0



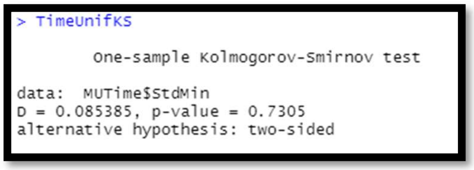
We now compare the mean and variance of standardized scoring minutes of our data with the mean and variance of a standard uniform distribution, which are 1/2 and 1/12 respectively.



```
> MeanStdTime <- fav_stats(MUTime$StdMin)[[6]]
> MeanStdTime
[1] 0.4882443
> VarianceStdTime <- fav_stats(MUTime$StdMin)[[7]]^2
> VarianceStdTime
[1] 0.07331052
```

Our data’s mean is pretty close to the expected value for a standard uniform random variable. Meanwhile, our variance value is slightly smaller than the expected variance. One possible explanation for this low variance value is because our distribution graph is not as flat as the regular uniform curve, as we observed more goals toward the middle and fewer goals than expected at the extremes of the distribution. Thus, this means the goal data points are closer to the center, which results in a narrower spread.

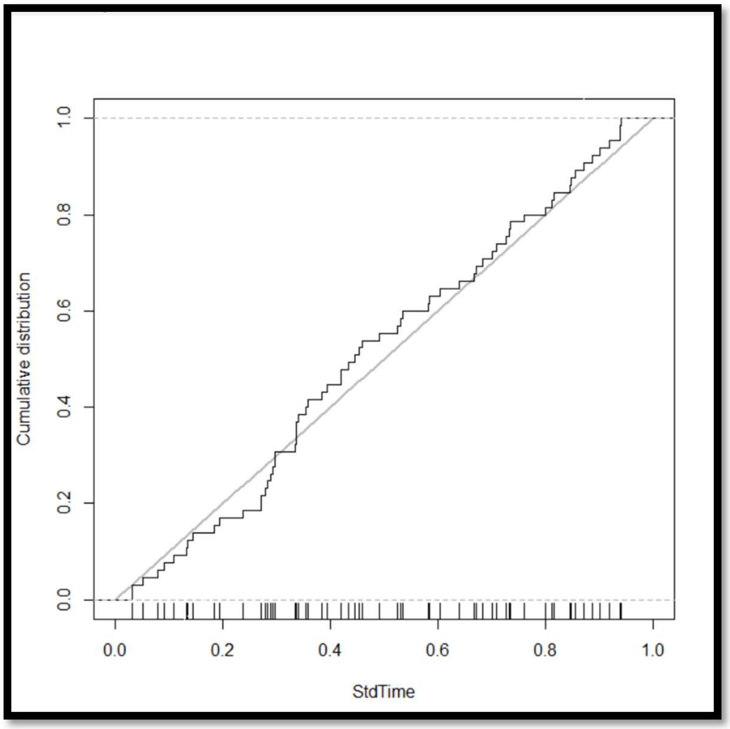
We are interested in finding out whether our data of Man Utd’s scoring time actually follow a standard uniform distribution. Since our distribution is continuous, we once again use a Kolmogorov-Smirnov goodness of fit test to answer this question. Just like the KS test from the previous section, the null hypothesis for this test is there’s a good fit between our data and the specified distribution (uniform in this case), and the alternative hypothesis is the scoring time data don’t fit the uniform distribution. The test results are accompanied by a plot showing how the empirical and hypothetical cumulative distributions differ.



```
> TimeUnifKS

        One-sample Kolmogorov-Smirnov test

data:  MUTime$StdMin
D = 0.085385, p-value = 0.7305
alternative hypothesis: two-sided
```



A D-statistics of 0.0854 and a p-value of 0.73 indicate that there's no good evidence against the claim that the data is not consistent with a uniform distribution. As we, have a big enough amount of data points leading to high statistical power here, we can conclude that our data is consistent with the specified reference distribution. In context, the standardized scoring time for Man Utd during the 2018-19 season is uniformly distributed.

Predicting 2018-19 Season Results from Prior Data

Data Transforming

We first get a new data frame with matches from last year being filtered out, since we want to use data prior to 2018-19 to predict the results of last season. We now create Poisson regression models and use them to get the scoring rates for all teams at home and on the road. Poisson Regression is a member of a broad class of models known as the Generalized Linear Models (GLM). A generalized linear model has the general form

$$E(Y_i) = \mu_i = g^{-1}(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik})$$

There are three main components to a generalized linear model:

1. A random component, indicating the conditional distribution of the response variable Y_i (for the ith of n independently sampled observations), given the values of the explanatory variables. Y_i's distribution must be a member of an exponential family, such as Gaussian, Binomial, Poisson, or Gamma.
2. A linear predictor,

$$(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)$$

   which is a linear combination of the explanatory variables (the X 's), with the β 's as the regression coefficients to be estimated.
3. A canonical link function gg, which transforms the expected value of the response variable, $E(Y_i) = \mu_i$, to the linear predictor.

Poisson regression models are generalized linear models with the logarithm as the link function. It is used when our response's data type is a count, which is appropriate for our case since our count variable is the number of goals scored. The model assumes that observed outcome variable follows a Poisson distribution and attempts to fit the mean parameter to a linear model of explanatory variables.

The regression equation for a Poisson regression model is of the form

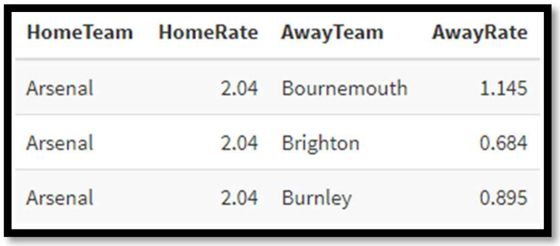
$$\mu_i = e^{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}}$$

or equivalently,

$$ln(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}$$

We first do Poisson regression to get every team's scoring rate when playing at home.

Our next step is to get the regression coefficients (all the β's) for every team from this model, alongside with the model's y-intercept, which will allow us to make predictions later on. We make a table of all possible matchups and all the home and away rates. Here's a quick glimpse at our table. It has 380 rows, representing the 380 total team matchups of the season. In each Premier League season, a club gets to play every other squad twice, once on their home pitch and once on the other team's field. As we can see from the table below, the first matchup is Arsenal - Bournemouth, at Arsenal. There is also a row for a game between these two teams, but in reverse order, with Bournemouth playing at home against Arsenal.



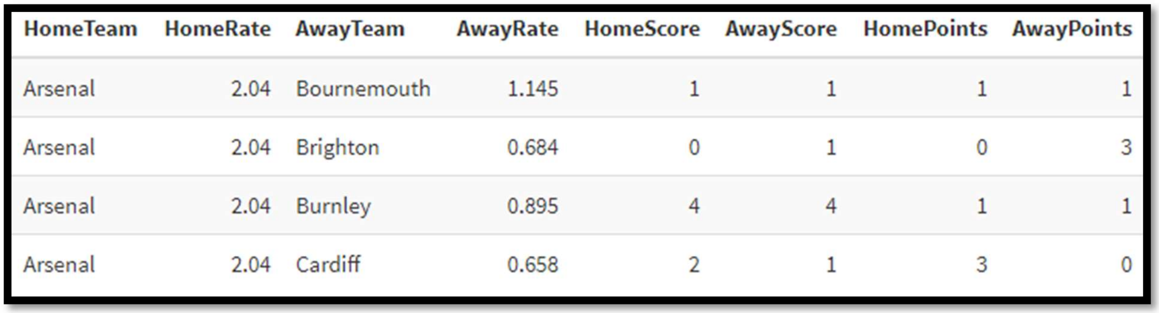
HomeTeam	HomeRate	AwayTeam	AwayRate
Arsenal	2.04	Bournemouth	1.145
Arsenal	2.04	Brighton	0.684
Arsenal	2.04	Burnley	0.895

Simulation

We would like to simulate the results of the 2018-19 seasons 10000 times. Our goal is to get the team's ranking, their total points, and their goal differential for each simulated season. Our current table has 380 rows representing all possible matchups of the 2018-19 seasons. To create of 10000 simulations, we duplicate this table 10000 times, and create a new table.

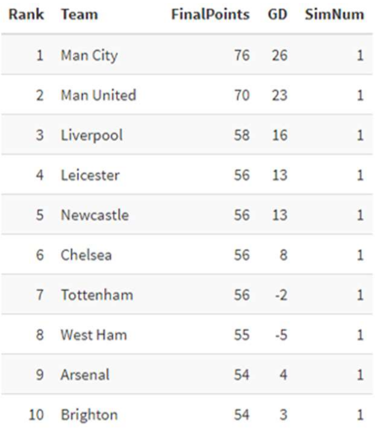
We then generate the number of goals scored for every home and away team in every row of our table. In addition, the number of points for every match outcome based on the teams' number of goals scored are also calculated, as a side gets 3 points if it scores more than its opponent, 1 point if it's a tie, and 0 points if the opposing roster has more goals.

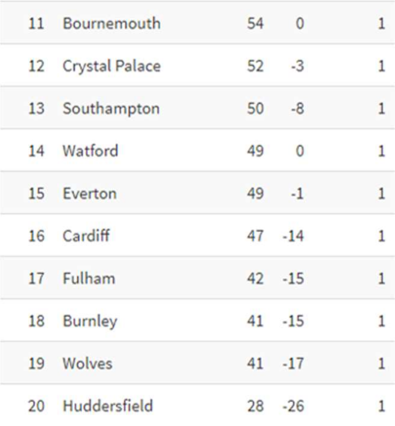
Here's a look at our table. The two columns *HomeScore* and *AwayScore* indicate our simulate match outcome for each matchup, as they are the number of goals scored for each club randomly generated using their scoring rates.



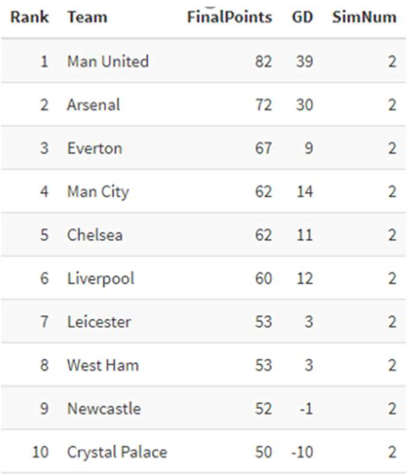
HomeTeam	HomeRate	AwayTeam	AwayRate	HomeScore	AwayScore	HomePoints	AwayPoints
Arsenal	2.04	Bournemouth	1.145	1	1	1	1
Arsenal	2.04	Brighton	0.684	0	1	0	3
Arsenal	2.04	Burnley	0.895	4	4	1	1
Arsenal	2.04	Cardiff	0.658	2	1	3	0

Since each season has 380 games and we duplicated the table 10000 times, the table has a total of 3800000 rows. We can say that every 380-row segment contains the result of 1 simulated season. We then get each individual simulation and also tally up the points, calculate goal differentials and get the team ranking for each season. The next page shows 2 sample simulated seasons we got by using this function. Next, we're interested in getting the result of each one of our 10000 simulations and put everything together into a data frame.

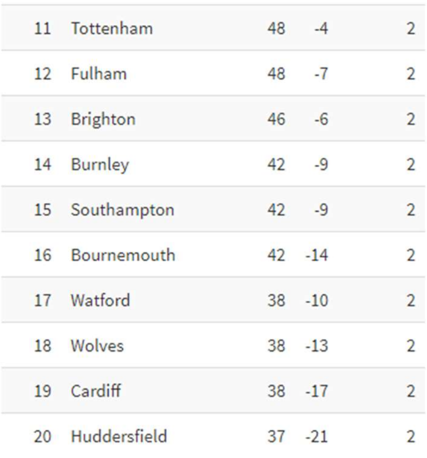




Rank	Team	FinalPoints	GD	SimNum
1	Man City	76	26	1
2	Man United	70	23	1
3	Liverpool	58	16	1
4	Leicester	56	13	1
5	Newcastle	56	13	1
6	Chelsea	56	8	1
7	Tottenham	56	-2	1
8	West Ham	55	-5	1
9	Arsenal	54	4	1
10	Brighton	54	3	1
11	Bournemouth	54	0	1
12	Crystal Palace	52	-3	1
13	Southampton	50	-8	1
14	Watford	49	0	1
15	Everton	49	-1	1
16	Cardiff	47	-14	1
17	Fulham	42	-15	1
18	Burnley	41	-15	1
19	Wolves	41	-17	1
20	Huddersfield	28	-26	1



Rank	Team	FinalPoints	GD	SimNum
1	Man United	82	39	2
2	Arsenal	72	30	2
3	Everton	67	9	2
4	Man City	62	14	2
5	Chelsea	62	11	2
6	Liverpool	60	12	2
7	Leicester	53	3	2
8	West Ham	53	3	2
9	Newcastle	52	-1	2
10	Crystal Palace	50	-10	2



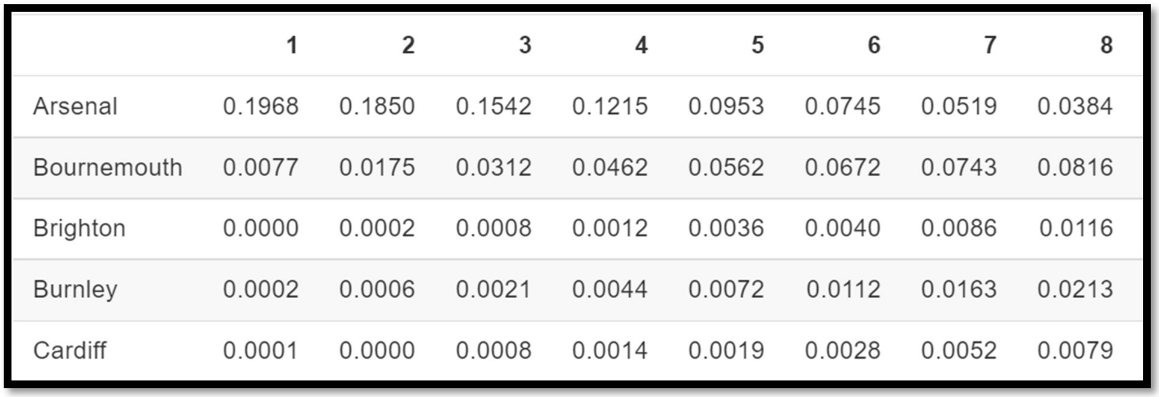
11	Tottenham	48	-4	2
12	Fulham	48	-7	2
13	Brighton	46	-6	2
14	Burnley	42	-9	2
15	Southampton	42	-9	2
16	Bournemouth	42	-14	2
17	Watford	38	-10	2
18	Wolves	38	-13	2
19	Cardiff	38	-17	2
20	Huddersfield	37	-21	2

Analysis

After the simulation is complete, we write the simulation table to a csv file and name so we can now import the data file and do some analysis.

Team Rankings

We now obtain a table of every 2018-19 team's chance of finishing at each position on the final standing table from our simulation, as shown by the below output table.



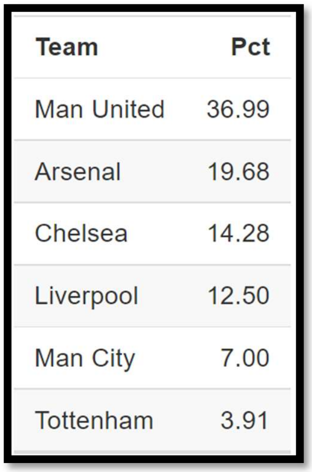
	1	2	3	4	5	6	7	8
Arsenal	0.1968	0.1850	0.1542	0.1215	0.0953	0.0745	0.0519	0.0384
Bournemouth	0.0077	0.0175	0.0312	0.0462	0.0562	0.0672	0.0743	0.0816
Brighton	0.0000	0.0002	0.0008	0.0012	0.0036	0.0040	0.0086	0.0116
Burnley	0.0002	0.0006	0.0021	0.0044	0.0072	0.0112	0.0163	0.0213
Cardiff	0.0001	0.0000	0.0008	0.0014	0.0019	0.0028	0.0052	0.0079

This table has 20 columns and 20 columns.

For example, if we look at row 1, Arsenal has a 0.1968 (19.68%) probability of finishing first, 18.5% chance of finishing second, and so on and so forth. To get the probability of Arsenal finishing at least at a certain position, we can just simply add up the probabilities of being at or above that position. For example, Arsenal's chance of finishing in the top 4 is the sum of their probabilities of ending at position 1, 2, 3, and 4, which is 0.1968 + 0.1850 + 0.1542 + 0.1215 = 0.6575 = 65.75%.

First Place

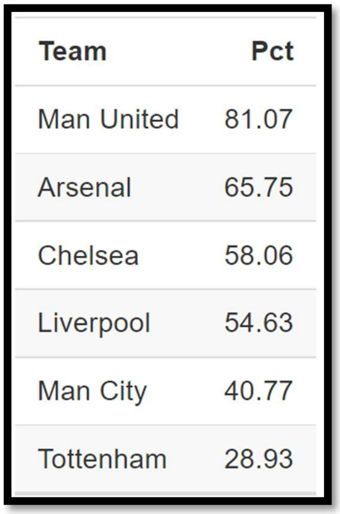
Just like many top sports leagues around the world, the team that finishes first at the end of each Premier League season will be crowned league champions and will get to take the trophy home with them. The table above shows the 6 teams with the highest chance of winning the 2018-19 season based on our simulation results.



Team	Pct
Man United	36.99
Arsenal	19.68
Chelsea	14.28
Liverpool	12.50
Man City	7.00
Tottenham	3.91

Unsurprisingly, they are the infamous Premier League's "Big 6" - Manchester United, Arsenal, Chelsea, Liverpool, Manchester City and Tottenham.
Man Utd leads the way by winning 36.99% of the seasons, or 3699 out of 10000 simulated seasons; followed by Arsenal, Chelsea, and so on.
Since we're using data from all Premier League seasons and Man Utd is the winningest club in the history of top tier English football, in terms of both number of titles and matches, it completely makes sense why they won the title race more than any other club in our simulations.

Top 4

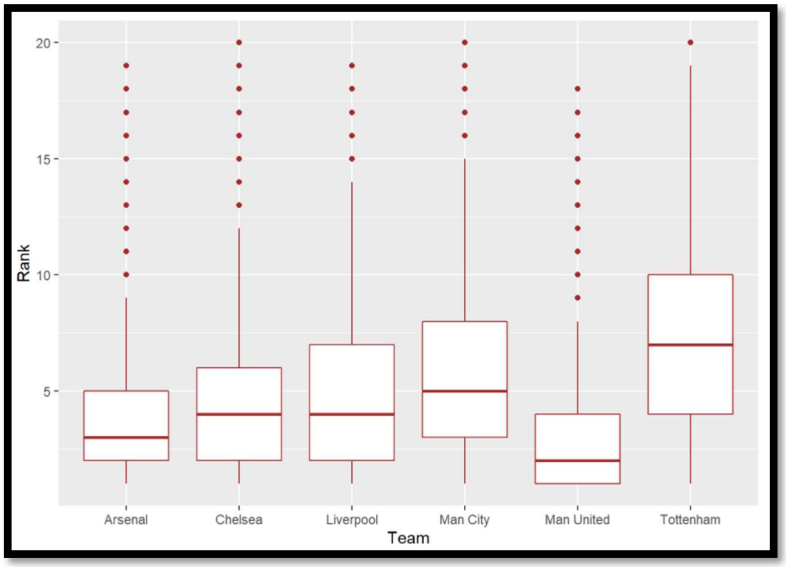


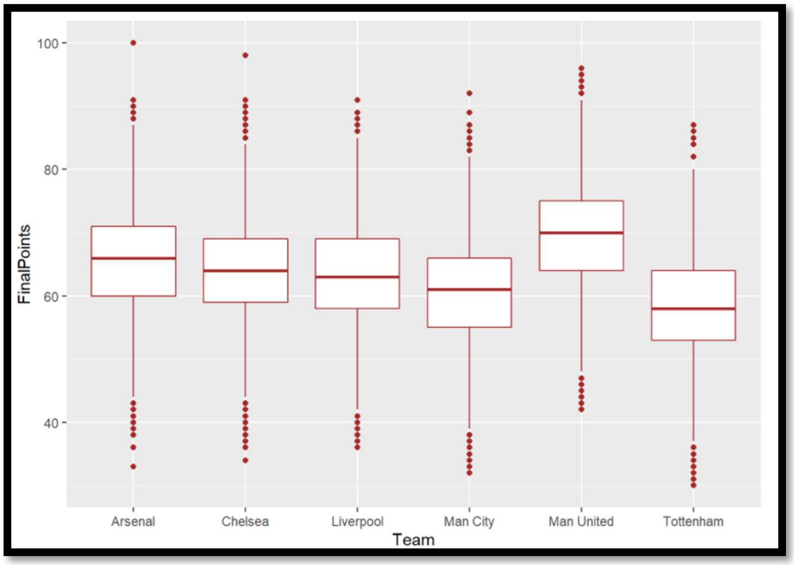
Team	Pct
Man United	81.07
Arsenal	65.75
Chelsea	58.06
Liverpool	54.63
Man City	40.77
Tottenham	28.93

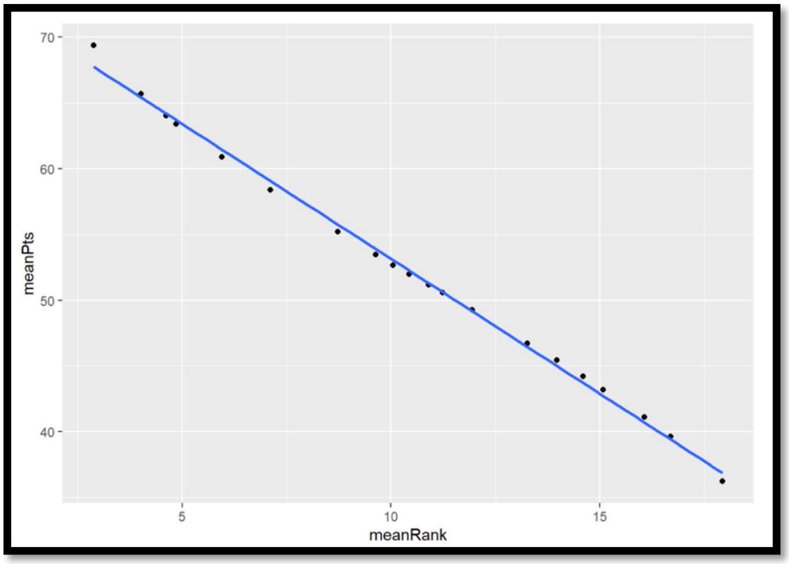
The significance for a ball club to finish in the top 4 of the Premier League is that they would punch their tickets to the UEFA Champions League, an annual competition contested by top football clubs in Europe. Based on the table above, the big 6 once again dominate this category, as Manchester United has the highest chance making the top 4 at 81.07. Arsenal, Chelsea and Liverpool each secures a Champions League spot in more than half of the 10000 simulations. In reality, the 4 teams that claimed the top 4 spots of the table last year in order were Man City, Liverpool, Chelsea, and Tottenham; followed by Arsenal at fifth and Man United at sixth.

Big 6

Since the Big 6 are head and shoulders above everyone else in our first two categories, we might as well want to look at their overall distribution of Ranking and Total Points from our simulation. Below the first two-figure are the side-by-side boxplots of their Final Ranking and Total Points distributions.
We can see that Man Utd's rank numbers are lower than other Big 6 clubs on average, meaning that United tend to secure a higher place on the table, and they also have higher mean final points than other teams. In contrast, Tottenham tends to have lower ranking spots and lower points than their fellow Big 6 competitors.
Unsurprisingly, there's a strong and linear correlation between team's ranks and total points, meaning that higher total points are associated with a higher (lower number) finishing position of the table, as illustrated by the last figure.

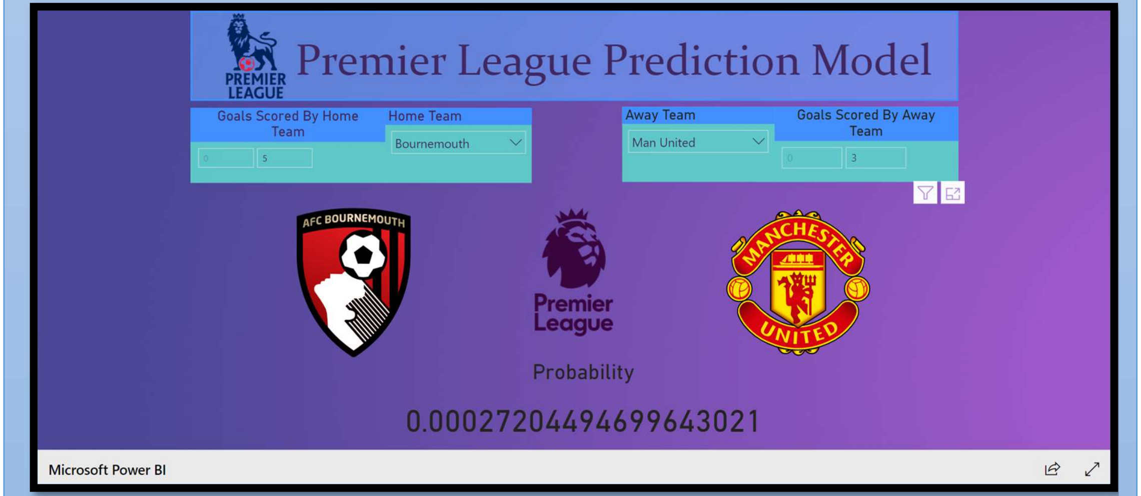

Rank
20
15
10
5
Arsenal
Chelsea
Liverpool
Man City
Man United
Tottenham
Team


FinalPoints
100
80
60
40
Arsenal
Chelsea
Liverpool
Man City
Man United
Tottenham
Team


meanPts
70
60
50
40
5
10
15
meanRank

Results

We extracted data for all the possible combinations of teams that played the EPL in last season and stored the results in an excel. Using PowerBi, we developed a web-report that will help us view the results concisely.



Premier League Prediction Model

Goals Scored By Home Team: 0 – 5

Home Team: Bournemouth

Away Team: Man United

Goals Scored By Away Team: 0 – 3

Probability

0.00027204494699643021

Microsoft Power BI

Link to online PowerBi report – https://bit.ly/3lsdGST

Link to GitHub repository – https://github.com/tanishka2001/SoccerMatchScorePredictionModel

Conclusion

Overall, we have found that Premier League goal scoring fits the characteristics of a Poisson process. Our first result was a Poisson distribution can be used to predict the number of matches with each number of goals scored. Additionally, the time between each individual goal in a season can be described by an exponential distribution. We also have evidence that the goal scoring time positions after being standardized are uniformly distributed.

We also used the data to predict what would happen in 2018-19. We got each team's goal scoring rate at home and away from home by doing Poisson regression, and then performed simulations using those rate parameters. Different team metrics like how many points each team got and what place each team finished were being kept track of from the simulations. We extracted data for all the possible combinations of teams that played the EPL in last season and stored the results in an excel. Using PowerBi, we developed a web-report that will concisely help us view the results.

References

1. www.football-data.co.uk/englandm.php
2. www.wikipedia.org
3. www.link.springer.com
4. www.statisticsports.com
5. www.towardsdatascience.com
6. www.stats.idre.ucla.edu