Smartodds

Graduate Quant Analyst Test

Section 1

To prepare the Dataset for the assignment, I have used the dataset available in CSV format here : https://www.football-data.co.uk/usa.php

Created 2 subsets from this Dataset:

- Fitset : All matches between 01/01/2012 and 12/06/2015
- Simset : All matches between 12/06/2015 and 23/10/2016 i.e: 2016 MLS season

Before creating the 2 Datasets, I created an additional file with a list of all team between 2012 and 2016 seasons with a column in it representing the conference they belong to. This file has been used to merge into the original Dataset to have a *home_conference* and *away_conference* columns with values of the conference of the home and away teams respectively. The Date column of the file has been converted to a Python compatible format to have filters and indexes in the further sections easily.

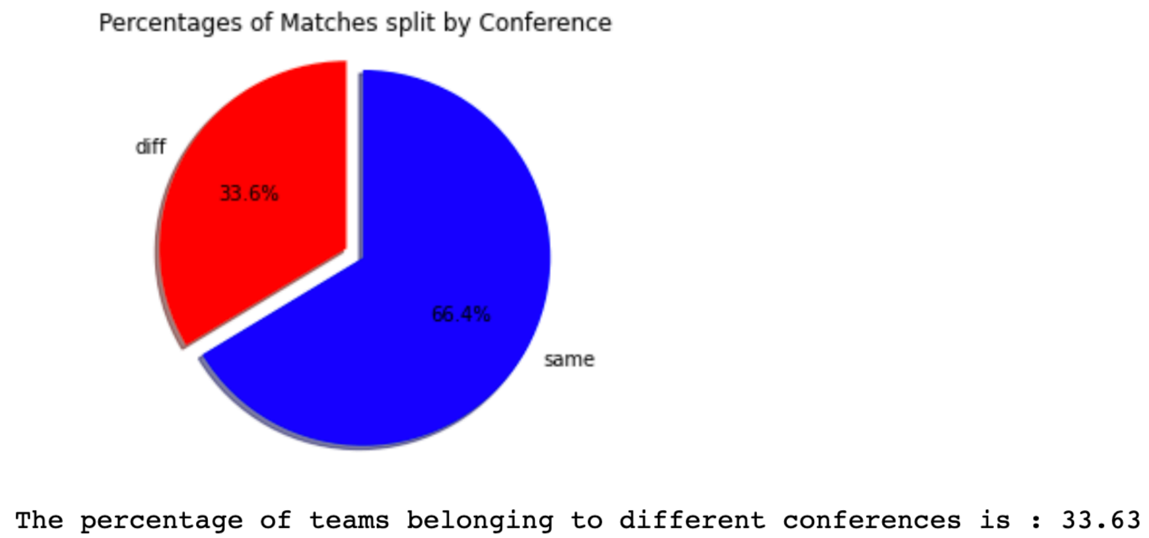
A column of total_goals has been added to the original dataset as well. This column represents the sum of Home Goals and Away Goals for each match. This will be used in finding the seasonality and in predictions in further sections.

These computations have been made even before splitting the dataset into training and test sets so as to avoid doing these for both datasets individually and rather save time, lines of code and computation time by doing it once before the initial split.

Section 2 – Descriptive Statistics

2.1

I have represented the percentages of matches where teams belong to different conference using a pie chart. The pie chart has blue and red colours only to have a consistency with the MLS league, which is the data we are dealing with here. This colour scheme will be more or less constant through this assignment.



```
The percentage of teams belonging to different conferences is : 33.63
```

Fig 1

The chart shows not only the percentage of matches in different conferences but also matches in the same conference. The matches in the same conference are almost twice as the matches between different conference.

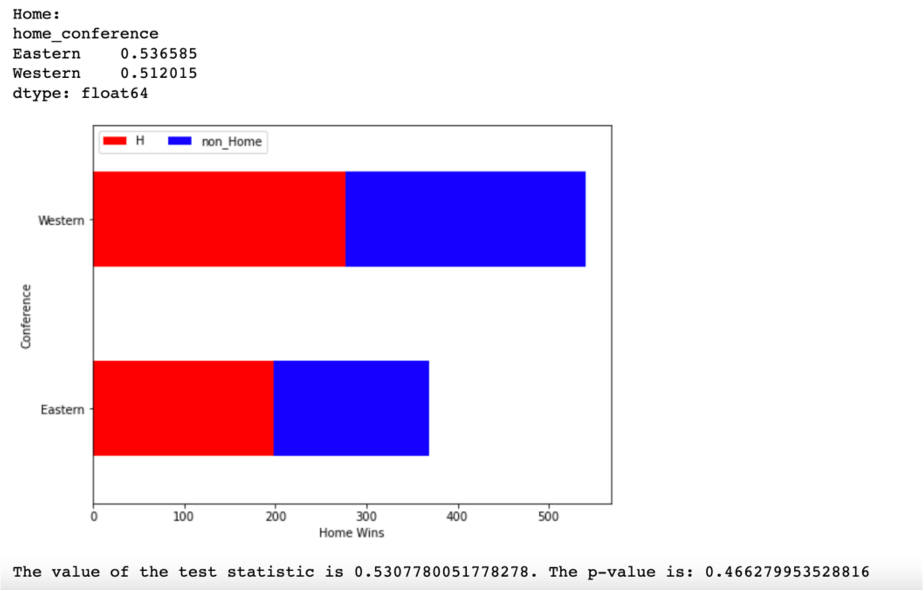
2.2

To compute statistical significance of the difference in home advantage between Eastern and Western Conference:

For the matches between teams in the same conference, The Eastern conference has a home win percentage of 53.65%(148 Home games won out of 265) and the Western conference has a Home win percentage of 51.2%(277 Home games won out of 541)

Our Null Hypothesis is that the difference in Home advantage between Eastern and Western conference is not significant.

Using Chi Squared test to test if the difference is statistically significant, also taking into account the number of games in each conference, we get a p-value of 0.466. If we assume a significance level of 0.1 (10%), the p-value is above that significance level, hence failing to reject the null hypothesis and denoting the difference is not statistically significant.



```
Home:
home_conference
Eastern    0.536585
Western    0.512015
dtype: float64
```

```
The value of the test statistic is 0.5307780051778278. The p-value is: 0.466279953528816
```

Fig 2

2.3

To find Seasonality in the Data, I have tried to find a yearly pattern in the average Total goals scored per month in a match along with the average Total goals scored at Home per month in a match and goals scored Away per month.

We will first analyse the average Total Goals scored per month in a game.

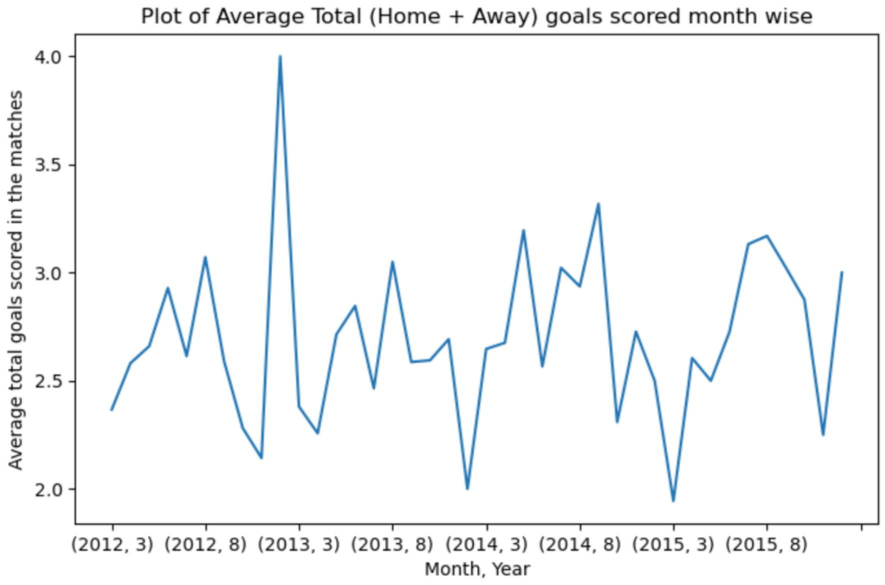


Fig 3

Figure 3 above, represents the Average Total goals scored in match , month wise from the start of season 2012 till the end of season 2015. We have the plot above for 4 full seasons. The first step is to always visualise and analyse the data visually. There is a slight seasonal aspect to the plot since we can see some reputative nature. Our dataset is from March to December every year, which makes our seasonal period as 10 months.

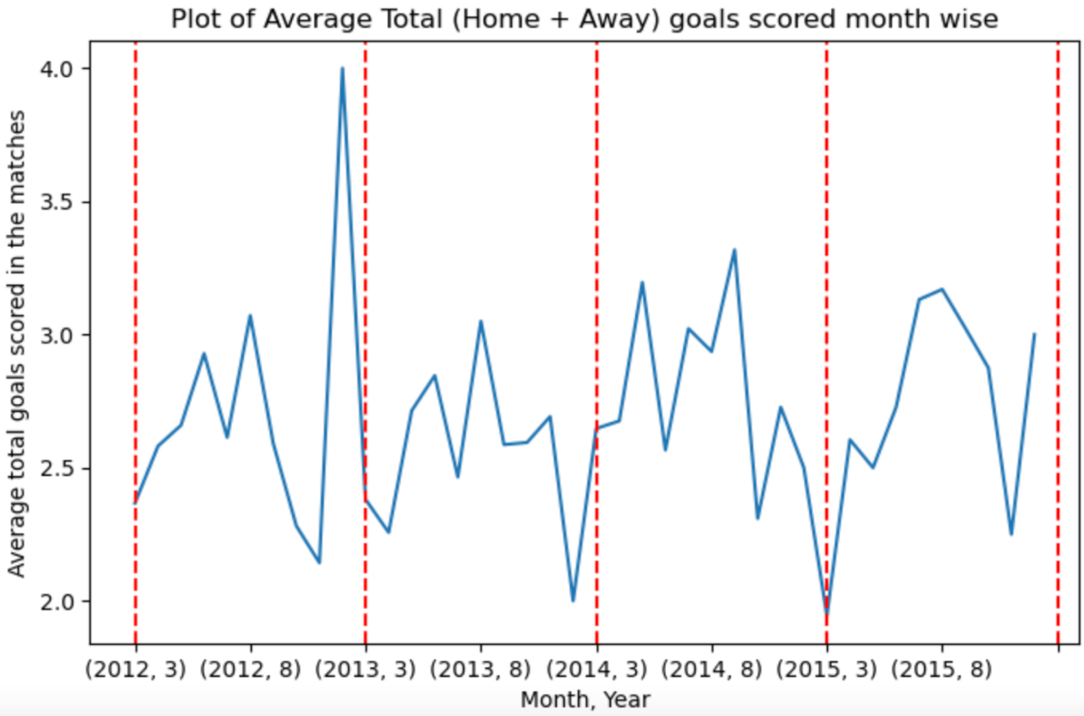


Fig 4

I have separated the data with red dotted lines for every season. This gives us an idea that there is one peak at the middle of every season and one dip towards the end. This is constant throughout the 4 seasons. This makes it quite clear that the behaviour is seasonal.

We will identify the trend. I have tried to fit a quadratic function to the series.

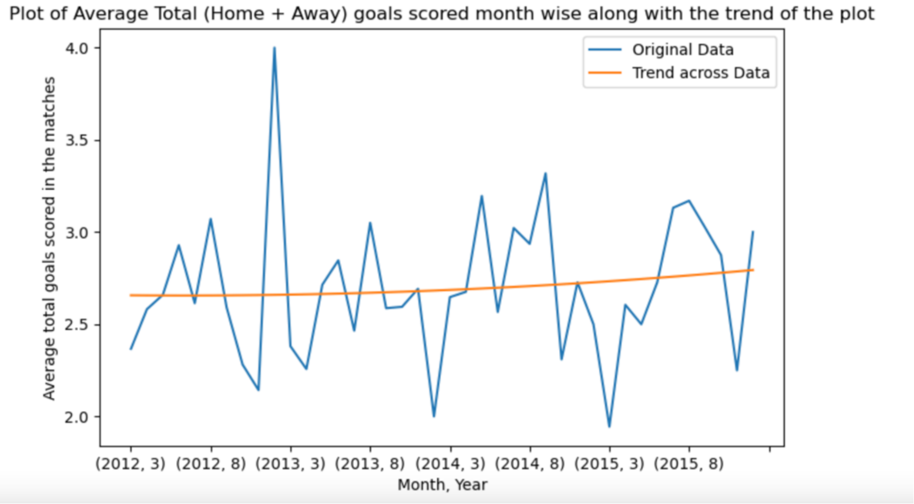


Fig 5

Although the trend was minor and near about constant, it was important to find it so I could subtract it from my original data to find the correct seasonal behaviour. Trend was then subtracted from the original dataset to get the below plot.

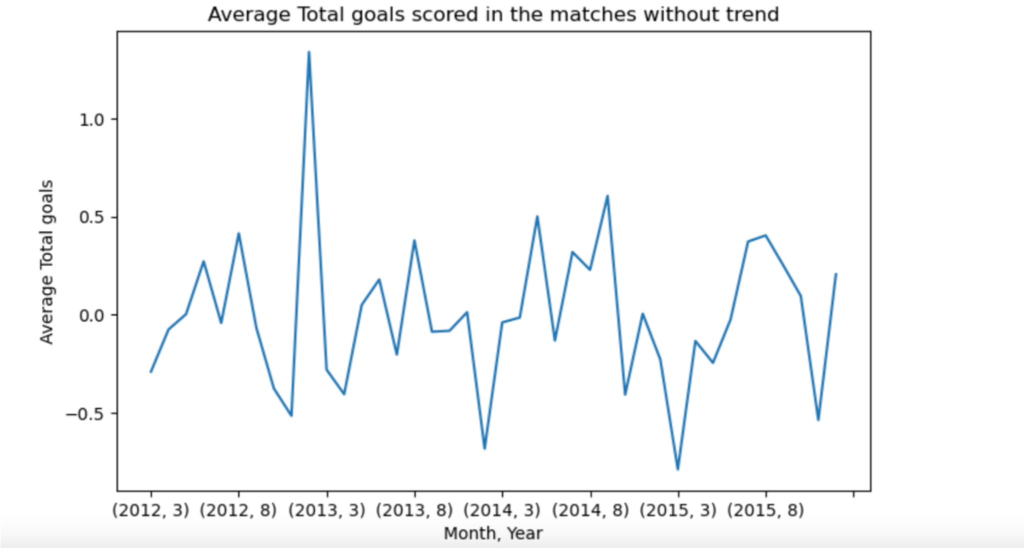


Fig 6

To find the seasonality, I considered an individual “wave” of length 10. The values of wave were obtained by averaging the respective values. The first average would be $x_{0,}x_{10}$, x_{20}… and so on. The x’s here represent the de-trended data. Hence we get 10 averages. These represent the average seasonal value for one year. These were repeated across 4 seasonal cycles to get the seasonal visualisation.

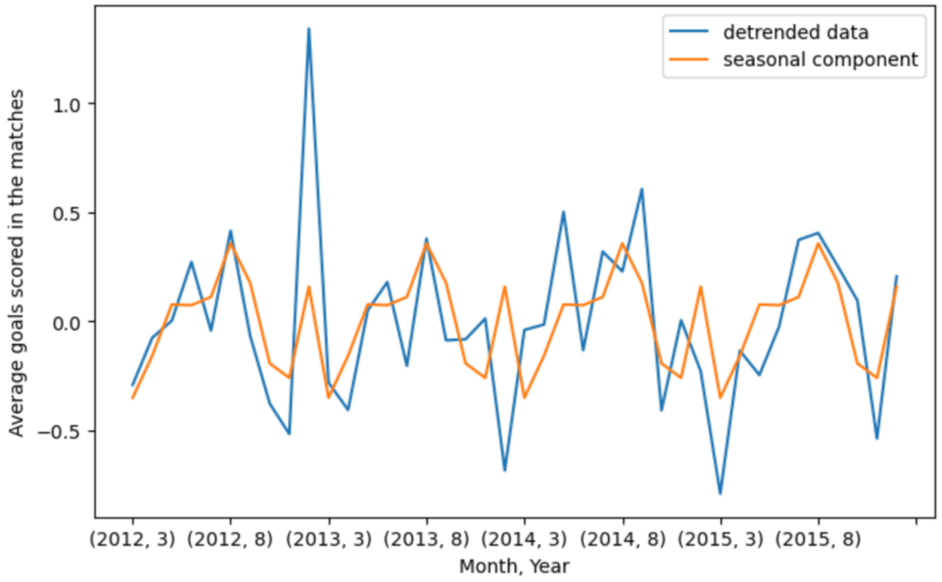


Fig 7

I also went on to plot cycle to further analyse the reasons and cause.

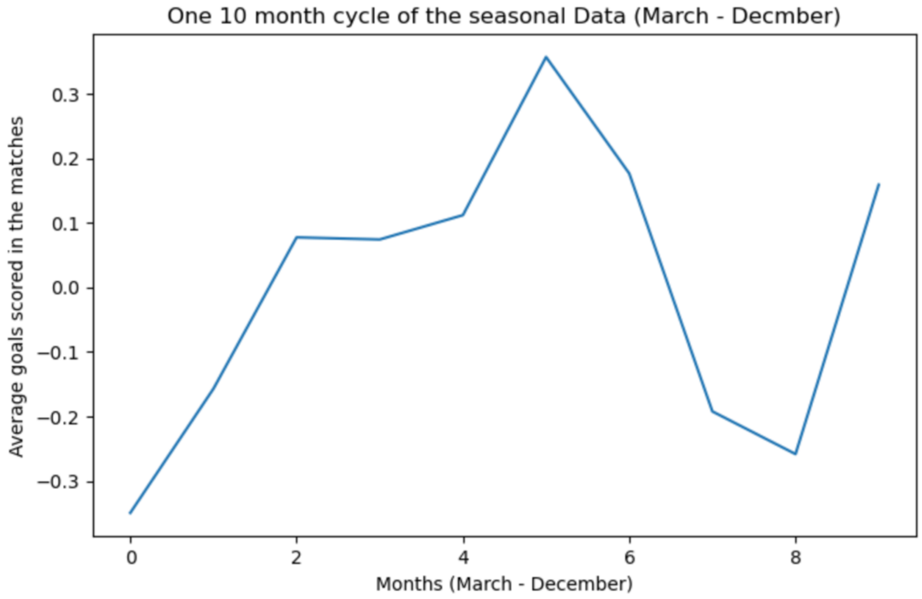


Fig 8

We see that 0 is the month of March in this one seasonal cycle and 5 is August where the goals peak.

August is typically the middle of the season, when teams have had time to settle into their playing styles and players have become more accustomed to playing together. This can lead to more cohesive team performances and more goals being scored.

Secondly, the weather conditions in August can be favourable for scoring goals. In many parts of the United States and Canada, August is a hot and humid month, which can tire out defenders and make it easier for attackers to find space and create scoring opportunities.

Finally, the schedule of games in August often includes a number of matches between teams that are fighting for playoff positions. These high-stakes games can lead to more attacking play and a greater willingness to take risks in search of goals.

A possible reason why we see the lowest scoring games in March is :
Firstly, March is the beginning of the season, and teams may still be working on building their chemistry and settling into their playing style. This can lead to more cautious play and fewer goals being scored. Additionally, the weather conditions in March can be less favourable for attacking play, with colder temperatures and, in some regions, inclement weather that can affect the quality of play.

The MLS schedule tends to include a number of international breaks in March and December, when players are called up to play for their national teams in various competitions. This can lead to disruptions in team chemistry and a lack of continuity in the starting line-ups, which can in turn affect the number of goals being scored.

November is a low scoring month since the playoffs are a high-pressure situation, and teams tend to focus more on defence rather than offense. The teams know that even one goal conceded can be the difference between advancing to the next round or being eliminated. Therefore, teams are more cautious and tend to play more defensively, which can result in lower-scoring games.

I have performed a similar sort of analyses with Home and Away goals separately to see how they individually contribute to the above conclusions, as seen below.

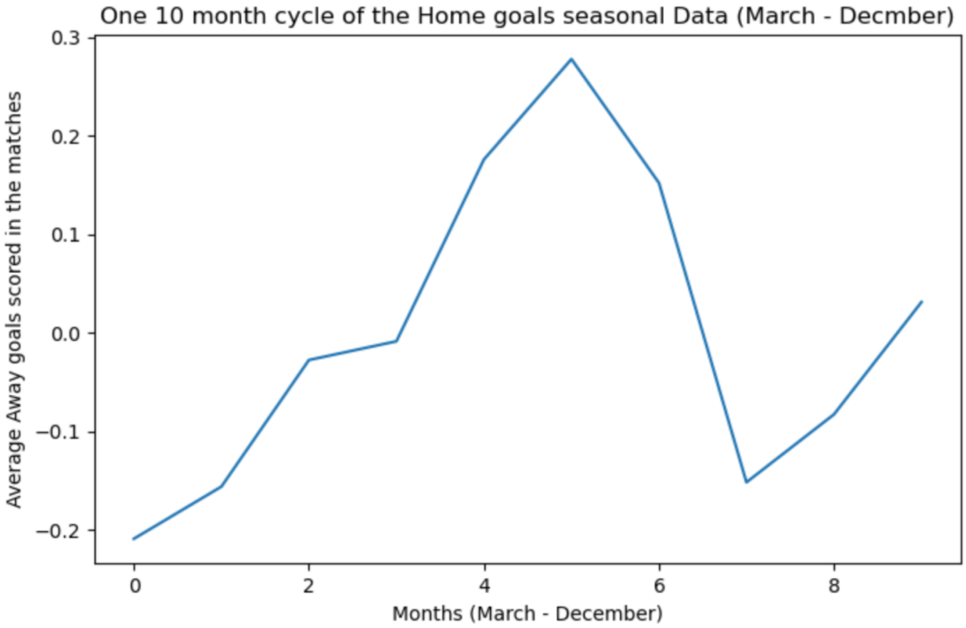


Fig 9

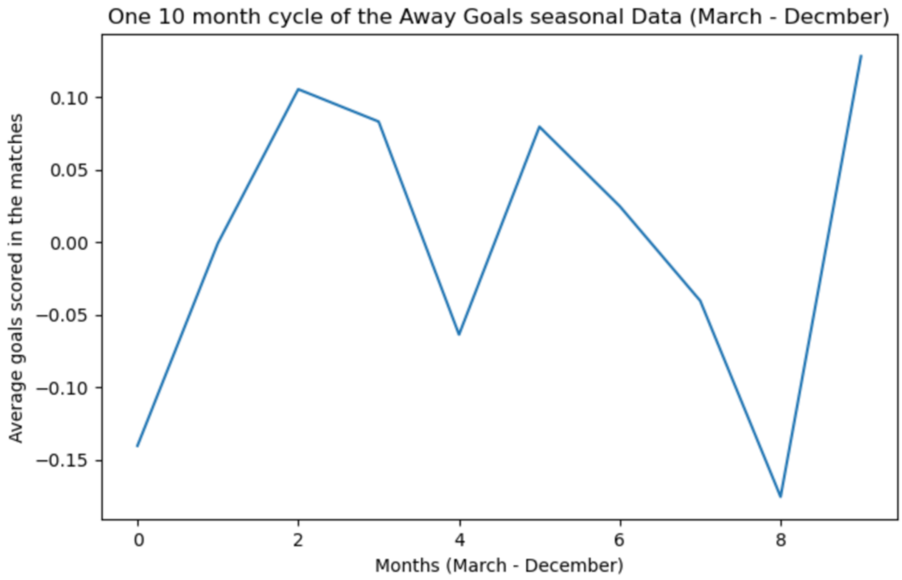


Fig 10

Analysing these graphs we see a similar behaviour especially with the Home goals. The away goals have a peak around the months of May. This does not indicate that the away teams are scoring at a higher rate compared to Home teams, this only indicates that away teams are scoring more goals. There could be multiple reasons for that like the weather conditions getting more pleasant in certain parts of North America, Travel schedule but none for certain. This could be a small factor in the sharp increase in total goals in May.

Section 3 – A simple Model

3.1

Given the equations :

$$X_k \sim \text{Poisson}(\lambda_k)$$

$$Y_k \sim \text{Poisson}(\mu_k)$$

Where

$$\ln \lambda_k = \alpha_{i(k)} + \beta_{j(k)} + \gamma + \eta/2$$

$$\ln \mu_k = \alpha_{j(k)} + \beta_{i(k)} + \gamma - \eta/2$$

The parameters α_i, β_i, γ, and η are the hyperparameters to the mentioned Poisson models with λ_k and μ_k as the mean parameter of the respective models.
X_k and Y_k represent expected number of goals by home and away teams in a football game.

The parameter of a Poisson distribution is the mean of the distribution.
For example, Given a lambda of 1.5, over large iterations a team is expected to score 1.5 goals on an average per game.

In an inference sense, these parameters could represent the following:

- α_i represents the ability of the i-th team to score. An expected number of goals from the attacking team or the average number of goals the attacking team would score in the k-th match. Since in the equations alpha is a parameter to estimate natural log of λ_k, $e^{\alpha i}$ would better represent the attacking ability of the i-th team. Higher value of α_i means higher the scoring rate of a team.
- β_i represents the defensive ability of the i-th team. The defensive ability of a team can be depicted by the average number of goal the team concedes per game. In lines with the above explanation, β_i is a parameter to estimate the natural log of λ_k (or μ_k), hence $e^{\beta i}$ would better represent the defensive ability of the i-th team.
- γ represents the average rate of scoring across all teams. This represents a baseline in the equation. Other than Home and away factors, this could capture factors in the league such as location of the game, weather conditions etc. In lines with the above explanations, γ is a parameter to estimate the natural log of λ_k (or μ_k), hence e^{γ} would better represent the average scoring rate. This can be the overall intercept term in this Poisson regression model.

- η represents the home advantage of a team. It is added to positively affect the scoring rate of the home team while it has been subtracted from the away team to negatively affect the away team of this advantage. η has been divided by 2 in both cases to not double to effect of it on both sides. Adding half to one teams scoring rate and subtracting half from one teams scoring rate leaves us with an overall impact of η being the home advantage in the whole game. η is a hyperparameter to compute the natural log of λ_k (or μ_k), hence e^η would be a better definition of the home advantage.

  I have also calculated these values in the Jupyter Notebook as per the above definitions, although these have not been used in any modelling.

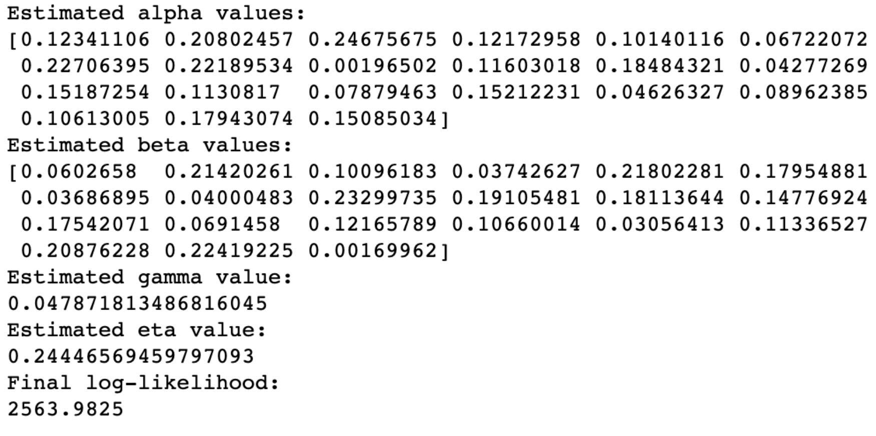
3.2

An identifiable model is a model whose parameters can be uniquely estimated from the data. In other words, the data contains enough information to estimate the model's parameters without ambiguity.

Since the model takes into account every teams strength(parameters) as relative to each other and not as absolute strengths, a sum to zero constraint could be useful on α_i and β_i i.e.: sum of all α_i and β_i are zero. This would make the model identifiable and ensures meaningful predictions.

Another option is to use a prior distribution on the parameters that imposes constraints, where the parameters are assumed to follow a distribution with a mean of zero. By introducing these constraints, we can ensure that the model's parameters are uniquely identifiable, and the estimation of the model's parameters can be carried out effectively. This needs to be done to avoid redundancy or overfitting in the estimation of these parameters

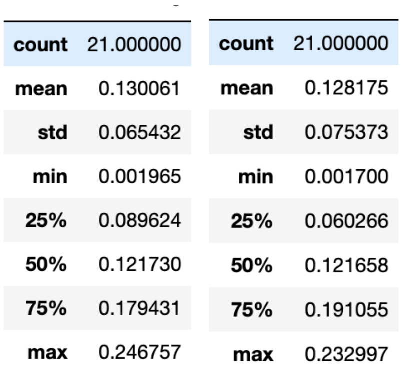
3.3



```
Estimated alpha values:
[0.12341106 0.20802457 0.24675675 0.12172958 0.10140116 0.06722072
 0.22706395 0.22189534 0.00196502 0.11603018 0.18484321 0.04277269
 0.15187254 0.1130817  0.07879463 0.15212231 0.04626327 0.08962385
 0.10613005 0.17943074 0.15085034]
Estimated beta values:
[0.0602658  0.21420261 0.10096183 0.03742627 0.21802281 0.17954881
 0.03686895 0.04000483 0.23299735 0.19105481 0.18113644 0.14776924
 0.17542071 0.0691458  0.12165789 0.10660014 0.03056413 0.11336527
 0.20876228 0.22419225 0.00169962]
Estimated gamma value:
0.047871813486816045
Estimated eta value:
0.24446569459797093
Final log-likelihood:
2563.9825
```

Fig 11

Figure 11 shows the estimated values of α_i, β_i, γ, and η and Figure 12 shows summary statistics of alpha and beta.



count	21.000000	**count**	21.000000
mean	0.130061	**mean**	0.128175
std	0.065432	**std**	0.075373
min	0.001965	**min**	0.001700
25%	0.089624	**25%**	0.060266
50%	0.121730	**50%**	0.121658
75%	0.179431	**75%**	0.191055
max	0.246757	**max**	0.232997

Summary statistic of α_i, β_i

Fig 12

The final Log-likelihood of the model is : 2563.9825

The parameters converge to values of a local minima if not a global minima.
The parameter initialisations have been using a uniform distribution of values between 0 and 0.25. This was done purely by a brute force method to obtain the lowest likelihood value possible.
I have set a seed to the NumPy module to regenerate these values. The seed is 408 which represents a random number and not a calculated one. The expression of likelihood has been derived from the Dixon and Coles paper.

$$L(\alpha_i, \beta_i, \gamma, \eta; i = 1 ,\dots k) = \Pi \{ \exp(-\lambda_k) \lambda_k^{Xk} \exp(- \mu_k) \mu_k^{Yk} \}$$

I have taken the log values to turn the product notation into summation and to get computationally efficient values.

Section 4 – Model extensions

4.1

Below are 4 extensions / improvements that could be carried out without additional date:

- **Time Dynamics** : One way to handle time dynamics could be to include a time variable. For instance, we could include a time trend variable that captures changes in the scoring rates over time. This would allow us to account for any changes in team form as the season progresses. Not just across the year time dynamics but also during the day time dynamics could be useful in successfully predicting the result of the game. For example – the weather conditions could favour teams playing in the afternoon or evening as compared to morning or late evening games.
- **Team Form** – Currently the data trains on multiple seasons of data to compute alpha and beta values for each team. But through the course of multiple seasons, a team tends to go through many changes in personal, mentality etc. Instead of calculating alpha and beta over multiple season it should be calculated on their last k games as a k moving average.
- **Home and away** – Every team plays differently at their Home and away fixtures. The factors of fans motivating or acting as an additional player or the factor of familiarity with your Home ground makes a difference. Hence having alpha and beta values for each team for Home and Away differently could give better predictions.
- **Head to Head** – It is common for a team that plays very well against most teams, to falter against one constant opponent. Hence an additional parameter to compute lambda and Mu, that takes into account how a team has performed against another certain team in the past.

4.2

Team Form

To incorporate a team's form into the mode, each team will now have 2 additional parameters. Let's call these parameter delta and phi.
Delta and Phi will be very similar to alpha and beta, representing the same thing alpha and beta represent.

Steps to incorporate 2 new hyperparameters into the model:

- First the model will calculate alpha and beta along with gamma and eta as per model A.
- Once we have the values of gamma and eta, we will use the same values of gamma and eta to calculate Delta and Phi by replacing alpha and beta from the earlier equation.
- Delta and Phi will be calculated using the same equation and given values of gamma and eta but this time taking into account the last k games of the respective teams

only. Let’s say we take 10 games as a form parameter to decide the values of Delta and Phi.

- This will be done by Creating an additional Likelihood function. This new function will estimate Delta and gamma running the loop on a truncated dataset.
- Truncating the dataset to get last 10 games of each team is fairly easy in python.
- Once we have trained to get Delta and Phi, the final value of alpha would be an average of the alpha value and the delta value for that team and final value of beta would be an average of beta and Phi for that team.
- Let’s say the original alpha value of Team A was 0.2. In the last 10 games Team A has been playing much better scoring more goals i.e.: their scoring rate or their scoring form has been much better. Due to this we compute the Delta value to be 0.24.
- Finally while making predictions we will use the alpha value of the average of 0.2 and 0.24 i.e.: 0.22. This would increase the value of Lambda and in-turn increase the outcome of their scoring rate.
- In a similar manner we can incorporate Phi and average it out with the initial Beta value.

4.3

There are a lot of additional factors that could be considered to make a model and predictions better.

Factors such as:

- The **number of key player injuries** to any team during a particular match. Key players being absent from a team could have huge impacts. Any key player has significant contributions towards the outcome of a game. That key player being missing would mean that the team is less threatening in the position of the missing player. This could compromise a team’s ability to score or defend.

- How long a coach has been in charge of the club with the same players during his tenure. Statistically speaking, a team that has lesser changing managers and lesser changing players would have less variance in the outcome of their matches. Lesser variance would lead to better predicted results for **an unchanged team.**

- Every team wants to improve and in order to do so they buy better players or sell lesser players during the transfer market. The **impact of transfers** on any squad would be a good metric to judge how a team will perform in the future games.

4.4

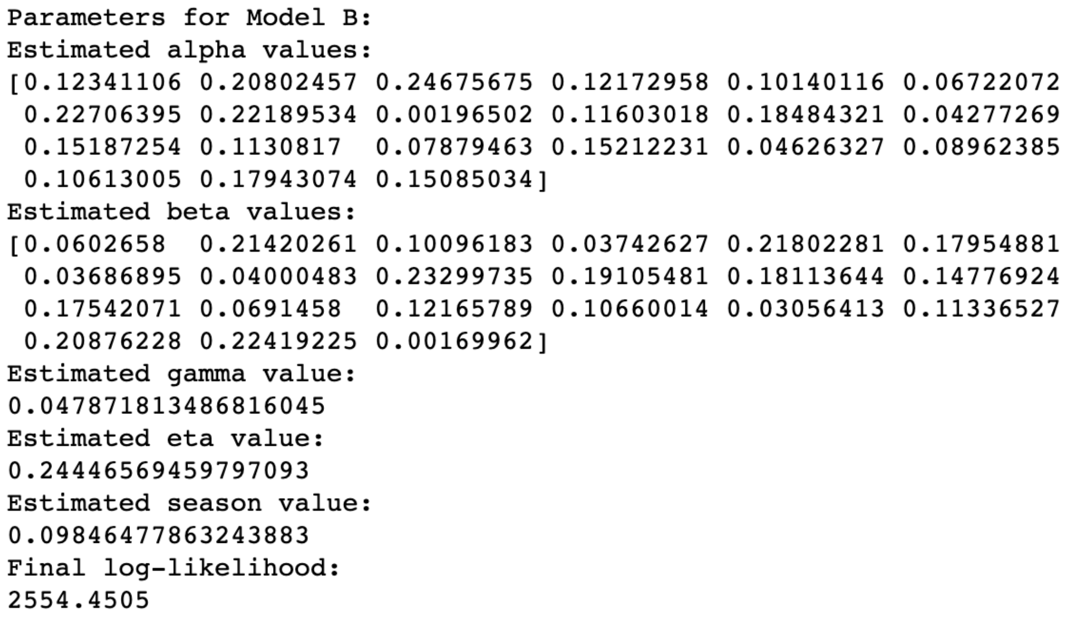
One way to fit a new model that takes the seasonal component into account is using ARIMA models to find the autoregressive and Moving average components.

A simpler way in the current case is to use a seasonal variable within our data. Given the Seasonality above, We can see that the months of March, April October and November have significantly lower goals than the average, while August and September have higher than average goals and May June and July have no shift from the average.

Hence a variable that gets subtracted in the low scoring months from our equation and gets added in the high scoring months.

We will introduce a variable called seasonal having values of {-1, 0, 1} denoting lower, average and higher vales respectively. This variable will now contribute towards the estimation of our parameters for or new model, B.

After creating and training the model, our hyper parameters are :



```
Parameters for Model B:
Estimated alpha values:
[0.12341106 0.20802457 0.24675675 0.12172958 0.10140116 0.06722072
 0.22706395 0.22189534 0.00196502 0.11603018 0.18484321 0.04277269
 0.15187254 0.1130817  0.07879463 0.15212231 0.04626327 0.08962385
 0.10613005 0.17943074 0.15085034]
Estimated beta values:
[0.0602658  0.21420261 0.10096183 0.03742627 0.21802281 0.17954881
 0.03686895 0.04000483 0.23299735 0.19105481 0.18113644 0.14776924
 0.17542071 0.0691458  0.12165789 0.10660014 0.03056413 0.11336527
 0.20876228 0.22419225 0.00169962]
Estimated gamma value:
0.047871813486816045
Estimated eta value:
0.24446569459797093
Estimated season value:
0.09846477863243883
Final log-likelihood:
2554.4505
```

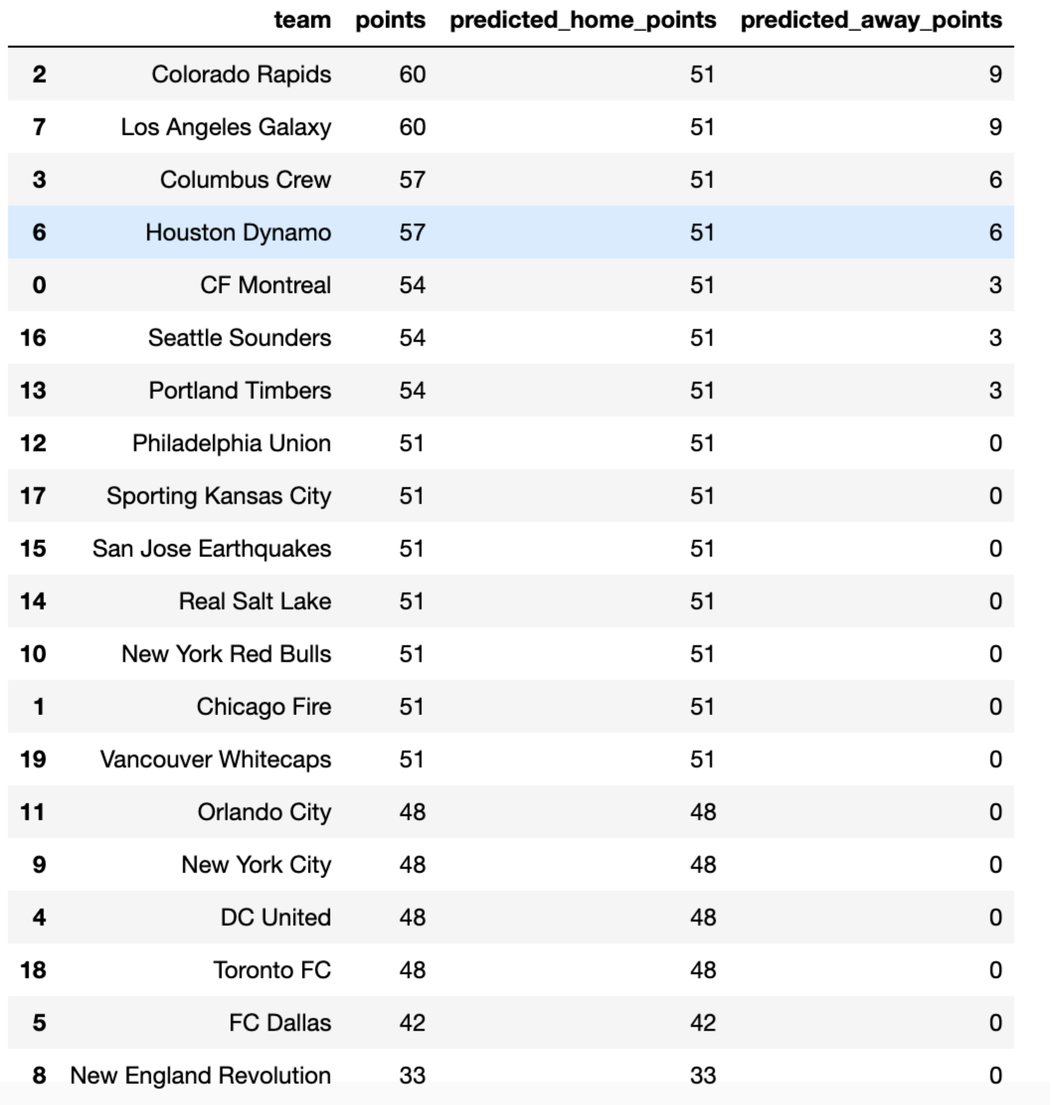
Fig 13

Checking the performance of both the models on the same Training set (fitset) we observe that the performance of the 2 models are the same. The additional seasonal component has no improvement towards the output of the model on the training set.

Hence we will be using Model A as our Best model.

5.3

The league Table was constructed as a whole and not as the eastern and western conference separately to not make any assumptions of the unsaid.
The league Table is displayed below:



	team	points	predicted_home_points	predicted_away_points
2	Colorado Rapids	60	51	9
7	Los Angeles Galaxy	60	51	9
3	Columbus Crew	57	51	6
6	Houston Dynamo	57	51	6
0	CF Montreal	54	51	3
16	Seattle Sounders	54	51	3
13	Portland Timbers	54	51	3
12	Philadelphia Union	51	51	0
17	Sporting Kansas City	51	51	0
15	San Jose Earthquakes	51	51	0
14	Real Salt Lake	51	51	0
10	New York Red Bulls	51	51	0
1	Chicago Fire	51	51	0
19	Vancouver Whitecaps	51	51	0
11	Orlando City	48	48	0
9	New York City	48	48	0
4	DC United	48	48	0
18	Toronto FC	48	48	0
5	FC Dallas	42	42	0
8	New England Revolution	33	33	0

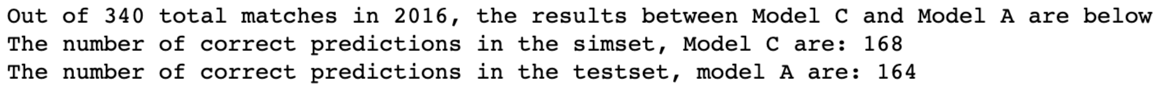
Fig 14

According to the league Table, the table topper of the league would be : **Colorado Rapids**.

5.4

simset_predictions DataFrame contains the predictions of Model C from the file.

Model A (my best model) and Model C were compared by primarily comparing the number of correct result predictions made by each model on 2016 season and the total overall expected goals computed by both models vs the actual Total Goals throughout the season 2016.



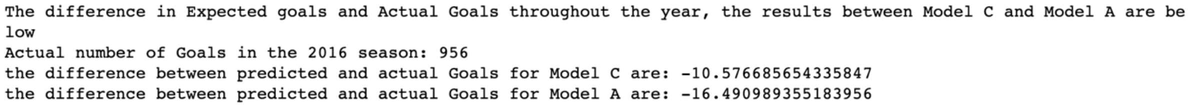
```
Out of 340 total matches in 2016, the results between Model C and Model A are below
The number of correct predictions in the simset, Model C are: 168
The number of correct predictions in the testset, model A are: 164
```

Fig 15

The above Figure depicts that model C had more correct predictions in the 2016 season compared to Model A.

Model A had a correct prediction percentage of 48.23%
Model C had a correct prediction percentage of 49.41%

In terms of win probabilities, Model C performs better than Model A by small margin.



```
The difference in Expected goals and Actual Goals throughout the year, the results between Model C and Model A are be
low
Actual number of Goals in the 2016 season: 956
the difference between predicted and actual Goals for Model C are: -10.576685654335847
the difference between predicted and actual Goals for Model A are: -16.490989355183956
```

Fig 16

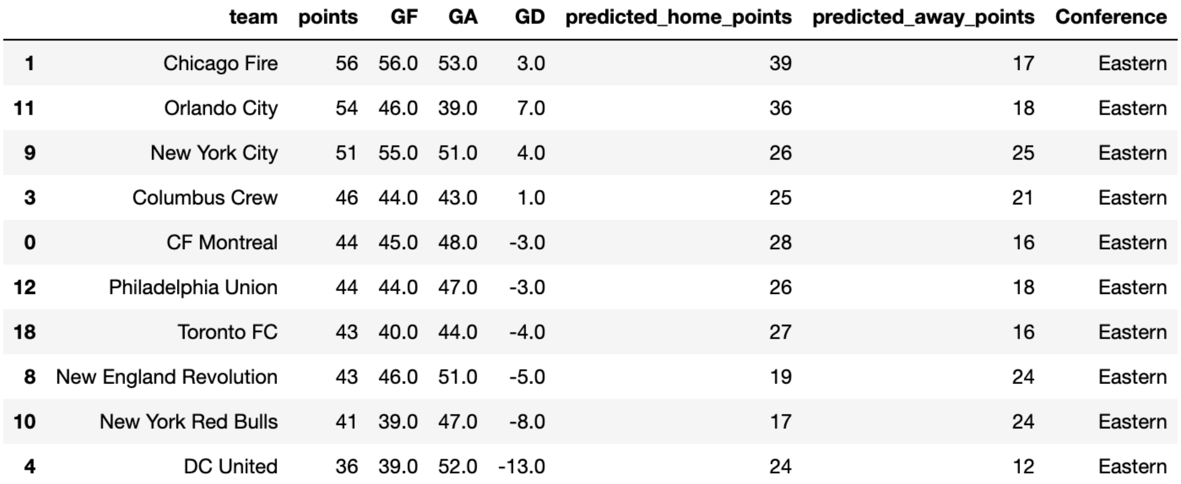
Model C predicted 945.42 expected goals by Home and Away teams combined throughout the season.
Model A predicted 939.50 expected goals by Home and Away teams combined throughout the season.
The actual goals scored in 2016 were : 956.

In terms of expected Goals, Model C performs better than Model A by a small margin.

Section 6 – Simulation

6.1

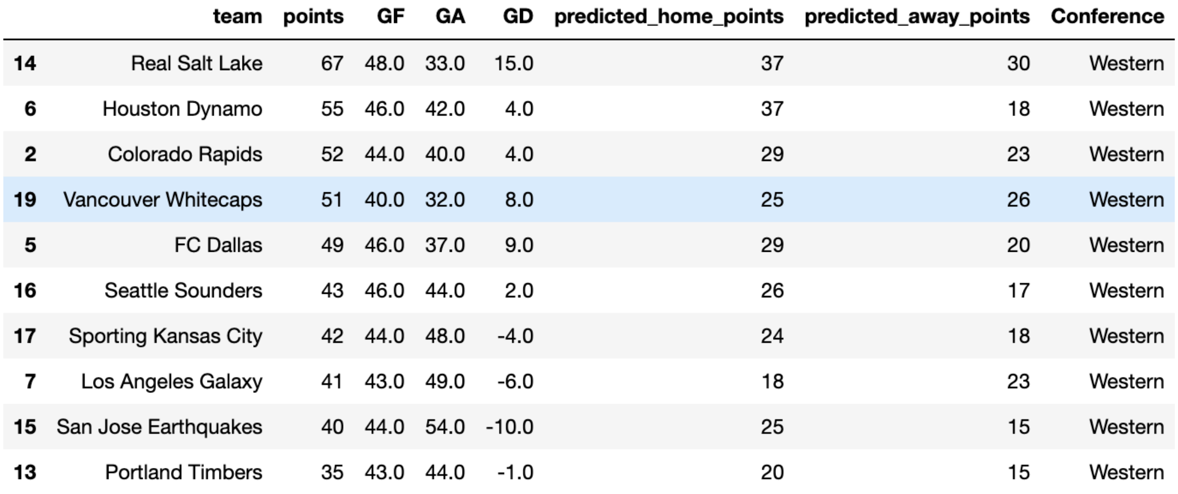
Eastern Conference League Table



	team	points	GF	GA	GD	predicted_home_points	predicted_away_points	Conference
1	Chicago Fire	56	56.0	53.0	3.0	39	17	Eastern
11	Orlando City	54	46.0	39.0	7.0	36	18	Eastern
9	New York City	51	55.0	51.0	4.0	26	25	Eastern
3	Columbus Crew	46	44.0	43.0	1.0	25	21	Eastern
0	CF Montreal	44	45.0	48.0	-3.0	28	16	Eastern
12	Philadelphia Union	44	44.0	47.0	-3.0	26	18	Eastern
18	Toronto FC	43	40.0	44.0	-4.0	27	16	Eastern
8	New England Revolution	43	46.0	51.0	-5.0	19	24	Eastern
10	New York Red Bulls	41	39.0	47.0	-8.0	17	24	Eastern
4	DC United	36	39.0	52.0	-13.0	24	12	Eastern

Fig 17

Western Conference League Table



	team	points	GF	GA	GD	predicted_home_points	predicted_away_points	Conference
14	Real Salt Lake	67	48.0	33.0	15.0	37	30	Western
6	Houston Dynamo	55	46.0	42.0	4.0	37	18	Western
2	Colorado Rapids	52	44.0	40.0	4.0	29	23	Western
19	Vancouver Whitecaps	51	40.0	32.0	8.0	25	26	Western
5	FC Dallas	49	46.0	37.0	9.0	29	20	Western
16	Seattle Sounders	43	46.0	44.0	2.0	26	17	Western
17	Sporting Kansas City	42	44.0	48.0	-4.0	24	18	Western
7	Los Angeles Galaxy	41	43.0	49.0	-6.0	18	23	Western
15	San Jose Earthquakes	40	44.0	54.0	-10.0	25	15	Western
13	Portland Timbers	35	43.0	44.0	-1.0	20	15	Western

Fig 18

6.2

Using 10000 simulations, the probability that LA Galaxy finish in the top 2 in their conference is : 0.0192

6.3

To construct a confidence interval for the probability estimate, we can use the normal approximation to the binomial distribution with mean *np* and variance as *np(1-p)* to get :

$$\hat{p} \pm z_{\alpha/2} (\hat{p} (1 - \hat{p})/n)^{0.5}$$

Since we need a confidence interval of 95%, this gives us a significance level of 5% i.e.: 0.05

The 95% Confidence Interval of LA galaxy finishing in top 2 is: [0.01651039 0.02188961]

To estimate the number of simulations needed for Monte Carlo error rate of 0.1% :

$$n = z^2_{\alpha/2}\, \hat{p} (1 - \hat{p}) / \varepsilon^2$$

where ε is the Monte Carlo error rate.

The number of simulations obtained are : **72,340**