Masters Comprehensive Examination: Report

MASTERS PROJECT:

ANALYSIS OF BARCLAYS PREMIER LEAGUE SOCCER DATA - 2005-06 TO 2014-2015(10 YEARS)

Advisor: Prof. D. Stott Parker, Jr.

Ritesh Varyani

Graduate Student

Computer Science

UCLA

1. Introduction:

In this project, we study and analyse the data for the Barclays Premier League Soccer for the last ten seasons in the United Kingdom. The Barclays Premier League Soccer is the highest level of domestic soccer competition in the United Kingdom and is widely known across the world because of its competitive nature and temperamental game. A typical season of the Barclays Premier League Soccer starts in August and ends in May. This dataset is across 10 seasons, from 2005-06 to 2014-15. The dataset has about 3800 rows representing all the matches played during the course of ten seasons. It has a lot of features, which include the month, year the match was played, the officiating referee for the match, the number of cards shown by the referee, fouls conceded by the teams, shots taken, goals scored and half-time and full-time results.

We also try to analyse and support or refute popular notions in soccer, which include theories like indirect influence in the games, how the teams fare at different times of the season and which teams are actually the resilient teams (have a winning attitude in spite of losing halfway into the game). We end our analysis with the future scope of the project and what could be further determined in styles of play of different teams or in general, English Soccer.

2. Source of the Dataset:

The data for the project is an open-source dataset taken from http://www.football-data.co.uk/englandm.php [1]. The data was not consolidated as a single data-set and appropriate cleaning, pre-processing and consolidation of the data into a single data set has been performed. This data is across 10 seasons of the Barclays Premier League Soccer in the United Kingdom. The data is from the season 2005-06 to 2014-15.

3. Data Description and Pre-Processing:

The data we analyse is of 10 seasons of the Barclays Premier League Soccer in the United Kingdom from the season 2005-06 to 2014-15. We have different .csv files representing each season of the Barclays Premier League and we have consolidated the data into one .csv file.

This pre-processing also involves, filling up NA values from the official statistics of the League. The source of these statistics is http://www.worldfootball.net/referees/eng-premier-league/ [2]. After our pre-processing we end up with data which has about 3800 rows representing the matches played over the course of ten seasons and has about 24 different features.

Each record (a match) includes the features like HomeTeam, AwayTeam, FullTimeHomeTeamScore, FullTimeAwayTeamScore, FullTimeResult, HalfTimeHomeTeamScore, HalfTimeAwayTeamScore, HalfTimeResult, OfficiatingReferee, HomeTeamShots, AwayTeamShots, HomeTeamYellowCards, AwayTeamYellowCards among others.

4. Objectives of Analysis of the Data

For this project, our goals are manifold. We try to gauge different parameters which have affected the teams throughout the ten seasons in the League. We try to reason whether there are any indirect referee influences on the game depending on the different refereeing styles-number of cards shown, number of fouls given. We try to find the teams which are resilient in nature and have better results in spite of losing first half of the game. We also try to analyse how different teams fare during different times in their season.

5. Analysis of the Data:

We have about 36 different teams which have been a part of the league for the two seasons. However, only 8 of the teams have been consistent throughout the ten seasons. Our main analysis lies on these 8 different teams where we compare how these teams have performed in the league for ten seasons and also against each other. We split our analysis into different modules and draw conclusions from each of these analysis.

Module 1: Find Consistent Teams and the Goals scored by them in all Seasons

In our first part of analysis after pre-processing we try to determine the teams which have been consistent throughout the ten seasons. It turns out, that from 36 teams, only 8 have been successful in not getting relegated during those seasons and have played at this top level consistently. We also find the goal scored by these teams across ten seasons and analyse the plot.

Sr. No.	List of Consistent teams across 10 seasons
1	Arsenal
2	Aston Villa
3	Chelsea
4	Everton
5	Liverpool
6	Manchester City
7	Manchester United
8	Tottenham

figure 1: List of Consistent Teams across ten seasons(teams which have played for all ten seasons)

We get the consistent performers of the league and have tabulated the result above. These teams have appeared in all ten seasons of the league and have never been relegated during all seasons.

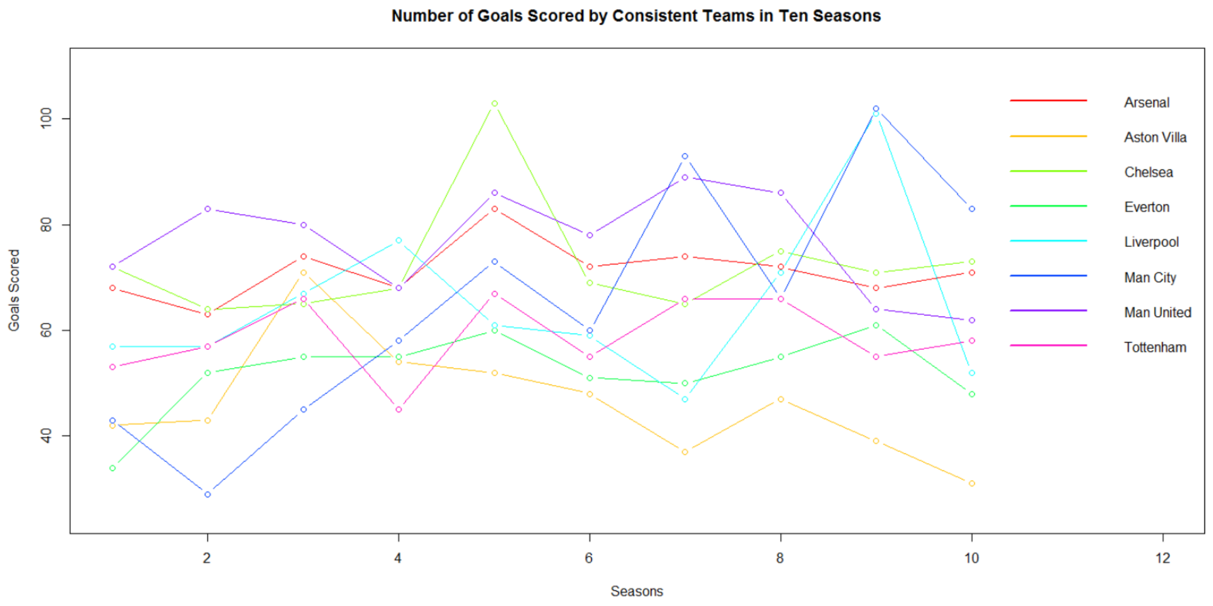
Number of Goals Scored by Consistent Teams in Ten Seasons

figure 2: Number of Goals Scored by Consistent Teams in Ten Seasons

We have the following graph which has number of goals scored across ten seasons by the teams which have been consistent throughout the ten seasons.

We see from the graph that mid-table teams like Aston Villa have consistently performed low in terms of scoring, especially for the last 6 seasons. Another key observation from the graph is the overall scoring of goals for all the 8 teams has slowly increased, which shows that teams have gone on a more attacking approach, from one season onto the next.

Module 2: Study of Resilient Teams

We define Resilient Teams as teams which are losing at half time and win the game at full-time. For this, we get the number of matches for all the 36 teams for which this has happened. We plot the following graph.

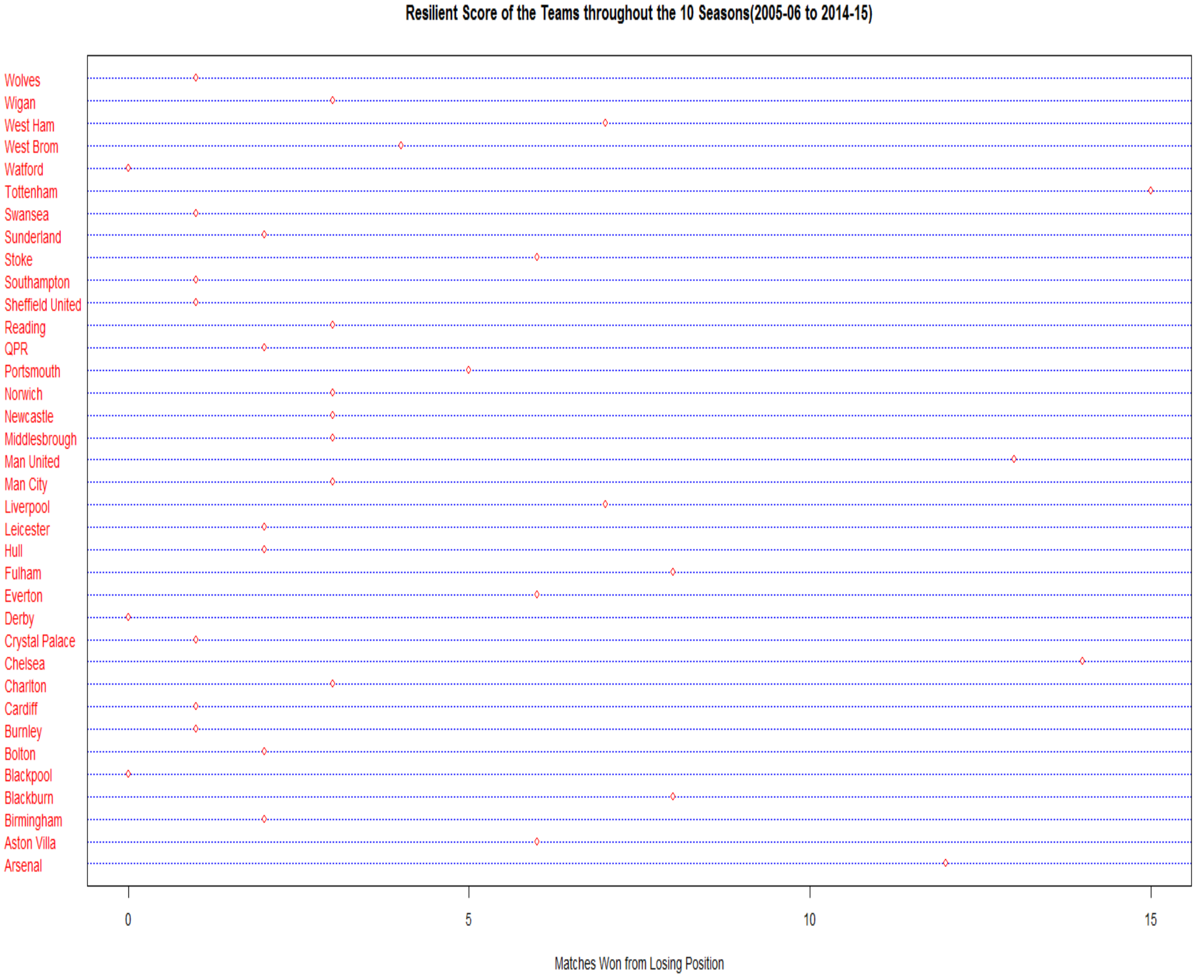

figure 3: Resilient Score of the Teams for the 10 Seasons.

From the graph, we see that the most resilient teams are actually the consistent performer s of the league. Teams like Arsenal (12 wins), Chelsea (14 wins), Manchester United (13 wins) and Tottenham (15 wins) are the best resilient teams across ten seasons.

We also make see that the teams like Blackpool, Derby and Watford have all had 0 wins from a losing position.

Module 3: Discipline Record for Teams

Yellow and Red Cards in Soccer play a very important role. Some of the teams actually base their team strategies based on set-pieces like free kicks and penalties, from which they try to score and also play in a way in which the other team concedes a foul to them.

However, from the results that we have obtained from the consistent teams, we do understand that the top-flight teams; teams which are consistent throughout the ten seasons, almost have defaulted more or less in the same way. We plot the two tables with Yellow and Red Cards during the game:

Team	Number of Yellow Cards Earned for all Ten Seasons	Number of Yellow Cards Earned per Game
Arsenal	586	1.542
Aston Villa	643	1.692
Chelsea	617	1.623
Everton	577	1.518
Liverpool	538	1.416
Manchester City	625	1.645
Manchester United	582	1.532
Tottenham	557	1.466

figure 4: Number of Yellow Cards Earned for Ten Seasons and Per Game by the Consistent Teams

Team	Number of Red Cards Earned for all Ten Seasons	Number of Red Cards Earned per Game
Arsenal	33	0.0869
Aston Villa	25	0.0678
Chelsea	38	0.1
Everton	27	0.0711
Liverpool	24	0.0632
Manchester City	34	0.0895
Manchester United	27	0.0711
Tottenham	31	0.0812

figure 5: Number of Yellow Cards Earned for Ten Seasons and Per Game by the Consistent Teams

Module 4: Referees

In this analysis, we see how much are referees are stringent with issuing cards to players. There are often times where the issuance of cards have changed the course of the games. Here, we try to find out what is the exact variance between different referees in terms of cards shown. We consider only the experienced referees.

We define the experienced referees as the referees who have officiated in more than 100 games. There are a total of 19 such referees. We compare these performance with the help of following two plots.

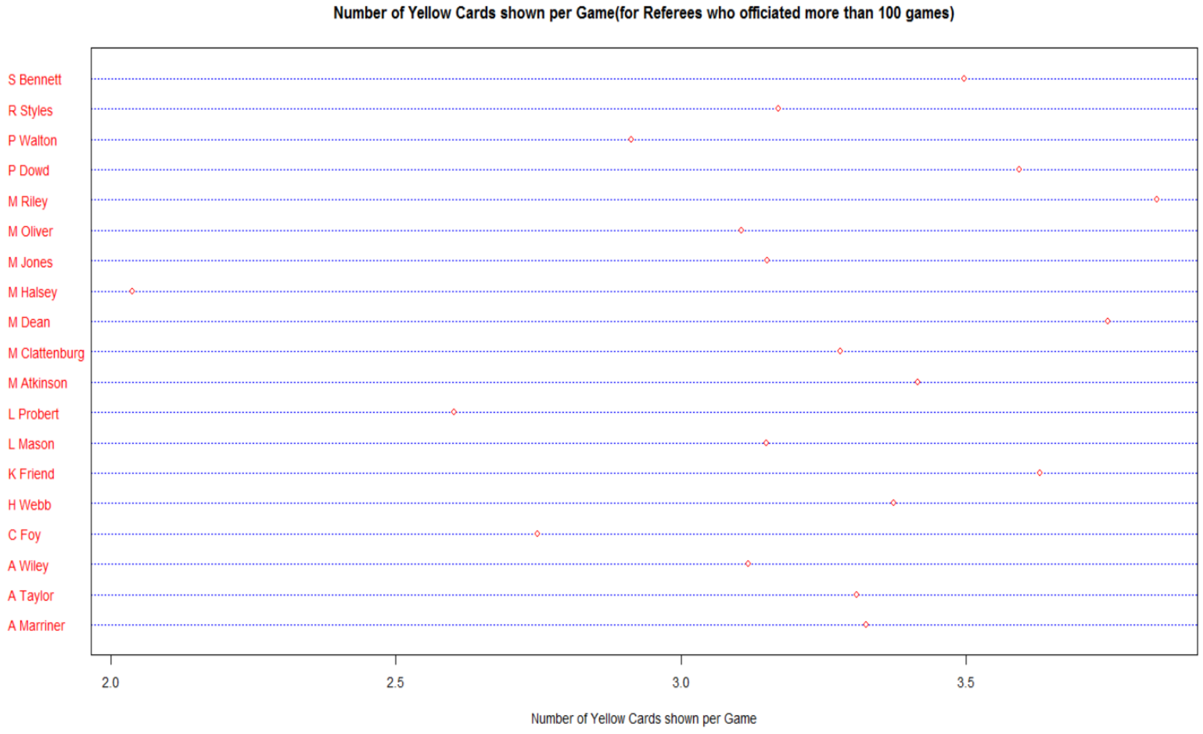

figure 6: Number of Yellow Cards shown per Game by the Referees who have officiated more than 100 Games

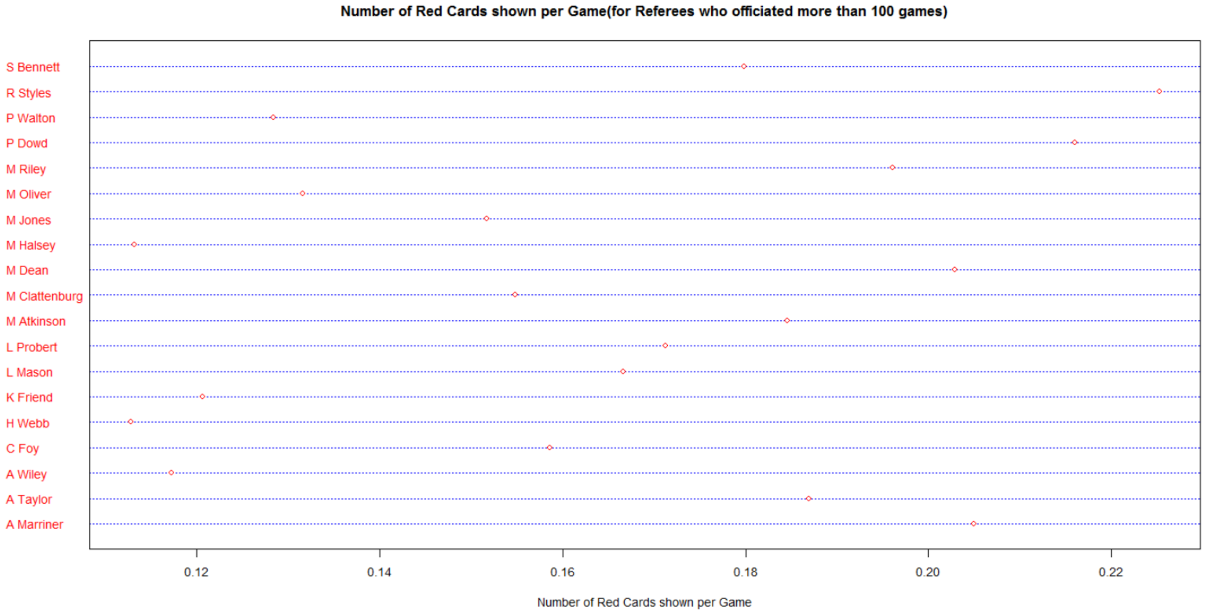

figure 7: Number of Red Cards shown per Game by the Referees who have officiated more than 100 Games

From our observations of the two plots, we see that one referee in particular- P. Dowd has given out a huge number of red and yellow cards as well. On the other end of the spectrum, we see that M. Halsey has given out the least number of red as well as yellow which actually goes to show the huge variance in terms of how stringent the referees are.

The overall variance is anywhere between 2 to 4 yellow cards and 0.10 to 0.25 red cards issued per game by the referees.

Module 5: Fouls Conceded

As we have talked about, the number of fouls conceded sometimes align with the interests of the opponent team. Over here we analyse, the mean number of fouls conceded by the consistent teams for all the ten seasons, when they have played as a Home Team as well as an Away Team.

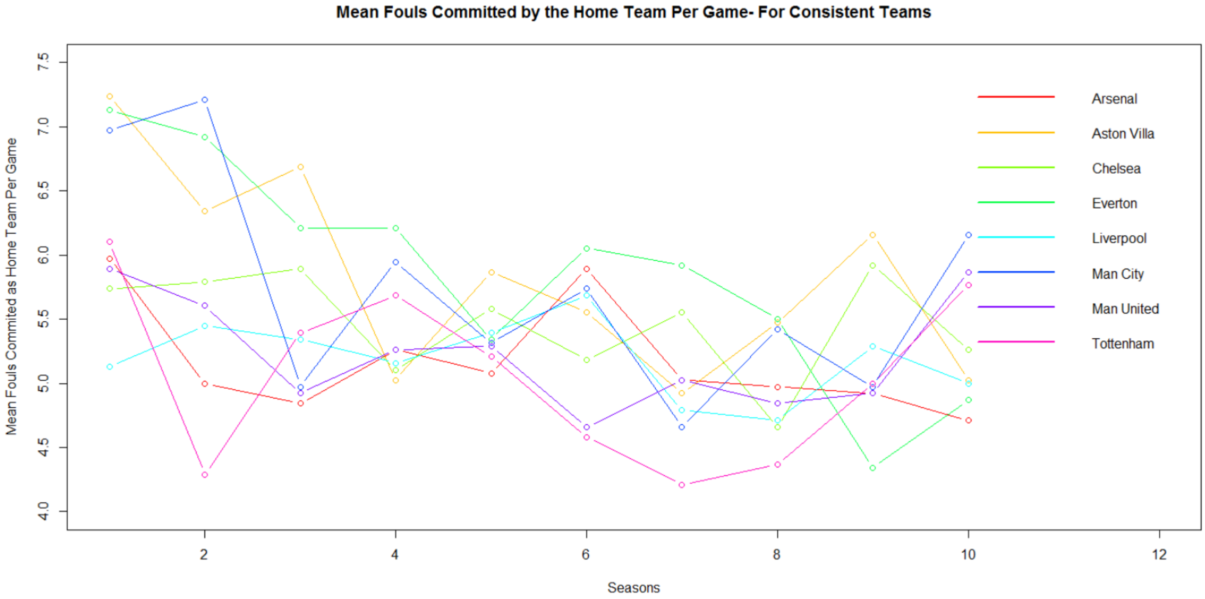

figure 8: Mean fouls committed by the consistent teams when they played a home game(for 10 seasons)

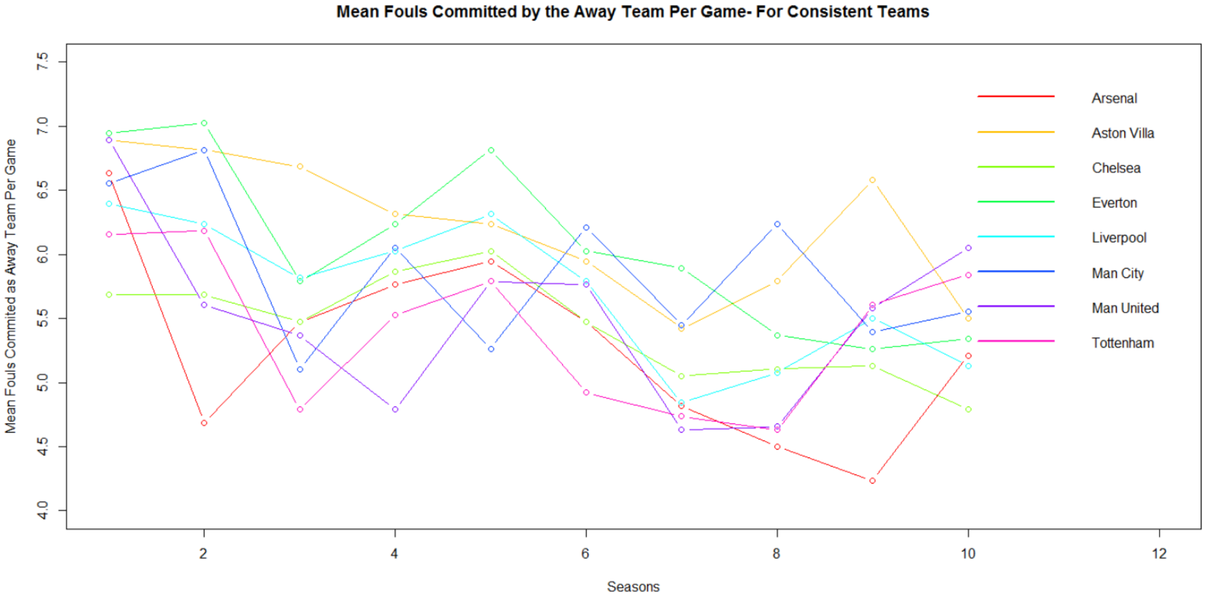

figure 9: Mean fouls committed by the consistent teams when they played an away game(for 10 seasons)

If we compare the two graphs, we do observe that the number of fouls conceded by a single team when they play at home or away are almost similar. One interesting trend which we observe is as the seasons have progressed, we see a decline in the number of fouls conceded per game and it shows, that teams have become more disciplined and stream lined in their approach.

This can also be interpreted as the referees becoming more tolerant and ‘advantages’ are being given out, potential fouls are not being given- defining the ‘English style of play’.

Module 6: Monthly Analysis of the Consistent Teams

Our last part of analysis involves monthly analysis of the winning percentage of the teams across ten seasons. In Soccer, there is a very popular notion about teams which perform well and certain months and fail to perform up to expectation I others. We try to support or refute such theories with our analysis with monthly analysis of 5 out of the 8 consistent performers in the league.

Arsenal:

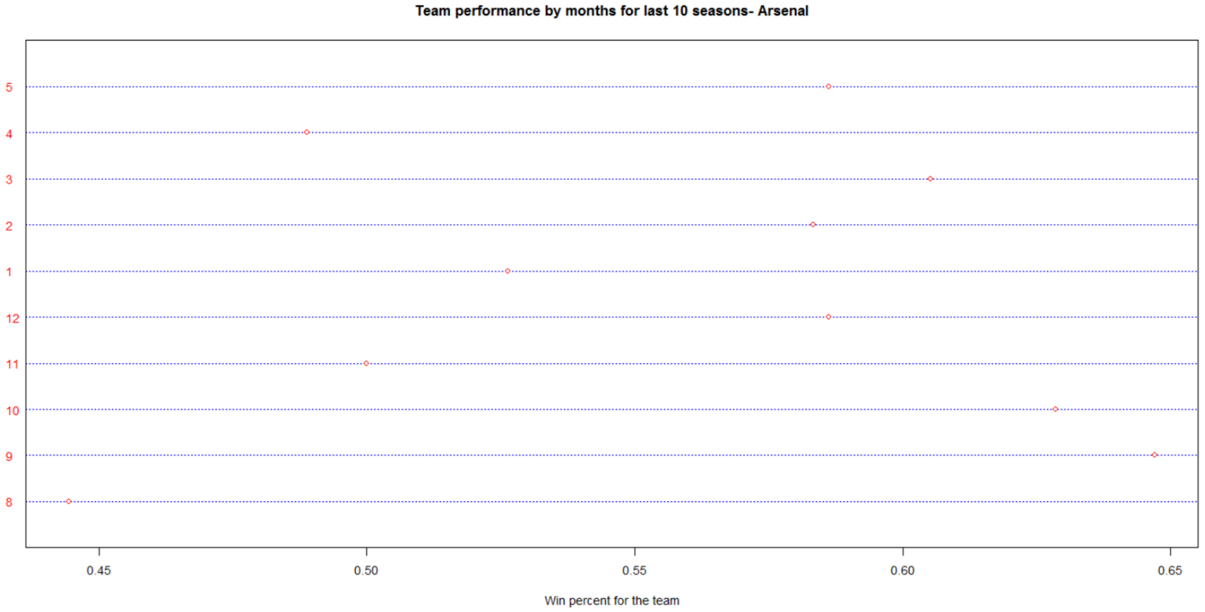

figure 10: Win percentage of team Arsenal by months for the last ten seasons

From the above plot about Arsenal, we see that they are more or less consistent performers with win-percentage between 49% to 65% barring one month(8)- August, during which they have a win-percentage of less than 45%. This shows, that they often have bad starts to the season and an average finish with ~58% win percentage in the last month (5) May.

Chelsea:

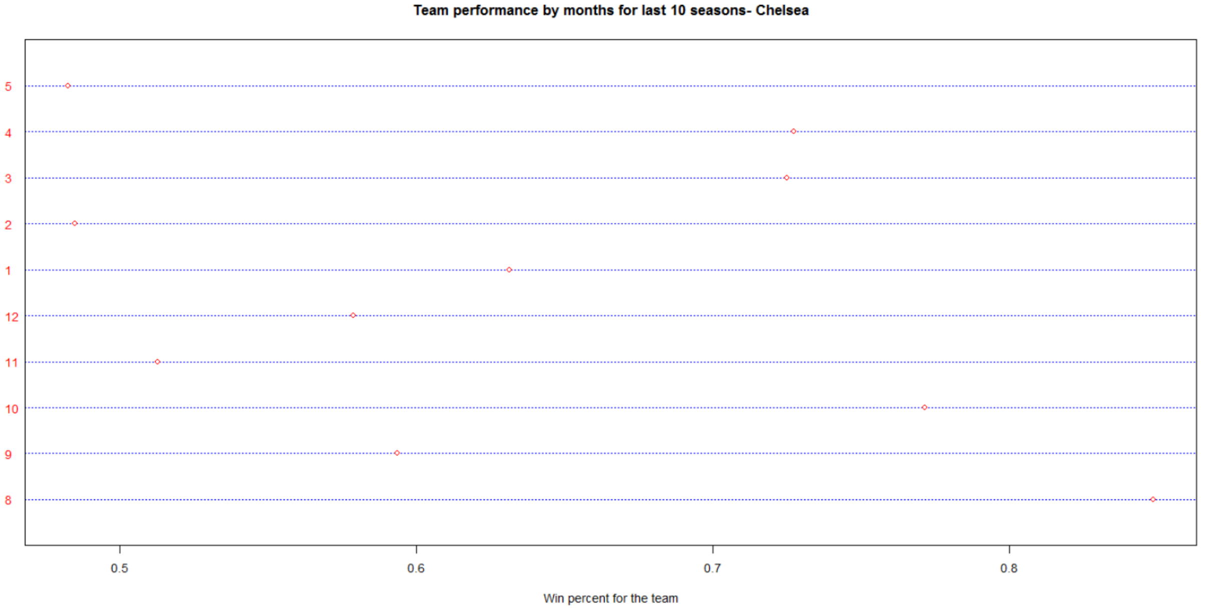

figure 11: Win percentage of team Chelsea by months for the last ten seasons

Chelsea are one of the teams whose win percentage never drops below 48% for any month and goes as high as ~84%. They are one of the consistent performers of the league. They do tend to end their season with not a great winning percentage (about ~48%) but they are one of the best starters in terms of winning(a win percentage of ~85% in the month of August).

Liverpool:

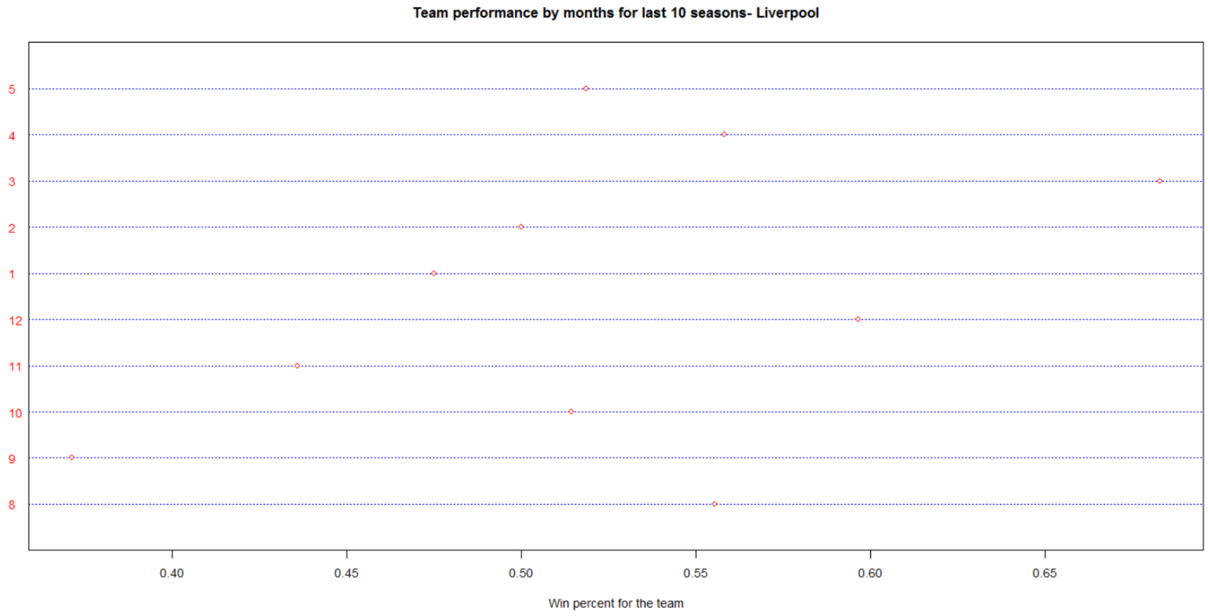

figure 12: Win percentage of team Liverpool by months for the last ten seasons

Liverpool are historically known as one of the best clubs but recently, their form dips every other season. They also have a large variance in their winning percentage from ~35% to ~70%. Their average win percentage is 50-55%.

Manchester United:

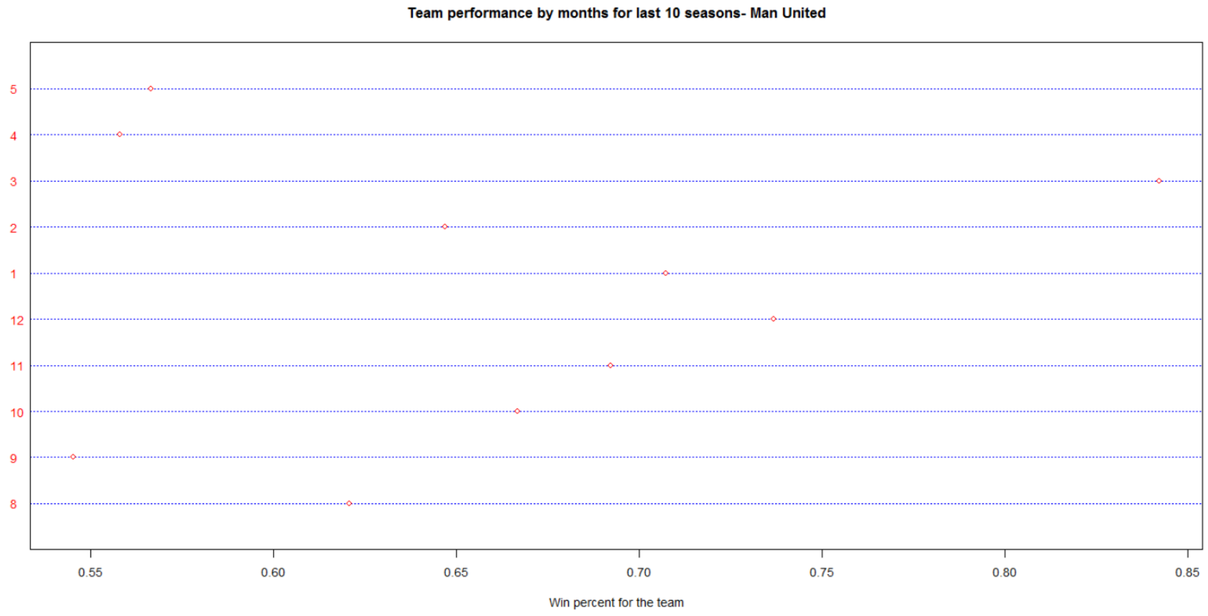

figure 13: Win percentage of team Manchester United by months for the last ten seasons

Manchester United are one of the teams who have consistently performed well over a period of 20 years. Their win percentage for any month are no lower than 50% and they win as high as ~85% of their games. Their average win percentage is ~70% which is evident of the fact that they have performed consistently well. Although they end the season during the last two months at a low winning percentage, that might be an indication of experimentation or giving the fringe players a go to have some match practice, because of the massive lead that they may have achieved over their opponents throughout the season.

Tottenham:

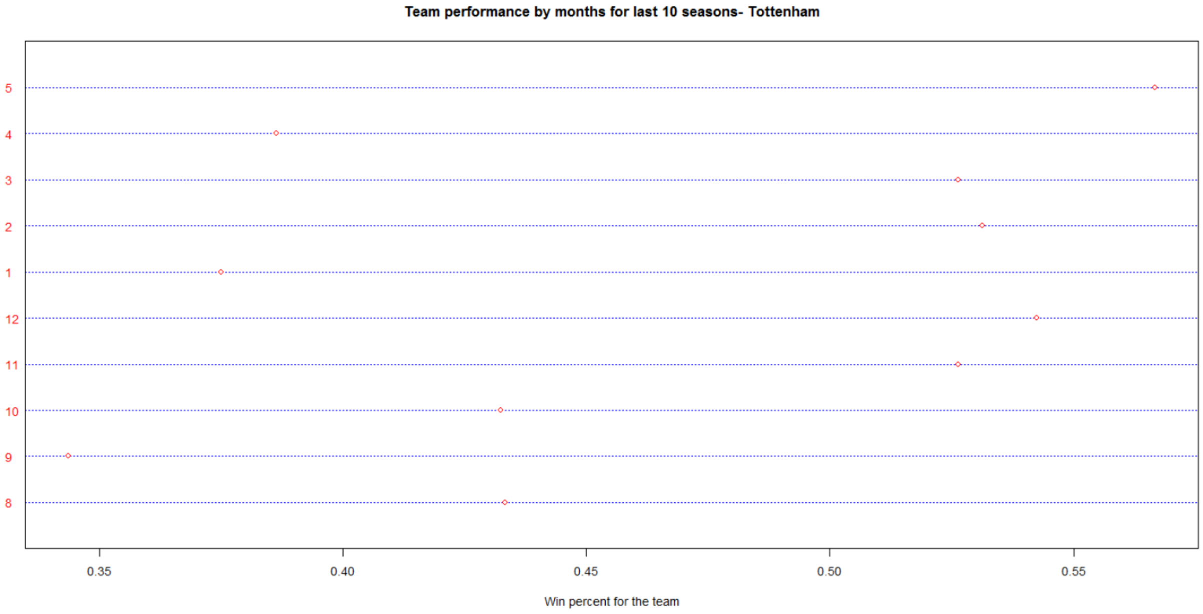

figure 14: Win percentage of team Tottenham by months for the last ten seasons

Tottenham are one of the average players among the consistent teams in the league and that is actually emulated well by their win-percentage graph for different months. They tend to perform really well or go downhill altogether. Their win-percentages are varied from ~34% to ~57%.

Conclusion:

Throughout the project, we try to analyse the different and unique aspects of the game. There is a lot written on specific theories in soccer, but it is hardly backed by data and relies more on intuition. We analyse those issues and try to refute or support them with our analysis. There are a lot of open-source data available on soccer for many years; but a lot of them are inconsistent and sometimes, have missing information. We have bridged all those gaps, by consolidating one single database.

We have analysed our project on different levels and have had some unique and uncommon observations from resiliency results and also with monthly analysis of different teams for the ten seasons. There were a total of 36 different teams which had participated in the league for the last 10 years. But only 8 were a part of it for all the ten seasons. We call these teams as ‘consistent teams’. This describes how much is the league dynamic. The resilient plots showed us that the consistent teams are actually very resilient and there are 5 consistent teams in the top 8 teams which show resiliency. Tottenham, a team which has never won the league throughout the last 10 seasons, is the most resilient of all teams, winning 15 matches from a losing position. They are a team which are not highest scorers in any seasons, but their mentality has kept them in the fold and having a winning attitude, which is what have made them a consistent performer.

From the discipline record of the consistent teams, we do see that these teams rake almost the same in terms of cards received per game. However, Chelsea has a significant difference in terms of number of red cards received per game. They receive an average of 0.1 red cards per game as compared to ~0.08 red cards for other consistent teams. Also, they are third in terms of yellow cards received; which shows that overall, Chelsea have been the least disciplined of the teams that we analysed.

In terms of fouls conceded, teams like Tottenham and Arsenal have conceded the least number of fouls from the consistent teams, both home and away, which indicates that they rely more on strategic and composed game-play rather than a ruthless and foul-conceding one.

Referee decisions directly or indirectly play an important part in soccer, because they are instantaneous (no television replays) and cannot be changed. We have seen from our analysis that how the referees are stringent in giving out cards varies individually. Referees like M. Halsey do not give out cards often and let the game play flow; however, a referee like P. Dowd has given out more red and yellow cards(combined) than any other referees, which indicates frequent stoppages in play and more of a non-direct impact(by a referee) on the game.

Our monthly analysis of teams for last 10 seasons brought a string of interesting conclusions. A famous theory which indicates that Arsenal wear out during the second-half of the season is actually refuted by the analysis. Although, their win-percentage goes down drastically during the month of April, overall they are a stable team during the second half of the season. However, they are the worst starters of the season with a very low winning percentage. The win-percentages of teams like Chelsea are always in the higher ranges and they always start their season with a very high win-percentage. Historically great teams like Liverpool have performed mediocrely during the last 10 seasons; whereas a team like Manchester United almost, consistently perform well throughout the seasons, barring a couple of months. Like Liverpool, Tottenham has been average but they try to maintain their winning momentum by trying to continue their winning form for consecutive months; however, when they tend to lose, they lose big, which makes them an overall average performer in the league.

These results as graphs and show us how the teams have performed for the last 10 years and gives us a fair idea of what the team ideologies are and how they tend to play a season.

Future Scope of the Project

We have analysed different aspects of the data in our project. However, there is a lot more which can be worked on. In soccer, it is very common to have a style of play. In fact, more often than not, teams are characterised by their style of play in the league. The styles could be ‘direct’ where in the teams relentlessly attack at the risk of getting scored against, ‘tiki-taka’ where the game relies on possession of the ball and making short passes, wearing the opponent out and then scoring, or ‘parking the bus ’ where teams defend throughout the game and look to score only on counter-attacks or set-pieces. From these statistics the styles of play of different teams can be predicted and then verified with what style the team employs actually. We can also study the indirect referee influences on the game and exactly how have the game results been affected because of them.

References:

[1] England Football Results Betting Odds | Premiership Results & Betting Odds. 2015. England Football Results Betting Odds | Premiership Results & Betting Odds. [ONLINE] *http://www.football-data.co.uk/englandm.php*

[2] Premier League 2015/2016 » Referees. 2015. Premier League 2015/2016 » Referees. [ONLINE] *http://www.worldfootball.net/referees/eng-premier-league/*