Introduction

There is so much information available now that the challenge [sic] is in deciphering

what is relevant. The key thing is: What actually wins [soccer]

matches? – (Magowan,2011)

Soccer is the most played and watched team sport in the world by a large margin (Dawson et al., 2007). Professional soccer league takes the form of a round-robin tournament in which each team plays each other team both at home and away, with three points awarded for each win and one for each draw. The winner of a league is the team with the most points at the end of the season. In 2014-2015 season, the English Premier League (UK) was watched on television by a cumulative audience of 3 billion (Elder, 2017). Soccer's global popularity continues to rise, as it attracts more fans and investors around the world (Dawson et al., 2007).

No large, freely available data set recording match data beyond the result exists (McCaskill, 2016). Historically, the data published for soccer games in even the most watched leagues in the world has been limited to the result, the teams involved, and the number of goals scored by each team.

For the analysis, English Premier League data was aggregated from various online portals. The data was studied, manipulated & analyzed with the objective to showcase the efficacy of different data manipulation & analysis techniques in answering the following business questions:

- Has the soccer playing style evolved over the years? Are there more draws now than before?
- Does change in ownership of a club &/or arrival of a new manager for a team impact its performance?
- How significant is the home field advantage (Playing games at home ground) in the English Premier League now?

The Dataset consists of 194040 observations spread out over 129 years (1888 - 2016). The table below [Table 1] shows the 12 variables in the dataset and their description.

For this analysis both Season and the Team are used as dimensions. Primary variables Result, Goal Difference, FT, Total Goals in combination with Win %age and Draw %age, which are derived via data manipulation techniques are used as measures. The data is available in differently formatted .csv files (129), thus the first task was to aggregate the entire data and create a Data Warehouse to extract relevant data. Congruently, for successful analysis it is prudent to have uniform formatting throughout the data and this was achieved using R Programming Language.

Variables	Description
Date	Match Date (yyyy-mm-dd)
Season	Match Year
Home	Home Team
Visitor	Away Team
FT	Final Score
Home Team Goal	Goals Scored By Home Team
Visiting Team Goal	Goals Scored By Visiting Team
Division	League Division
Tier	League Tier
Total Goals	Total Number of Goals Scored
Goal Difference	Difference in goals scored by Home Team and Visiting Team
Result	Full Time Result (H=Home Win, D=Draw, A= Away Win)

Table 1: It shows the variables present in the dataset with their descriptions

Results:

This study followed a structured approach to drive the solutions for the business questions. The unstructured data was first processed using the R programming language to address the objectives, such as, loading & extracting all the data from .csv files, transforming the data to have identical formatting, removing the missing values & finally saving the aggregated data in a single .csv file for further manipulation.

Data Preparation:

The dataset is a list of the results of each game played in the soccer league for each season. For many purposes, especially for looking at performances by teams over seasons, the data was aggregated with Season as dimension. First, the observations with missing value were removed from the dataset. The dataset was then categorized to depict a league table for each season. In total, 129 league tables were created for each season from 1888 till 2016. To put it in a context, a single season consist of 380 games between 20 teams. The manipulated dataset was then aggregated and indexed in such a way that, a year (e.g., 2013) could be used to retrieve the league table for that year (season) [see Table 2].
Further, for analysis of individual team performances which is pre-requisite to answer the business questions, the dataset was exported and loaded as a SQL database.

Team	Games Played	Win	Draw	Lost	Goals Scored	Goals Given	Goal Difference	Points	Position
Manchester City	38	27	5	6	102	37	65	86	1
Liverpool	38	26	6	6	101	50	51	84	2
Chelsea	38	25	7	6	71	27	44	82	3
Arsenal	38	24	7	7	68	41	27	79	4
Everton	38	21	9	8	61	39	22	72	5
Tottenham Hotspur	38	21	6	11	55	51	4	69	6
Manchester United	38	19	7	12	64	43	21	64	7
Southampton	38	15	11	12	54	46	8	56	8
Stoke City	38	13	11	14	45	52	-7	50	9
Newcastle United	38	15	4	19	43	59	-16	49	10
Crystal Palace	38	13	6	19	33	48	-15	45	11
Swansea City	38	11	9	18	54	54	0	42	12
West Ham United	38	11	7	20	40	51	-11	40	13
Sunderland	38	10	8	20	41	60	-19	38	14
Aston Villa	38	10	8	20	39	61	-22	38	15
Hull City	38	10	7	21	38	53	-15	37	16
West Bromwich Albion	38	7	15	16	43	59	-16	36	17
Norwich City	38	8	9	21	28	62	-34	33	18
Fulham	38	9	5	24	40	85	-45	32	19
Cardiff City	38	7	9	22	32	74	-42	30	20

Table 2: The English Premier League Table for the 2013 Season. Similarly, 129 league tables were created for each year ranging from 1888 - 2016.This table was created using *SSRS tool* in Visual Studio.

- **Has the soccer playing style evolved over the years? Are there more draws now than before?**

In 126 Years (1888-2014), English Football Has Seen 13,475 Nil-Nil Draws (Roeder, 2014). In 1890, just 17 percent of games were drawn, and in 1977, 626 games out of 2,028, or 31 percent, were draws [Figure. 1]. The proportion of draws climbed initially until the First World War. Afterwards it remained relatively constant till the 1960s, when it rose again.

While this number is down slightly today, 26% games were drawn in 2016, we're near the historical high.

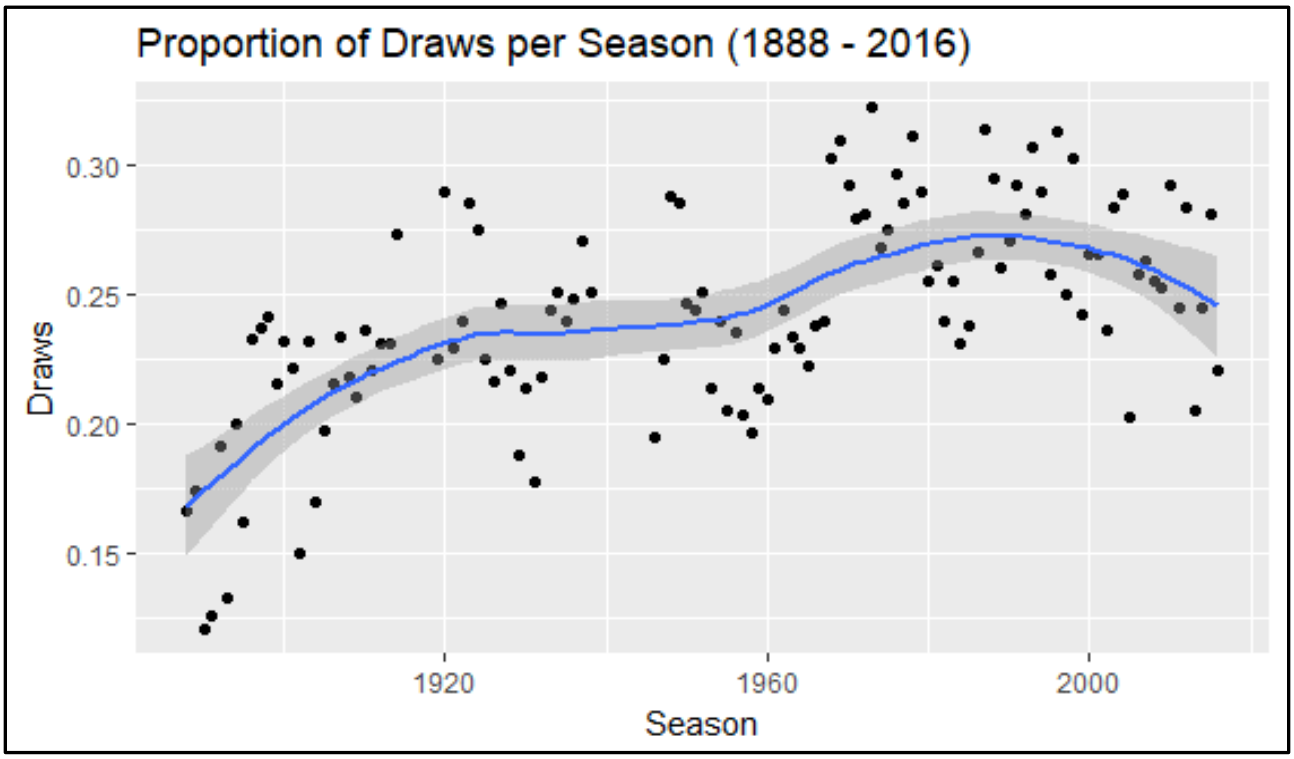

Figure 1: The proportion of draws per game over time for the top tier of the English league is shown with a loess smoother.

- **Does change in ownership of a club &/or arrival of a new manager for a team impact its performance?**

Figure 2a., below shows the win percentages of Manchester United for last 129 years. Sir Alex Ferguson, who was appointed the manager of Manchester United in 1986 till 2013, is often deemed as the best manager the premier league ever had. Manchester United won all major trophies, 38 out of total 41 ("Manchester United F.C.", 2017), during his 26 years with Manchester United.

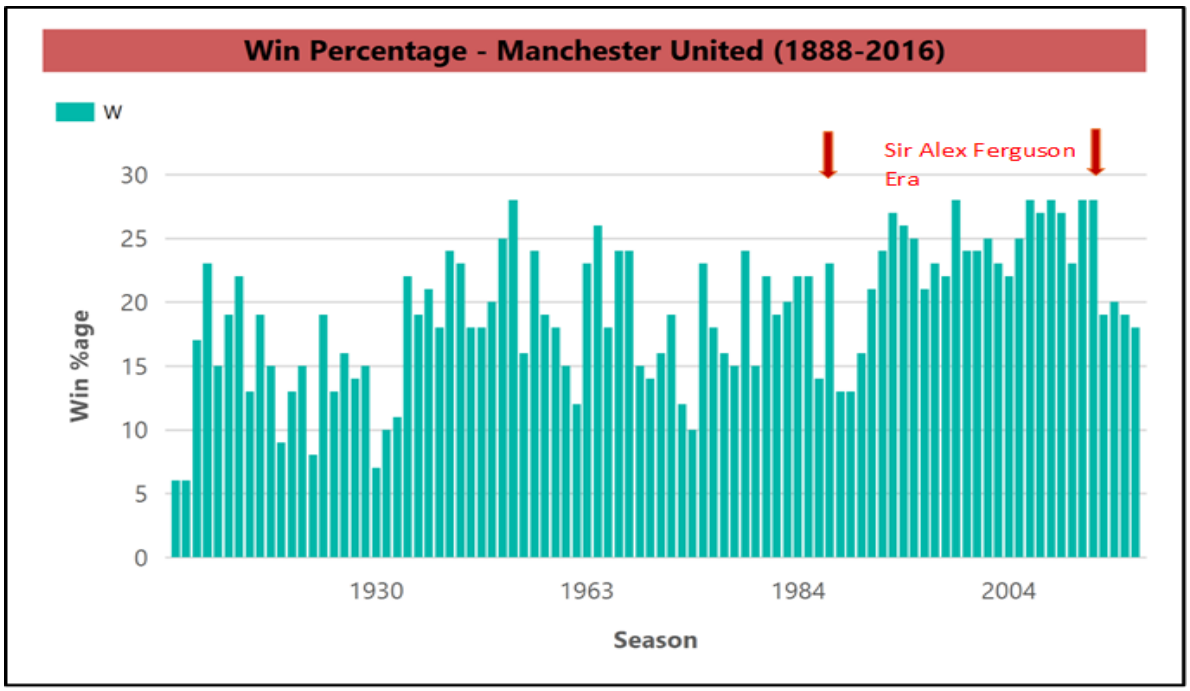

Figure 2a: The Figure shows the Manchester United Win %age (i.e. Games won per Season / Games played per Season) for each season from 1888 – 2016 *using SSRS*.

Figure 2b. shows the changes in fortune of Chelsea F.C. ("Chelsea F.C.", 2017) when multi-billionaire Roman Abramovich became the club owner in 2008 and invested heavily.

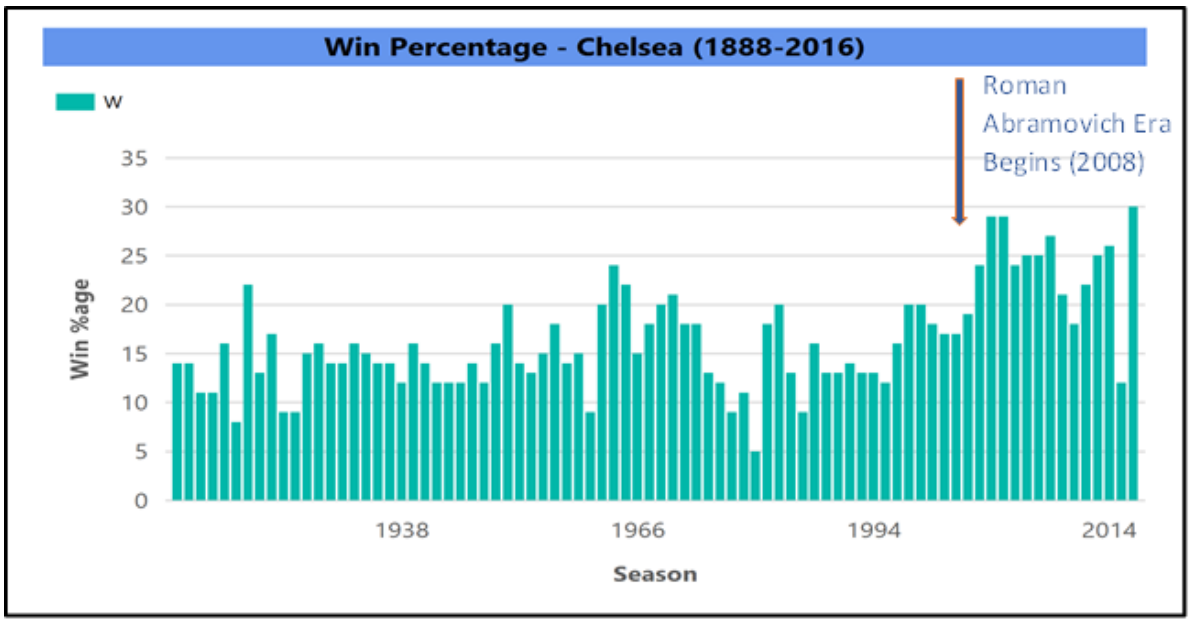

Figure 2b: The Figure shows the Chelsea Win %age (i.e. Games won per Season / Games played per Season) for each season from 1888 – 2016 using SSRS.

Similarly, Figure 2c. portrays the dramatic increase in Win %age for Manchester City F.C. ("Manchester City F.C.", 2017) when Abu-Dhabi based Abu Dhabi United Group Investment and Development Limited completed a takeover of Manchester City in 2008.

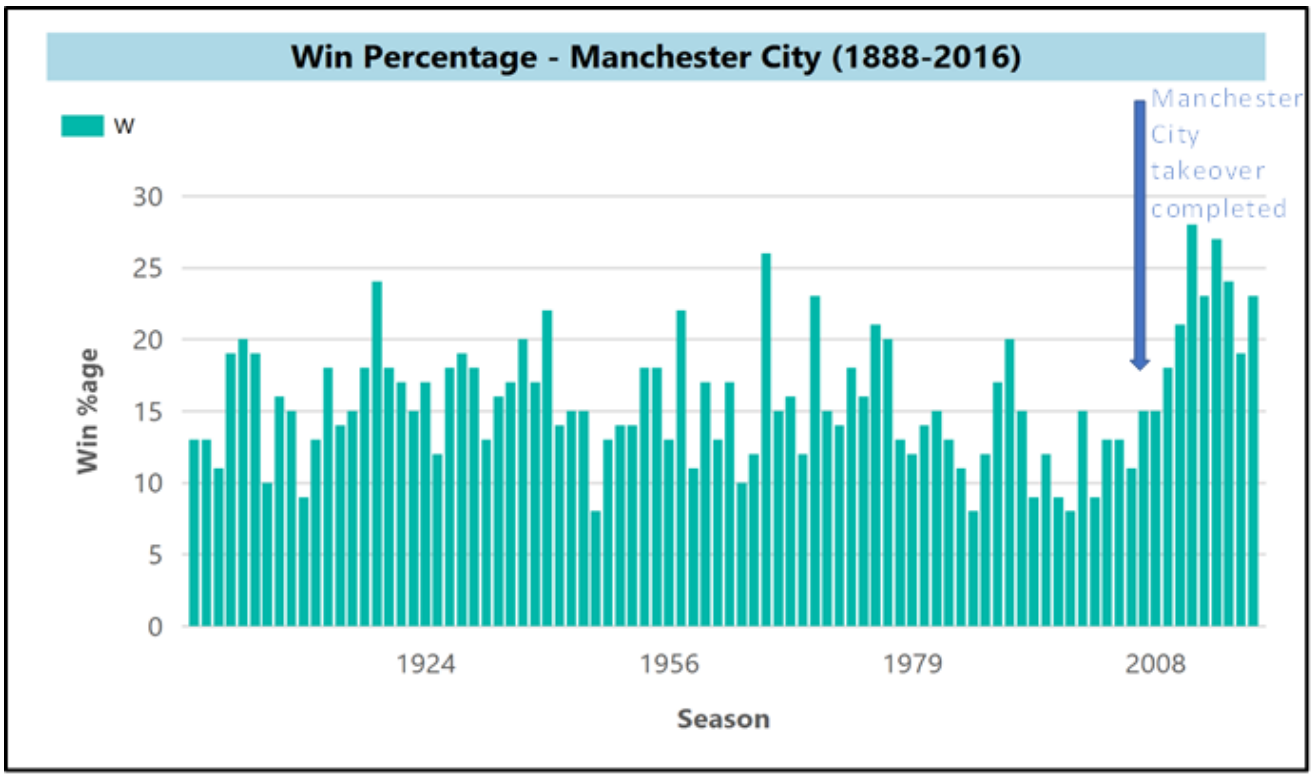

Figure 2c: The Figure shows the Manchester City Win %age (i.e. Games won per Season / Games played per Season) for each season from 1888 – 2016 using SSRS.

- **How significant is the home field advantage (Playing games at home ground) in the English Premier League now?**

  In the early days of English football, [Figure 3.] about 60 percent of games were won by the home team. Now, the home team wins only about 40 percent of games. The pattern effectively shows the gradual decline in home advantage for teams.

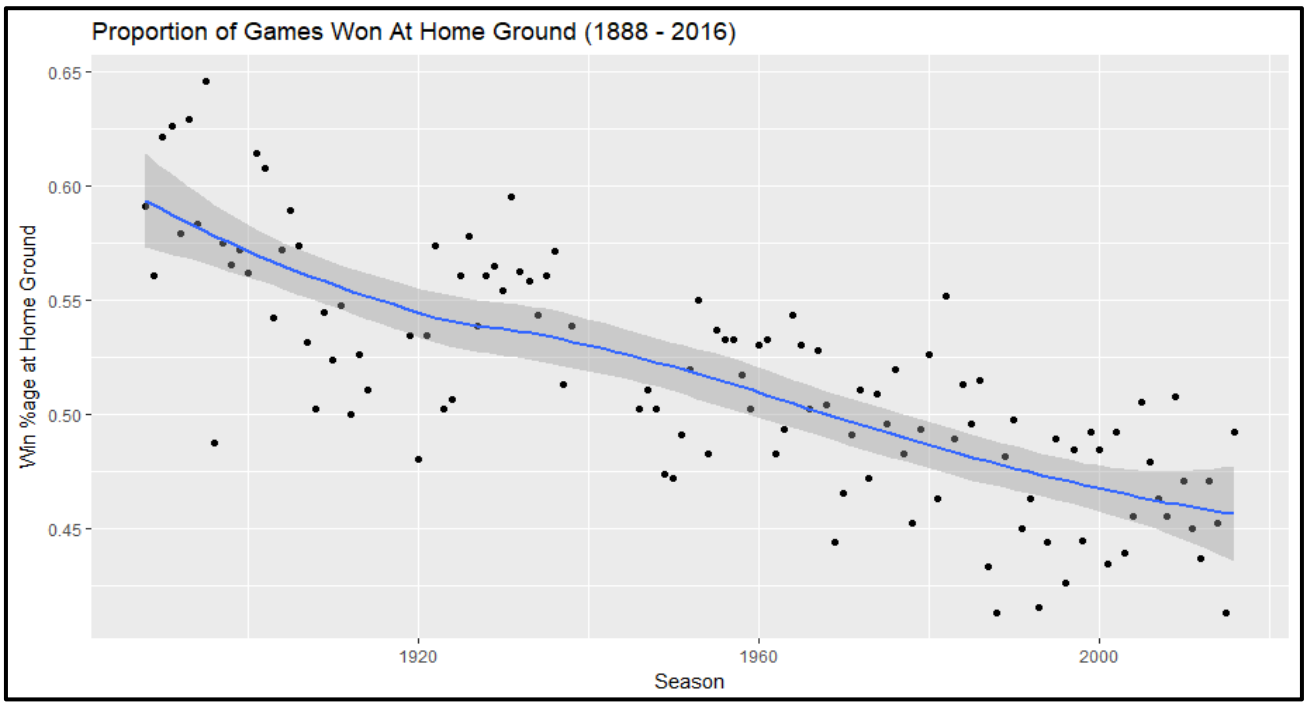

Figure 3: The proportion of games won at home ground by teams in the English league is shown with a loess smoother.

Discussion:

Nearly 200,000 English soccer games have been played in the top four leagues since 1888, the days of Jack the Ripper and Queen Victoria (Roeder, 2014), with 1-0 the most common final score. This result has occurred in 16 percent of the total games. In 85,694 games, close to half the total games, at least one of the teams forgot to score at all. More than quarter of all games, 47,412, in total have ended in a draw. Percentage of draws [Fig. 1] have increased manifold from 1888, when the game was centered around attacking football (Roeder, 2014), to 2016 as gradual realization of importance to defend gained prominence. Since football is a difficult beast to predict a few changes - like the arrival of a billionaire investor or the departure of a talented coach or manager - can change very quickly the fate of a team as illustrated by Fig. 2a., 2b. and 2c. The departure of Sir Alex Ferguson, Manchester United F.C.'s Manager from 1986 – 2013, in 2013 has left the club devoid of winning any major trophy till now corroborating his importance to team & club ("Manchester United F.C.", 2017). Similarly, from Fig. 2b and 2c. we can see the difference that a sudden cash flow makes to a - until then - rather average team. The Win %ages and trophy cabinet for both Chelsea F.C. [8] and Manchester City F.C. [9] depict a different picture after their respective takeovers by cash-rich organizations.

Is Home-field advantage in English football is disappearing? has been the ever-present question (Roeder & Curley, 2014) amongst soccer pundits for many years. The answer to which is yes [Figure 3.]. But more important is to find the reason for this dramatic shift?". Many reasons have been offered for shifts in advantage (Roeder & Curley, 2014), but one convincing

explanation is the reduced bias towards the home team by the match referee gradually over the years due to increase in money, exposure, professionalization, organization, scrutiny, oversight, monitoring and evaluation of the league.

Conclusion:

The above is preliminary analysis of pertinent issues ingulfing soccer. The approach and the data manipulation & visualization techniques used here can be reproduced on any other domestic soccer leagues or international soccer teams to derive meaningful results. To provide a holistic view of evolving playing style, the increase in Draw %age & decrease in Home advantage together with the variation in the average goals scored by both the visiting team and home team over the seasons needs to be analysed. This would firstly help us in depicting whether the increase in Draw %age is due to decreased scoring of goals and thus can be used to corroborate the fact that indeed the style of play has evolved over the years. Secondly, the scoring trends (both average away goals and home goals scored by a team over seasons) can also help shed some light on the decreasing home advantage. Whether decreasing home advantage is due to increased number of goals scored by visiting teams during recent seasons would be a good verification of our reduced bias towards visiting team. The shift in the dynamics of a soccer club upon arrival of a rich owner which increases the spending capacity of a club to buy talented players, support staff &/or talented team manager is corroborated by my own dedicated viewing of the game for past many years and by soccer history which is rife with numerous examples of the same.

References:

1. Magowan, A. (2011, November 23). Football Tactics: Can key statistics help prove a player's value? Retrieved January 03, 2018, from http://www.bbc.co.uk/blogs/thefootballtacticsblog/2011/11/how_statistics_shaped_a_hollyw.html
2. Dawson, P., Dobson, S., Goddard, J., & Wilson, J. (2007). Are football referees really biased and inconsistent? evidence on the incidence of disciplinary sanction in the English Premier League. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*(1), 231-250.
3. Elder, R. (2017, January 19). The English Premier League's viewership drop spells danger for the future of sports on TV. Retrieved January 03, 2018, from http://www.businessinsider.com/heres-what-the-english-premier-league-viewership-drop-means-for-the-future-of-live-sports-2017-1.
4. McCaskill, S. (2016, November 24). How Analytics And Digital Are Powering Man City's Quest For Success. Retrieved January 03, 2018, from http://www.silicon.co.uk/data-storage/bigdata/man-city-digital-tech-football-201114?inf_by=5a4d2251681db82c2b8b4770.
5. Roeder, O. (2014, October 16). In 126 Years, English Football Has Seen 13,475 Nil-Nil Draws. Retrieved January 03, 2018, from https://fivethirtyeight.com/features/in-126-years-english-football-has-seen-13475-nil-nil-draws/
6. Roeder, O., & Curley, J. (2014, October 08). Home-Field Advantage Doesn't Mean What It Used To In English Football. Retrieved January 03, 2018, from https://fivethirtyeight.com/features/home-field-advantage-english-premier-league/
7. Manchester United F.C. (2017, December 16). In Wikipedia, The Free Encyclopedia. Retrieved 19:46, December 20, 2017, from https://en.wikipedia.org/w/index.php?title=Manchester_United_F.C.&oldid=815676266.
8. Chelsea F.C. (2017, December 20). In Wikipedia, The Free Encyclopedia. Retrieved 20:10, December 20, 2017, from https://en.wikipedia.org/w/index.php?title=Chelsea_F.C.&oldid=816339998
9. Manchester City F.C. (2017, December 20). In Wikipedia, The Free Encyclopedia. Retrieved 20:14, December 20, 2017, from https://en.wikipedia.org/w/index.php?title=Manchester_City_F.C.&oldid=816319916