Tahmeed Tureen

Title: A Time Series Visualization of “Big” English Premier League Teams

SI 330: Data Manipulation | University of Michigan, Ann Arbor

Background

The English Premier League (EPL) or The Premier League is one of the world’s most popular soccer leagues. In the past, the EPL was primarily dominated by Liverpool F.C. This team was consistently competing for the title with the occasional championship run from Manchester United and Arsenal F.C. However, at the turn of the 21st century, “smaller” teams such as Chelsea F.C. and Manchester City started to have more of an impact on the league. This sudden evolution of such teams have brought more excitement and unpredictability to the EPL. For example, Chelsea F.C. won the EPL title in the 2016-2017 season, however, they finished tenth overall in the 2015-2016 season. This year, Manchester City is 11 points (~3.5 wins) clear at the top of the EPL table above historic giants, Manchester United. This type of unpredictability and competitive atmosphere between the teams make the English Premier League stand apart from the likes of the Spanish and German leagues.

Motivation

The purpose of this project is to explore the winning trends for seven of the “biggest” EPL teams from the start of the 21st century until last year’s season. The seven teams analyzed in this study are Chelsea F.C., Manchester United, Manchester City, Arsenal F.C., Tottenham Hotspurs, Liverpool F.C., and Everton F.C. These teams were chosen because they are all recognized as top-flight teams in the EPL and have historic rivalries amongst each other. As a soccer fan, I want to answer the question of whether or not there is a clear dominant team in the EPL in the modern era. In the past, the answer was either Liverpool or Manchester United. However, nowadays, the answer is not so clear. To answer this question, I analyzed the wins of each of these teams from the 2000-2001 season to the 2016-2017 season using a time series visualization of the trends.

Data Sources

Two separate data sources were used for this project. The first data source is from an online soccer repository that contains data on every single English soccer league including the EPL as well as lower tier leagues. The source can be viewed here. 17 CSV files and a txt readme document were downloaded from this repository. Each CSV file consists of match results for every single game (approximately 380 entries) for a specific season. The most important variables from this data were the home team (HomeTeam), away team (AwayTeam), full-time home team goals (FTHG), and full-time away team goals (FTAG). The second data source is from a different online repository that consists of large amounts of data regarding almost every soccer league in the world including the World Cup. The source can be viewed here, and an example of the EPL source can be viewed here. Although, the website is open for everyone's usage, the data for the EPL was available only as HTML files. 17 HTML files were downloaded from this repository which were manually converted to txt files. Each file consisted of the league table for each season of the EPL. Since the files are in HTML format, there is no number of records for the data. The important variable from this dataset is the name of the club who were champions for a specific season. This can be identified by observing the key word "Champions" at the side of the league table. It is important to note that none of the numbers in the txt files are denoted with variable names, so it is not clear to a non-EPL fan what the numbers mean for each team.

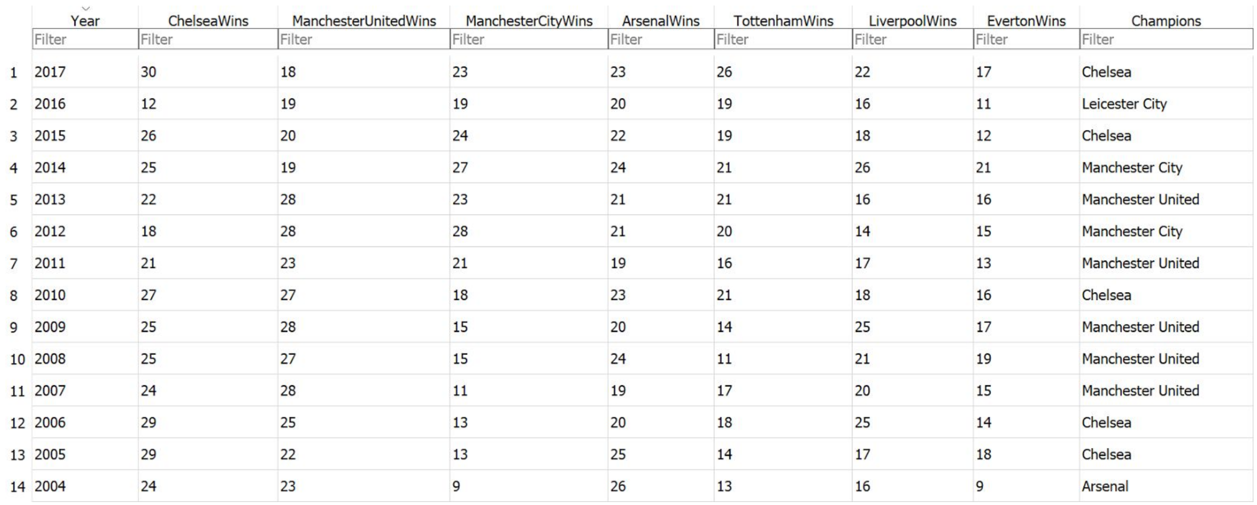
Data Manipulation Methods

For the purposes of data manipulation, Python3 along with the CSV module, Regular Expressions, and SQLite were used. First, all 17 CSV files were filtered and manipulated to output seven new CSV files. Consider Chelsea for example, the CSV files were read row by row and examined whether or not the string “Chelsea” was in the HomeTeam or AwayTeam attribute. If “Chelsea” was in HomeTeam, then it checked if the FTHG was greater than FTAG. If that were the case, Chelsea’s win counter was incremented by one. This process was repeated for each season (file) and each team until seven new CSV files were created. The newly outputted CSV files can be found in the “Manipulated_Data” folder. The HTML files were manually changed to txt files using a text editor. The new txt files were then read into Python and converted to the string data structure. Regular Expression methods were then used to strip the names of the champions for each season (each txt file) and outputted to 17 CSV files. These files can be found in the “Manipulated_Data” folder. Finally, the petl module and SQL queries were used to combine the new datasets into a clean database. First, all of the CSV files were converted to database tables. Then one by one, each of the tables were combined with the key variable “Year” until a final database table was created. Each row of the database table represents the EPL season and each column represents a specific team’s win count for that season. There were some small challenges with the Regular Expression parsing because the txt files contained champions for every single league in England. Fortunately, the Premier League table was at the very top, so I was able to limit the matching to just one group. Another issue with matching was that the EPL team names are either one word or two words. To resolve this issue, I first explored each txt file and noticed that only two champions had one word names.

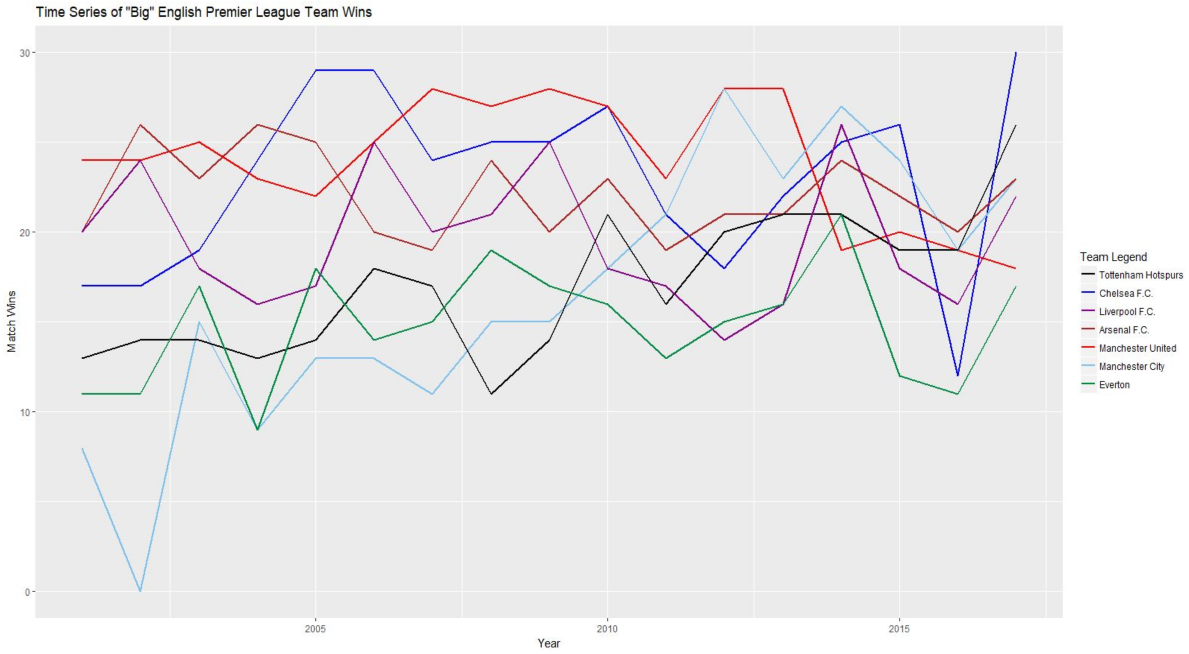
Therefore, in my source code I added if-else blocks that treated these two teams differently since their matched output behaved differently. There were no missing data in either data source.

Analysis & Visualization

The EPLSeasons table from the epl_teamWins database and the statistical software R was used to create a time-series visualization of each team's winning trends for the past 17 years to analyze the new manipulated data. Part of the data table and the visualization are shown below.



	Year	ChelseaWins	ManchesterUnitedWins	ManchesterCityWins	ArsenalWins	TottenhamWins	LiverpoolWins	EvertonWins	Champions
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	2017	30	18	23	23	26	22	17	Chelsea
2	2016	12	19	19	20	19	16	11	Leicester City
3	2015	26	20	24	22	19	18	12	Chelsea
4	2014	25	19	27	24	21	26	21	Manchester City
5	2013	22	28	23	21	21	16	16	Manchester United
6	2012	18	28	28	21	20	14	15	Manchester City
7	2011	21	23	21	19	16	17	13	Manchester United
8	2010	27	27	18	23	21	18	16	Chelsea
9	2009	25	28	15	20	14	25	17	Manchester United
10	2008	25	27	15	24	11	21	19	Manchester United
11	2007	24	28	11	19	17	20	15	Manchester United
12	2006	29	25	13	20	18	25	14	Chelsea
13	2005	29	22	13	25	14	17	18	Chelsea
14	2004	24	23	9	26	13	16	9	Arsenal



This is a very interesting project because the winning trends show that almost all of the team's performances are inconsistent throughout the years and there is no longer a clear dominant side in the EPL like in the past. The visualization shows that both Manchester United and Liverpool have not been dominating the league like they have in the past. As a matter of fact, Manchester United seems to be declining in terms of total wins each season in recent years. Something that I personally found interesting was that Manchester City had zero wins in the 2001-2002 season. This means that they were relegated from the EPL the previous season. Since then, Manchester City have won two EPL titles and are on their way to win this season's title. This is an amazing comeback story as they are the only "big" team to have been relegated in the 21st century. The visualization also shows that Tottenham Hotspurs are emerging to be a real competitor for the title and that Chelsea F.C. has the record for most wins (30) in a single season in the 2016-2017 season. It is also important to note that the results are very interesting when both the table and the time-series is observed together. In almost every season, the team that had the most wins won the EPL title. However, there are instances where this is not true. For example, in the 2015-2016, Arsenal F.C. had the most wins amongst the "big" teams. However, Leicester City were crowned the champions for that season. This is a very interesting result in the analysis because Leicester City is not recognized as a "big" team nor are they favorites to compete for the title. This result solidifies the unpredictable and competitive nature of the EPL.

Links

Source #1: http://www.football-data.co.uk/englandm.php

Source #2: http://www.rsssf.com/