Time Series Visualization of “Big” English Premier League Teams

Tahmeed Tureen

SI 330: Data Manipulation, Fall 2017
University of Michigan, Ann Arbor

Motivation & Data Sources

Motivation: English soccer (football) has always been dominated by a select few teams.

Statistically speaking, it's arguable that Manchester United is the greatest soccer team in English history.

However, at the start of the 21st century, smaller teams started to make more of an impact in the league. This project attempts to visually explore the dominance of the old "big" teams and the new "big" teams for each year starting from the 2000-2001 season by analyzing total wins per season.

Data Sources:

- 17 CSV files (~381 entries each) from http://www.football-data.co.uk/englandm.php , an online repository containing game results for the English Premier League.
- 17 HTML files (converted to .txt files) from http://www.rsssf.com, an online repository containing English Premier League table results as well as other leagues.

Data Manipulation

- CSV Manipulation
  - The 17 CSV files were read using DictReader
  - Each row for each CSV was run through filtration using Python 3
  - Once the desired data structures were created, 7 new CSV files for seven separate teams were created
- Regular Expression
  - Regular expression matching was then used to strip data from the set of .txt files
  - The champions from each season was the desired data
  - One new CSV file was created which contained the champions for each season
- Database Creation
  - Initialized new database using SQLite via Python
  - SQL Queries were then used to combine all of the CSV files created in previous steps
  - A final database table was created via and was used for visualization via RStudio

Time Series Visualization of Wins for “Big” Premier League Teams