**Syracuse University**

**MS Applied Data Science Project Portfolio Milestone**

**2019-06-02**

Student Information

Name: Martin Alonso

SUID: 939538821

Email: malons01@syr.edu

Portfolio Table of Contents

SyracuseUniversityProjects

* IST659 Project
  + ist659project
    - app.R
  + .RData
  + .RHistory
  + Alonso.Rmd
  + Alonso.docx
  + Alonso\_Project 1 Deliverable
  + Alonso\_Project 2 Deliverable
  + IST 659 Data Admin Concepts and Database Management.Rproj
  + IST659 Project (1).png
  + IST659 Project (2).png
  + IST659 Project (3).png
  + IST659 Project (4).png
  + IST659 Project Description.pdf
  + IST659 Project Normalization Tables.xlsx
  + IST659 Project.png
  + Project Data Exploration – 1.sql
  + Project Data Exploration – 2.sql
  + Project Data Exploration – 3.sql
  + Project Data Exploration – 4.sql
  + Project Data Exploration.sql
  + Project Data Insert.sql
  + Project Procedures.sql
  + Project Table Creation.sql
  + scripts.R
* IST687 Project
  + .RData
  + .Rhistory
  + IST687 Data.xlsx
  + IST687 Project Code.R
  + IST687 Project Write Up.docx
  + IST687 Project.Rproj
  + Superimposed histograms example.R
* IST707 Project
  + .RData
  + .Rhistory
  + Alonso\_Chen\_March\_Madness.pptx
  + IST707 Markdown.Rmd
  + IST707 Master dataset.R
  + IST707 Project.Rproj
  + IST707\_Markdown.docx
  + Martin\_Alonso\_Project Report.docx
  + WCities.csv
  + WGameCities.csv
  + WNCAATourneyCompactResults.csv
  + WNCAATourneySeeds.csv
  + WNCAATourneySlots.csv
  + WRegularSeasonCompactResults.csv
  + WSeasons.csv
  + WTeamSpellings.csv
  + WTeams.csv
* IST718 Project
  + .ipynb\_checkpoints
    - IST718 Project ABDK (2)-checkpoint.ipynb
    - Tag MBA-checkpoint.ipynb
    - Untitled-checkpoint.ipynb
  + Alonso\_Brown\_Davis\_CourseProject.ipynb
  + Alonso\_Brown\_Davis\_CourseProject.pdf
  + Alonso\_Week5ProjectCheckIn.docx
  + Alonso\_Week7ProjectCheckIn.pdf
  + Project Checkpoint 1.pdf
  + Tag MBA.ipynb
* IST736 Project
  + Final\_Project\_Alonso\_Martin.pdf
  + Harper\_Tweets\_Project.py
  + Harper\_Tweets\_Simp.csv
  + harper\_tweet\_locations.txt
  + harper\_tweet\_locations\_clean\_backup\_file.txt

Program Learning Goals

There are four goals in the Applied Data Science program: Data collection, data analysis, strategy and decision-making, and Implementation. These goals are achieved throughout the program by working on projects and assignments in the different courses offered. To show I have achieved these goals, I have selected five projects presented in five courses. These projects demonstrate at first glance that I am capable of planning and setting up a database architecture, I can collect data and analyze it, I cab develop models to extrapolate from my data analysis, and I can use the results to provide business analysis and suggestions.

The first goal, data collection, is featured on every project, but, it is more prominent on both the project for IST718 Big Data Analytics and IST736 Text Mining. For the IST718 project, along with Rashad Davis and LaRue Brown, we tried to build a movie recommendation system. For this, we used data that was featured on [Movie Lens](https://grouplens.org/datasets/movielens/20m/), a website that contains a database of 20 million movie reviews, along with tags, genres, and additional data for over 27 thousand movies. In total, the Movie Lens files contained data saved across six comma-separated value (CSV) and text files, in easily digestible format that we only needed to download and plug into our project notebook.

However, these data were not enough. In order to further enrich our collected data, we scraped movie duration, director, cast members, and PG-ratings from [The Movie Database](https://www.themoviedb.org/?language=en-US) website using Python’s Beautiful Soup package. These data were returned in a JavaScript Object Notation (JSON) format, needing to be cleaned and tabulated in a more user-friendly format before they could be added and inserted into our project. These collected data allowed us to compile a large data set that allowed us to use different techniques to model user-movie preference and design a model capable of suggesting similar movies to a specific user, provided a list of five movies.

For the IST736 project, the data collection process did not involve clean, formatted CSV files. The objective of the project was to perform sentiment analysis on tweets that referred to current Philadelphia Phillies right fielder Bryce Harper. In order to gather tweet data, I first obtained access to the Twitter developer Application Programming Interface (API), scraping every tweet between November 1st, 2018 and March 1st, 2019. However, Twitter’s API is limited to scraping tweets from the past seven days, severely limiting my available data.

Searching through Python’s package libraries I found a library called [*twitterscraper*](https://github.com/kennethreitz/twitter-scraper). This package allows users to scrape tweets through command line functions, taking as inputs the start and end dates for which the user wishes to scrape the data, a key word (or words) that a user wishes the tweets to contain, an optional cap size on the number of tweets the user wishes to return, and the format the user wants the data to be returned into. Using the following parameters, I scraped data between November 2018 and March 2019 that contained any reference to Bryce Harper, returning over 200 thousand tweets in JSON format.

Though I successfully collected the required data, the tweets recollected were missing one key aspect for my project to be completed. I needed location data for where each tweet was sent from. Luckily, the data collected from the *twitterscraper* package contained each user’s Twitter identification number. Without looking into each user’s personal information, in order to keep with privacy regulations and keep the data anonymous, I collected each user’s city, state, and country of residence by developing a script that, through the Twitter API, returned the requested information provided I supplied each user’s ID number.

The final challenge for this collected data was that each user is capable of altering their location information, thus the data was in an unstandardized format. Users could optionally input where they were located and how their data was presented. Users in Philadelphia, Pennsylvania, for example, presented their location data the following ways:

* Philadelphia, PA
* Philadelphia, PA, USA
* Phillly, Penn, USA
* Philadelpha, Penn.
* Carsondelphia, Wentzylvania
* West Philly
* East Philly

This presented a challenge as I would need to use regular expressions to further clean the scraped data and homogenize it in a way that was more useful for my final project.

Though the data collection project of the first project was relatively easy, the collection for the second project was more challenging as it not only required scraping the data but also processing and collecting a larger subset of data that, even using Python’s *multiprocessing* package and IBM Watson’s online server, the data collection and processing took close to 96 hours to complete. Though this was taxing process, it helped me to identify difficulties in my process that I could later optimize, synthesize, and improve after the final project was submitted.

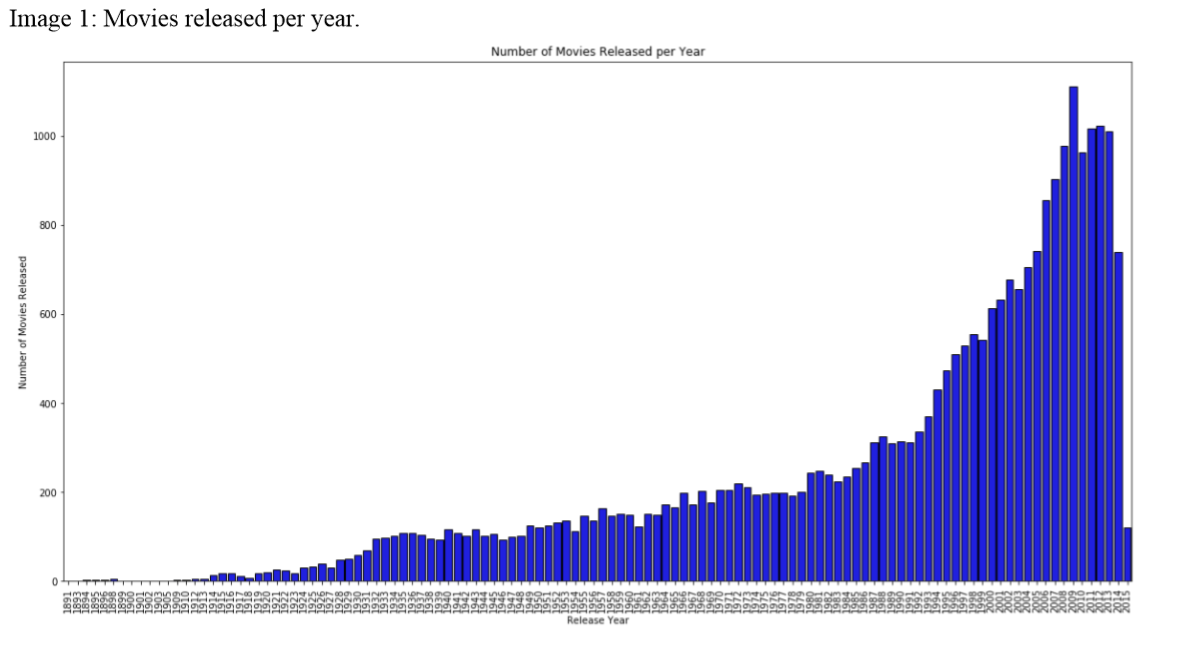
The second goal of the project involves data analysis. For this section, all of my projects, sans the project for IST659 Data Management and Database Administration Concepts, present evidence of detailed analysis that leads to both business insights and planning for building later models. Since we’ve already introduced the projects for IST718 and IST736, I’ll continue with these two before moving on to the projects for IST687 Introduction to Data Science and IST707 Data Analytics.

Starting with IST718, our first order of business was to check what data was contained in each of the CSV files downloaded from Movie Lens. The data contained six files with the following information:

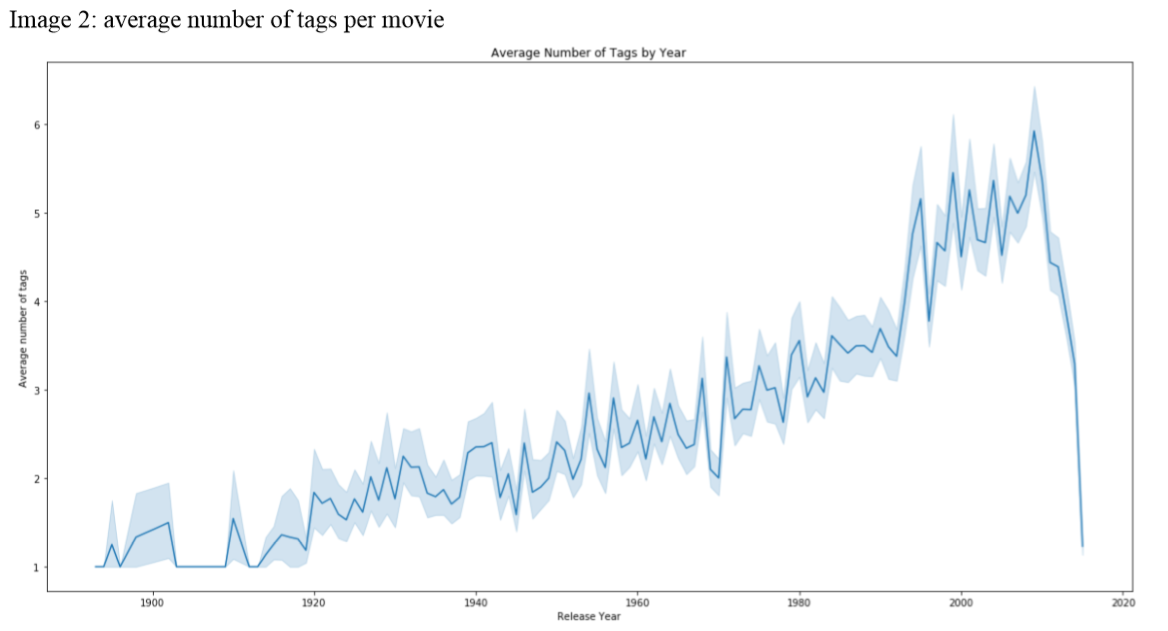
* Movies.csv: Movie ID, Title and Release Year, Genres
* Ratings.csv: User ID, Movie ID, Rating Score, Timestamp
* Tags.csv: User ID, Movie ID, Tag, Timestamp
* Genome-tags.csv: Tag ID, Tag
* Genome-scores.csv: Movie ID, Tag ID, Relevance Score
* Links.csv: Movie ID, IMDB ID, TMDB ID

For our project, we focused on working with the movies, ratings, tags, and links files. The first three gave us all the detailed information about the movie: it’s title, release year, genres, scores, tags, and when users rated and tagged them. The last file was necessary so we could scrape the movies in the file from The Movie Database, collecting additional details about the movies.

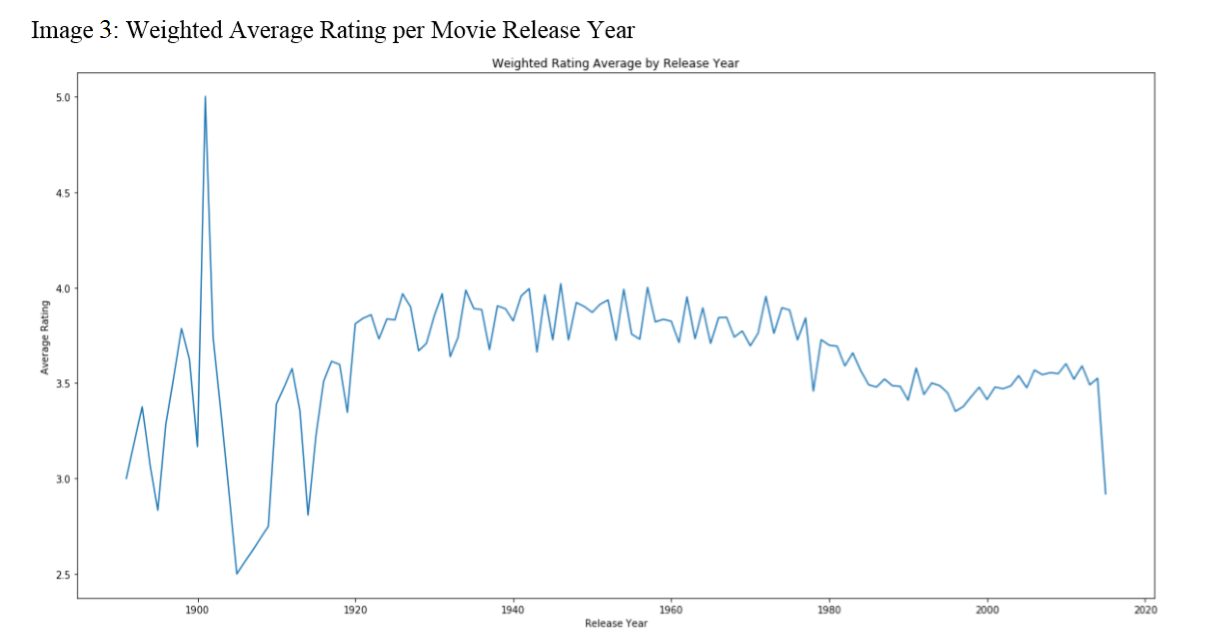
After first exploring the contents of each database, we started building a couple of graphs to better understand our data. We plotted a bar plot to view the number of moves released every year (image 1), while also reviewing the average number of tags each movie receives (image 2). But these weren’t the only point of interest we wanted to explore within the data.



It was also important to understand how movies were being rated, and whether rating varied across time, as we thought that older movies, especially those tagged as classical, would merit a higher rating. We weighed the average rating by number of movies released but we found no evidence of this (image 3).



Having explored the data, we noticed that how movies are rated or when a movie was released does not matter for a recommendation system. The important drivers are both the genre and number of similar tags that users give movies. By grouping genres, and then tags, we find that our data has a higher similarity score than relying on any other piece of information.



For my IST736 project, the analysis was a bit different. The objective of this project was to determine how different fanbases reacted to rumors that linked their teams to Bryce Harper. During the 2018-2019 MLB offseason, at least 10 teams were linked at some point to Harper. After obtaining the location data for all users and cleaning the data, I tabulated the number of tweets sent from each city. Table 1 shows the top 10 cities ordered by number of tweets sent during those four months.

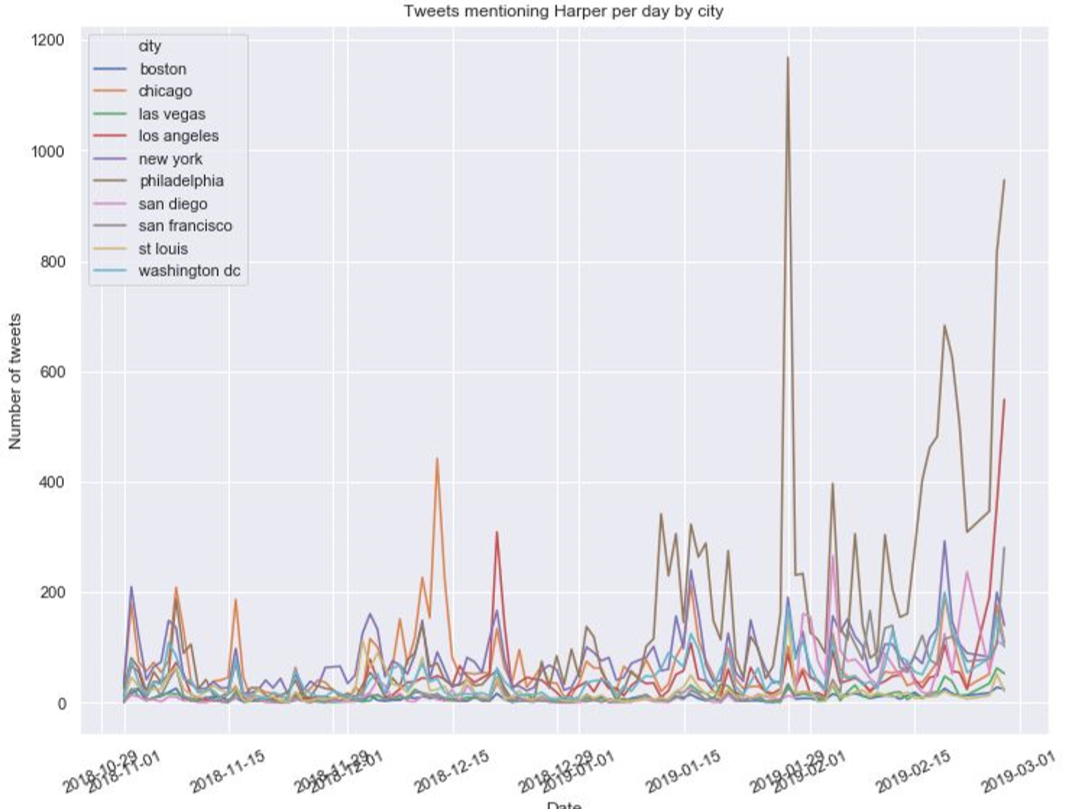
Table 1: Top 10 cities, post-cleaning

|  |  |  |
| --- | --- | --- |
| Rank | City | Number of tweets |
| 1 | Philadelphia | 56,762 |
| 2 | New York | 10,659 |
| 3 | Chicago | 8,330 |
| 4 | Washington D.C. | 6,662 |
| 5 | Lon Angeles | 6,114 |
| 6 | San Francisco | 3,696 |
| 7 | San Diego | 3,260 |
| 8 | St. Louis | 2,522 |
| 9 | Las Vegas | 1,695 |
| 10 | Boston | 1,231 |

Harper eventually signed a 13-year contract with the Philadelphia Phillies, but for most of the offseason, he was linked with the Chicago Cubs and Los Angeles Dodgers. This table, when I first looked at it, made me think that Philadelphia fans were enthusiastic when Harper signed with the Phillies.

But, in order to better understand the data, I wanted to explore how many tweets were being sent on a daily basis. Because I had each tweet’s timestamp, I was able to plot a time series graph showing how many tweets were being sent out from each city (image 4). For the majority of the offseason, most of the tweets were coming out of Chicago. The Philadelphia twitter sphere didn’t really start echoing their feelings towards Harper until before mid-January, sending out a large majority of tweets regarding Harper and the Philadelphia Phillies until Harper was signed.

Image 4: Tweets per day per city



Naturally, looking at these trend lines, I assumed that the majority of tweets coming out from Philadelphia would be positive. However, I was surprised to learn that that wasn’t the case. The tweets coming out from Philadelphia were mostly negative. Doing some sentiment analysis, I found that the majority of tweets were actually negative (image 5) and, to further incriminate Philadelphia, the majority of negative sentiment tweets came out of Philadelphia the day Bryce Harper signed (image 6).

Surprised by these results, I would then proceed to build a model that tried to determine where a tweet was sent from based on the sentiment of the tweet.

Image 5: Number of tweets by sentiment for stemmed and lemmatized words.

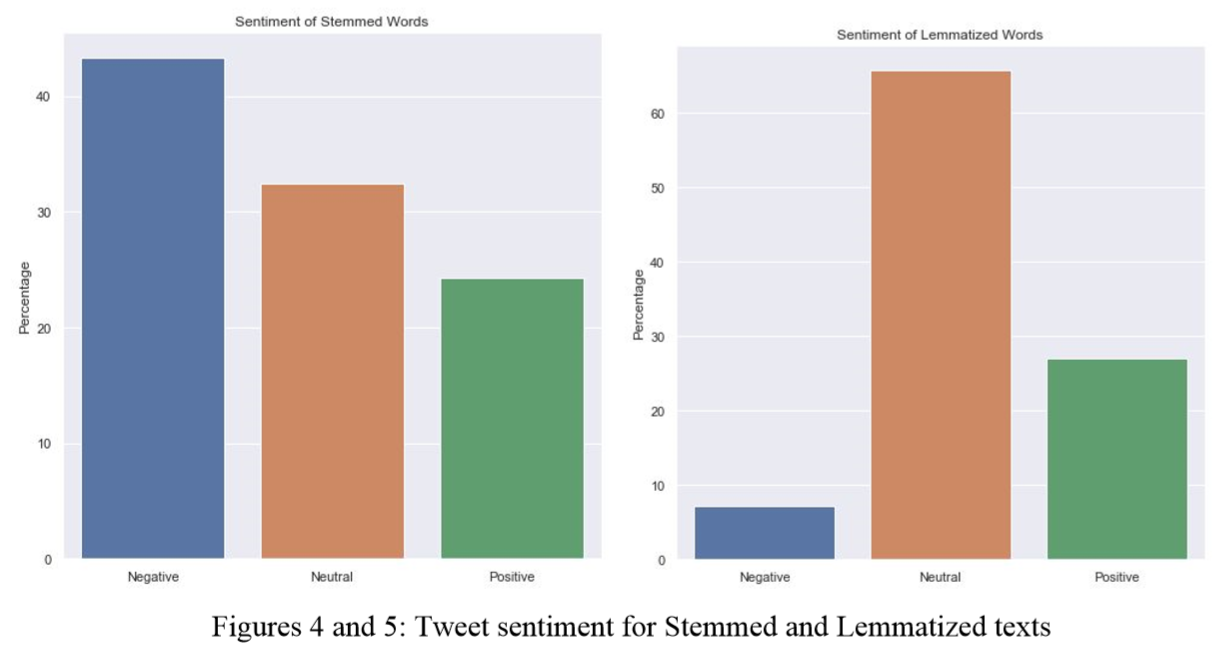
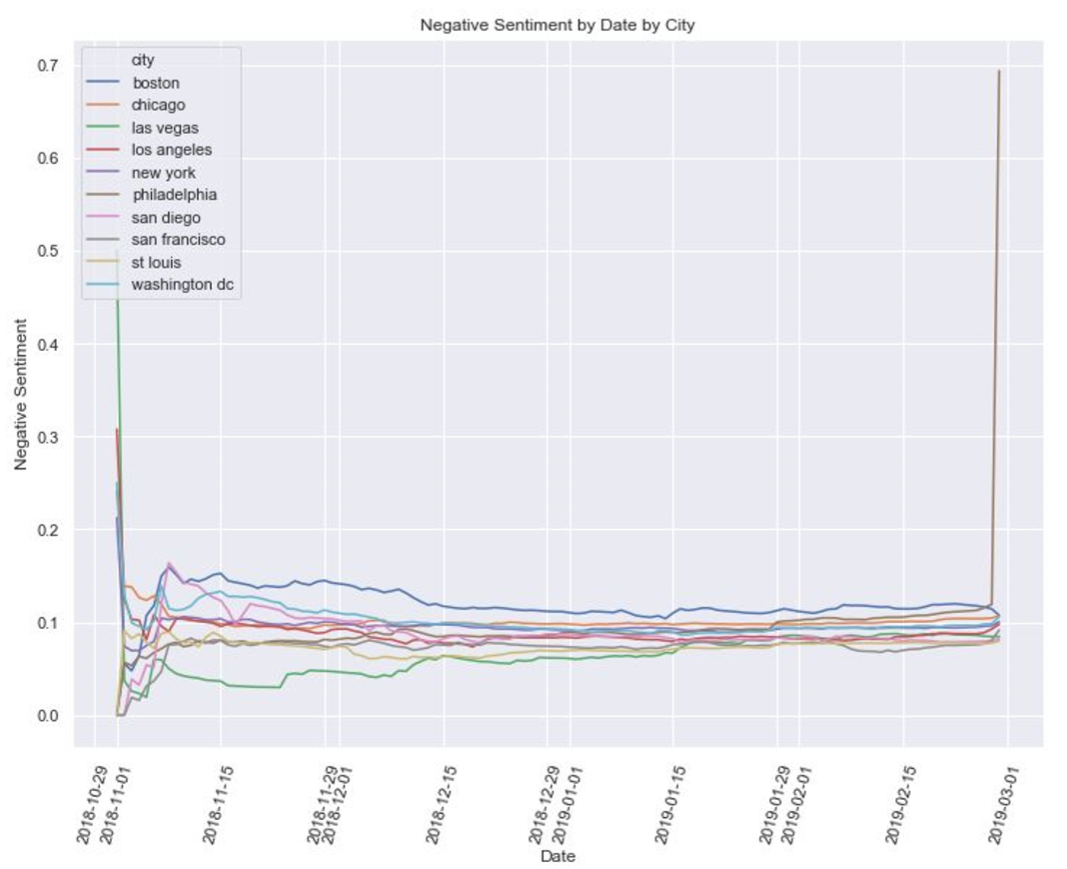


Image 6: Percentage of Negative Sentiment Tweets

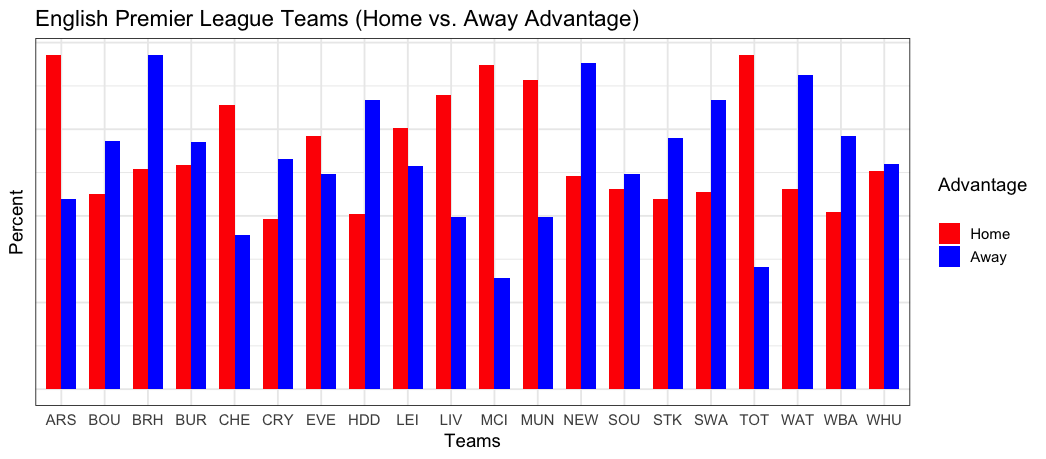


For my IST687 Introduction to Data Science project, the goal was to analyze English Premier League football matches. My group, consisting of Amani Alawneh, Michael Cerutti, Susannah Cleland, and myself, obtained the data from [www.football-data.co.uk](http://www.football-data.co.uk). The objective of this project was to, using several statistics from every match played, determine whether we could predict the winner of a match.

Analyzing the data for this project was not a challenge, as what we wanted to compare was the different statistics for home and away teams, as well as how they compare to winning and losing teams. We analyzed number of fouls committed by each team, as well as number of red cards and yellow cards received. We also analyzed the number of shots taken, shots taken on goal, corner kicks, and goals scored to determine average goals per shot, and goals at half-time, to determine how often a team wins and by what margin.

But this is not equal among all teams. As we can see in image 7, there are teams that have a higher home advantage, while there are teams that have a higher away advantage, though the latter are a rarer breed.

Image 7: Home and away advantages



By looking at these numbers, we found that home teams do have an advantage of sorts, receiving less cards and being called less for fouls committed. At the same time, home teams tend to have a higher advantage when it comes to shots taken (image 8). This means that either the home team tends to be more aggressive offensively, taking more shots, is more aggressive defensively, allowing less shots and, though could possibly commit more fouls, being called less on them. Or, it could be both, having home teams play more aggressive football on both sides of the pitch.

Image 8: Home advantage on shots taken.

