Joshua Douberly, Sean Lee, Nathan Williams

OA3801 Computational Methods II

September 2023

Final Project Executive Summary

An Analysis of Professional Sports Success in American Cities

1. **Introduction**

Americans live and die by the success of their sports teams; it is highly likely that most people know their city’s star quarterback better than their congressional representative. Fans fuel the big five multi-billion dollar sports franchises: Major League Baseball (MLB), National Football League (NFL), National Basketball Association (NBA), National Hockey League (NHL), and Major League Soccer (MLS). While every fan has their preference of sport, fans all across North America will claim their cities have the most storied legacies of sports success. While fans can bicker and compare recent championships, there are few objective ways to measure or visualize success. This study aims to provide insight into which cities or regions of the United States and Canada can legitimately lay claim to having the most professional sports success amongst the big five franchises: MLB, NFL, NBA, NHL, and MLS. We will use pure Win-Loss Percentage (W-L%), a direct ratio of wins to losses per year, number of championships won, and different ratios, such as championships per total number of seasons played. The study will provide results in a variety of ways, including Python-produced dashboards, evolving bar charts, pie charts, and animated heat maps.

1. **Data Collection and Cleaning Process**

Collecting data across all 5 major sports presents real challenges and inconsistencies. Although W-L% and number of championships won are easy enough to interpret across sports, there are intricacies to the data which make data processing much more time consuming than a simple read from CSV. Not only are the ages of the sports franchises vastly different, but franchises have the tendency to move cities throughout their history, which greatly affects our city-based analysis. Cities like New York have multiple teams for every sport, whereas cities like San Antonio may only have the San Antonio Spurs. While the NFL plays 17 games per season, as of 2021, the MLB plays 162 games per season, but this is assuming external events do not alter this; in recent history, events like COVID-19 and player strikes have shortened or all-together cut seasons short. Since the franchises present millions of dollars of opportunity for cities willing enough to invest, new franchises, like the Las Vegas Golden Knights, pop up out of the blue all the time (see Figure 1 below). All of these various data intricacies make city-based sports success analysis more difficult to conduct.

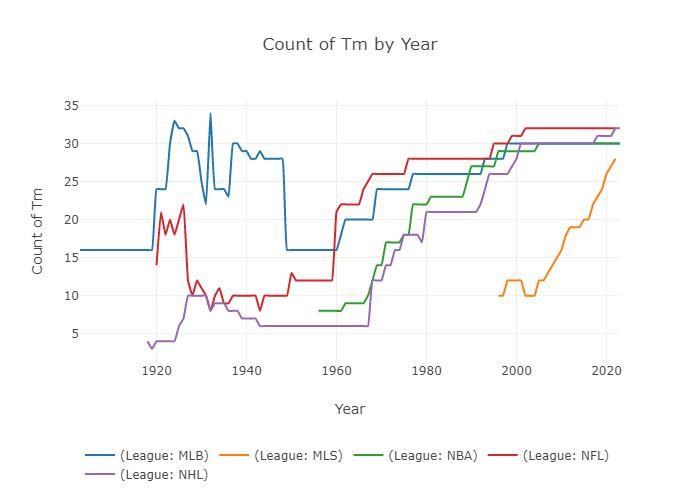


Figure 1: Number of Teams per Year per League - The widely fluctuating number of teams per league make year-by-year analysis more difficult.

The beginning states of each dataset are various forms of web-based data and the desired end state of our data is a DataFrame, called All\_Sports\_df, which was sorted by ascending Year, and stores as columns: Team Name (Tm), Win-Loss Percentage (W-L%), Latitude coordinates (Latitude), Longitude coordinates (Longitude), City name (City), State name (States), a binary Champion column (Champion), and a categorical League column (League). Since sports are highly researched statistics topics, each sport had options to export CSVs or XSLXs, but each website displayed results in vastly different ways. As an example of the different data storage methods, the NFL required reading in 160 different XSLX files and iterable concatenation, whereas the MLB required reading in 1 CSV and executing many pivot tables.

No matter the sport, data collection involved the same seven steps: locating a source, exporting as XSLX or CSV file(s) and reading in as a DataFrame, formatting the DataFrame to match the standard All\_Sports\_df template, account for data anomalies, adding in city location data through dictionary values, adding in championship through a dictionary, and exporting as a final XSLX. Data anomalies were mostly dealt with by manual addition because the databases would not account for necessary intricacies. For example, a franchise move would mean we must define new city data (Latitude, Longitude, City, States) after a certain year and a COVID-19 season would mean we must divide the total number of wins by less games than usual. After anomalies are accounted for, the formatting is essentially just various iterations of pivot tables and GroupBy functions in Pandas. Through our process, we learned in any real-world data that efficient web scraping techniques and Pandas GroupBy and pivot tables are critical to producing a usable DataFrame, and that data anomalies will always need to be accounted for because no database stores data exactly how one intends for it to be used.

1. **Visualizations and Dashboard**

Once the data was cleaned and compiled into a master spreadsheet that contained a row for each year of each team for each league, the data could be accessed by reading the csv in via pandas, and the visualization process could be started. To make the visualizations, we relied mostly on altair and then used streamlit to put all of the charts into a dashboard. We settled on using streamlit because it allowed us to create animated charts, which could be run by pressing a start button.

Our dashboard consists of three tabs: Primary Visualizations, Animated Graphs, and Data. The Data tab contains the master data frame, which allows the user to see exactly what data we are using and how we organized it. Next to the master data frame is a table that has summary statistics of the data to show exactly how much data we were using and the spread of the data between leagues. The table shows how many seasons of data we collected for each league, as well as the range of those seasons, the number of championships played and the number of teams in each league. Below that summary statistics table are three bar charts that visualize the summary statistics for easy, quick comprehension.

The Primary Visualizations tab contains all of the visualizations needed to answer the primary question of “which city has historically been the best at American sports?” At the top, we made three pie charts using altair which show the number of championships by city, the number of teams by city, and the championships won per total number of seasons played. We didn’t want to just claim that the city with the most championships is best because cities like New York have way more teams and seasons played then smaller cities like Atlanta. At the bottom of the Primary Visualizations tab, we broke down these pie charts into leagues to see which city was best at which sport. We also included a scatter chart which shows the average win/loss percentage of the teams in a city each year to see how the city performed over time.

The Animated Graphs tab has a pair of bar charts that run when the start button is clicked as well as an animated line chart. The top animated bar chart shows the number of championships won in each city, and the bottom bar chart shows the number of championships played each year. Both charts are color coded by the league. When the animation is run, the bars are incrementally added to for each championship played. This animated pair of charts shows a race between each city competing for the most championships. The animated line chart below shows almost the same thing as the bar charts above but instead in a single plot that also allows the user to see the winner for each year.

1. **Results**

Our data analysis focused on three main measures of success: average win/loss percentage, number of championships per city, and championships per season played in each city. Some of the results were surprising, others proved our predictions correct.

The three most successful regular season cities in the country in order (based on average win/loss percentage and seen more than 10 seasons played in) are San Antonio (0.604), Green Bay (0.575), and Oklahoma City (0.558). The success of these cities can easily be explained by them each only sponsoring one major professional team with consistently good teams.

The five most successful championship cities are New York City (54), Boston (40), and Chicago (30), Los Angeles (30), and Montreal (26). The top four are not surprising; each city has multiple teams that are established contenders and have hosted professional seasons numbering in the top five (New York Giants, New York Yankees, Boston Red Sox, New England Patriots, Chicago Bulls, Chicago Bears, Los Angeles Lakers, and Los Angeles Dodgers to name a few). However, Montreal is perhaps the most surprising result we found. Montreal is extremely competitive in hockey, and gathered a massive 26 Stanley Cups as of 2023. U.S. media probably cannot profit extensively from recounting the success of Montreal hockey teams, so they avoided our predictions.

The three most efficient professional sports cities in North America that have hosted more than 30 seasons are Montreal (0.157 Championships Per Season), Green Bay (0.127 Championships Per Season), and San Antonio (0.106 Championships Per Season). However, all of these cities can be considered small markets due to the aforementioned conditions. The most efficient big market city is Boston (0.092 Championships Per Season). The least efficient small market cities are Phoenix (0.008 Championships Per Season), Indianapolis (0.010), and New Orleans (0.010 Championships Per Season). The least efficient big market city is Atlanta (0.015 Championships Per Season).

Our most important result is not a statistic, but our combined data set(s) into coherent excel sheets. None of the five major sports leagues offered easily accessible, easily congestible information over the history of their organizations. The visualizations and statistics are helpful, but easily altered and created from the extensive work needed to create the wrangled data set.

1. **Conclusion**

Our project involved finding, cleaning, and combining over a century of data from five professional sports leagues, thousands of teams, and hundreds of cities. Hours of data wrangling resulted in a single, coherent dataframe of 9000 instances of year, team, location, win-loss percentage, championships won, and league. We turned and grouped the data into visualizations before moving our findings to a dashboard for easy accessibility and viewing. We discovered that big market cities that have hosted hundreds of seasons worth of sports have naturally won more championships, but are not necessarily the most efficient. We discovered small market teams with impressive win-loss percentages and even one with a massive number of championships won that nobody in the media really talks about. Overall, our project aimed and succeeded in finding the most successful sports cities, but inadvertently yielded what appears to be the first all-league data set of the past century and a half.