**DS2001: The Most Popular American Sports Team in the World (as of 2020)**

Grace Michael: [michael.g@northeastern.edu](mailto:michael.g@northeastern.edu)

Mark Hurley: [hurley.mar@northeastern.edu](mailto:hurley.mar@northeastern.edu)

**The Problem Statement:**

In the following report, a program has been built in order to determine which of the professional league sports teams in the United States is the most popular worldwide, based off of the most recent data. It is hypothesized that if the most popular U.S. sports team has the most play-off appearances, jersey sales, and twitter followers, then it can be inferred that they are the most popular team globally due to their successful seasons, high sales volume of merchandise, and large fan base on social media. The original hypothesis is that the most popular team is one of two of the most widely-recognized American sports teams: the New York Yankees and the Dallas Cowboys.

**The Data:**

Due to the outstanding popularity of American sports internationally, as well as the availability of U.S. sports data, the sports teams being analyzed in this report each belong to one of the major professional sports leagues in the U.S., which include the NBA, MLB, NFL, NHL and MLS. Among these sports leagues, the only foreign “competition” comes from Canada (whose top professional teams participate in American leagues such as the Toronto BlueJays and the Montreal Canadiens) and Europe (whose professional soccer teams tend to dominate American Major League Soccer, for example, Manchester United F.C.). Due to the fact that foreign sports teams are a minority next to the multitude of popular U.S. sports teams, in addition to the immense international popularity of leagues like the NBA and NFL, it can be inferred that the most popular sports team in the world can be determined from the following data.

Data for this project came from a variety of sources. This was due to the various categories we relied on for data. The data was organized based on the five major U.S. sports leagues: Major League Baseball (MLB), National Basketball Association (NBA), Major League Soccer (MLS), National Hockey League (NHL), and National Football League (NFL). This allowed us to compare statistics relative to teams within the same sport. Then, following the implementation of a scoring system within each respective league, we would be able to compare data across all leagues.

Most of our data, including Play-off Appearances and Twitter Followers, came directly from the leagues’ websites and social media. Since the leagues are responsible for all statistics in regard to their teams and have professional sports statisticians analyze the data and the games for correctness, it can be assumed that this information is not only accurate, but has been preserved ethically. This is further proven by the fact that many people, aside from those analyzing the data, are able to also view the games in real time. For leagues who have had their 2020 season finish, the 2020 season has been included in their league cluster. For those who have not, the most recent data has been used. In regard to social media, each team’s follower count was viewed and recorded on November 12th, 2020. Because the data was collected as a part of the project and the real-time number was used, the data can be considered ethical and correct.

Data for Jersey Sales was taken from Fanatics (the official partner for sports apparel of the MLB, NBA, NFL, and NHL) and the MLS Store (the official online shop for the MLS). Thus, it can be assumed that our data regarding jersey sales is accurate and ethical, coming from these two sources. Rather than sharing the exact sales volume, jersey sales were shared via a Top 25 ranking from 1 (the top selling jersey) to 25 (the 25th top selling jersey). This data was used due to the lack of online public availability of exact sales volume of jersey sales and is still representative of team popularity based on jersey sales. Because this data shared the ranking, the player and team of the jersey, it was easy to implement into a data table and assign the ranking positions to the teams. The jersey sales data from these sources does not include the unethical sales of bootleg jerseys or the sales of smaller sellers, which could either increase or decrease a specific jersey’s sales. However, these numbers represent the vast majority of ethical jersey sales and can be assumed accurate.

The MLS provided very little information as their league page was outdated, difficult to navigate, and did not provide strong and consistent historical data archives. Because of this, the Play-Off Appearances data for the MLS was taken from Wikipedia. This is not entirely ethical, and may be incorrect, due to the fact that the information can be altered by anyone. However, this information spanning the lifetime of the MLS was not available on any other data source, so it unfortunately had to be used in order to allow for its teams to be scored as accurately as possible.

It should be noted that for all data categories, data was only taken into account for current franchises. If franchises in a league were previously merged or abandoned, their old data, for example their number of play-off appearances, was not included in the dataset. Only current teams are being analyzed as the goal is to use the most recent data to determine who is currently the most popular team.

**The Methods:**

The previously mentioned data was implemented into a table with the following columns: ‘Team Name’, ‘Play-off Appearances’, ‘Twitter Followers’, ‘Jersey Sales’, and ‘League’. This data table was then downloaded as a CSV data file. The file was then imported into Jupyter Notebook as a dataframe (Figure 1). This dataframe included all five U.S. professional sports leagues: NBA, MLB, NFL, NHL, and MLS. By including the ‘League’ column, we were able to categorize the data by the five different leagues, enabling us to see how the teams in each league fare in relation to each other in the three data-analyzing categories: play-off appearances, jersey sales, and twitter followers. Then the top five organizations from each league will be pulled to receive scoring for the categories.

Once the dataframe was made, further analysis began. We used a single function as the basis for all of the categorical data analysis: the ‘stats’ function. This function took a dataframe, column, and league as parameters. By providing the function with the original dataframe, the name of the column which we’d be using as the category for analysis, and the name of the league we wanted to analyze, the function would output the top five teams based on the given column (category) and the given league. This was then used to analyze the requested statistics for each league.

In each of the following categories, a scoring system was applied to the top 5 teams in each league and in each category, and assigned points on a 5 point scale. The team placing 1st received 5 points, 2nd received 4 points, continuing down to the 5th place team which received 1 point. Any team ranking 6th or lower received 0 points. This was accomplished by implementing the ‘score’ function. After being used in the ‘stats’ function, a dataframe containing the top five teams assigned a score to said teams in a new dataframe column (‘Scores’). Eventually, a list containing all 15 dataframes (5 leagues, 3 categories) was inputted into the function, which creates a dictionary that puts the team name as the key and its total score as the value. The total score is a sum of all of the team’s scores from all of the dataframes, given that they placed in at least one of the 15 dataframes. If they did not, they received a total score of 0.

First, the number of play-off appearances were analyzed. To do this, we used the ‘stats’ function, taking the dataframe, the column name ‘Play-off Appearances,’ and the league from which we want the dataframe to represent, as inputs. With each resulting dataframe in descending order of play-off appearances, the dataframes looked similar to the one presented in Figure 2. The scoring system previously mentioned was applied at the end.

Next, the number of twitter followers for every organization in the league was analyzed. The process was identical to the one used for play-off appearances, except, this time, the input for the function was the column name ‘Twitter Followers’. The same scoring process was applied to the top 5 finishers, once again.

Finally, the jersey sales data needed to be analyzed. Aside from the ‘Score’ column, this analysis was conducted much differently than the other two categories. Since the data provided a list of the league-wide top 25 jerseys sold for each league, rather than each team and its respective number of jerseys sold, the rankings of the teams’ jerseys, if in the top 25, were implemented as a list in the ‘Jersey Sale Rankings’ column. For example, if a player was ranked 5th in their league for jersey sales, their rank of 5 was put into the ‘Jersey Sale Rankings’ list for their respective team. This was done for each player in the list of the top 25 jersey sales. If a team did not have any players in the top 25, they received a 0. Naturally, a team may have multiple rankings in the list if they had multiple players in the top 25 jersey sales. The individual rankings that each team had were in the dataframe as a string-list. This was rectified by turning the numbers into integers. To make integers in the ‘stats’ function, the dataframe and column were called on to equal itself with the addition of .astype(int). To account for the multiple jersey sale rankings within some teams, we implemented an additional scoring system. We did this by putting the column ‘Jersey Sale Rankings’ into a list, looping through all of the values in the column, splitting the data at commas, and then assigning scores to the rankings. The scores were only assigned if the integer of the jersey sale ranking was greater than zero. If it was, the number would then be added to the team’s cumulative score as the absolute value of 26 minus the number. This allows for all jerseys in the top 25 rankings to receive a score corresponding with their placement (high finishers receiving high points, no jerseys receiving zero). The first place player gave their team 25 points, while the twenty-fifth place player gave their team 1 point. For example, if a team had the #1, #10 and #15 ranking players for jersey sale rankings, they would receive a cumulative score of 50 (25 + 15 + 10). These total scores were put into a new list, then added to the dataframe as the column ‘Jersey Sale Rankings Score’ (Figure 3). Finally, the data from the ‘Jersey Sale Rankings Score’ column was put through the stats function and followed the same process as the other two categories.

The final analysis was conducted after the dictionary was created and renamed in the main. It was then used in the ‘visuals’ function to create the barcharts in Figure 5 and Figure 6. To have the results in the desired order, the dictionary was sorted to be in descending order based on total score (the values). The other input for this function, top, allows for us to determine how many teams are being shown in the chart (ex. Figure 5, top = 10).

In the main function, a small for loop was used to pull the top score and team from the dictionary. The result was shown by a print statement to return to the user the most popular team in regard to the scores of the categories compared across leagues (Figure 4).

The set-up of our functions and formulas allow for this program to be used on both future and previous years, as long as the necessary data is implemented in a similarly formatted CSV file. Alterations could be made in the main function as necessary.

**The Results:**

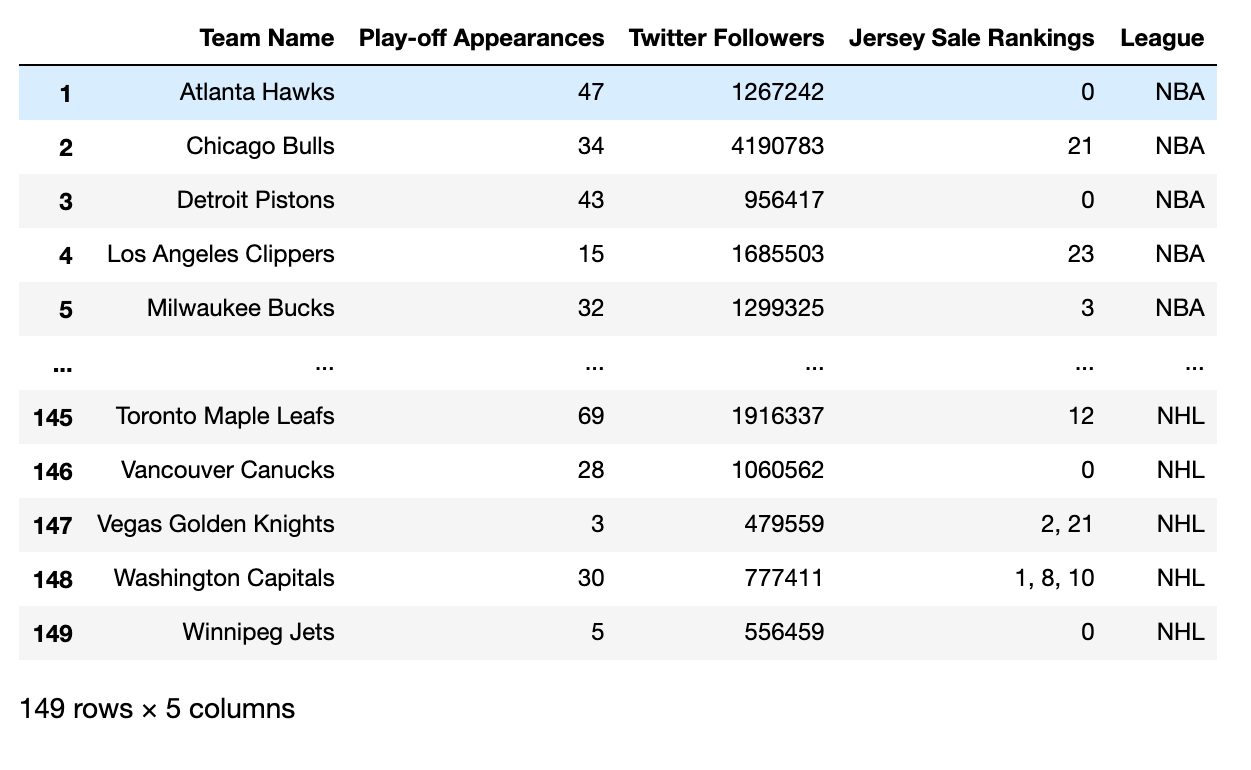
In regard to the hypothesis, mentioned in **The Problem Statement**, it was hypothesized that the most popular american sports team would be either the New York Yankees or the Dallas Cowboys. This is because internationally, people generally consider them to be the stereotypical ‘American’ sports teams.

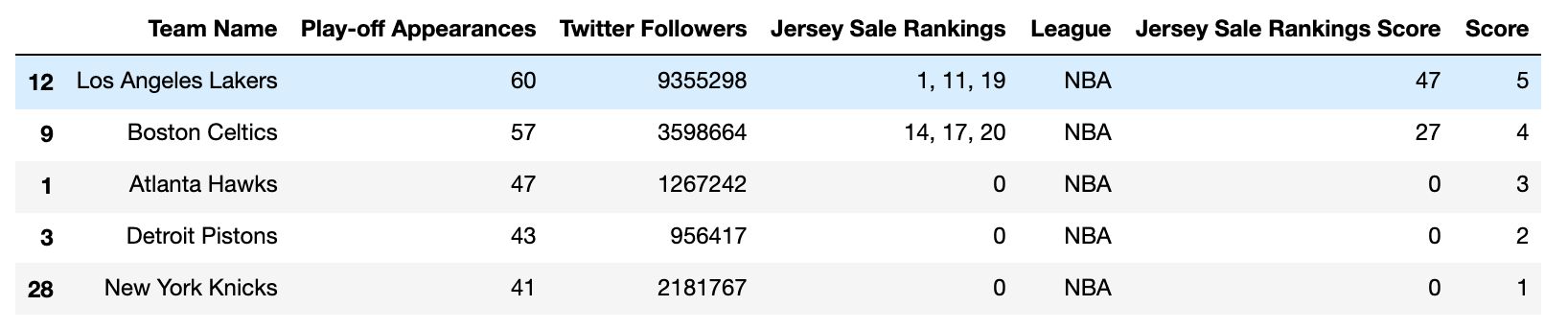
Our data concluded that the overall most popular American sports team was the Los Angeles Lakers. In both their league and overall rankings, they placed first with a total score of 15 points based on the scoring system (Figure 4). They were the only team in any league to place first in all three categories, evident by the score and it’s placement in the bar graph (Figure 5). This allowed them to win against the other contenders, including the hypothesized winners of the New York Yankees and Dallas Cowboys, both of which tied for 2nd place with 14 points each. As shown in Figure 6, total scores for most of the other teams had a large deficit in comparison to the top three. Once out of the top five, the scores are equal to or less than half of the score of the winner for the position of the Most Popular American Sports Team in the World (as of 2020).

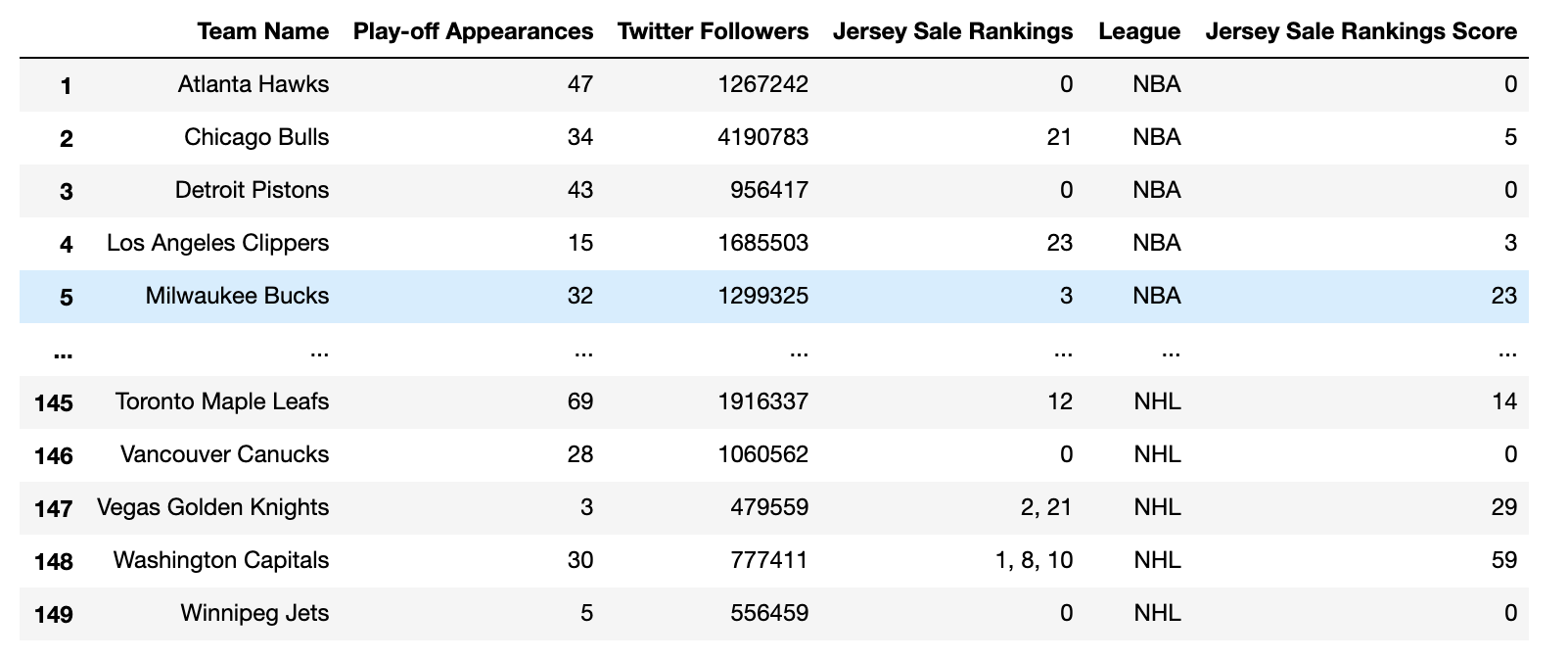
Our uses of scoring systems were a major strength within the program as they allowed for a variety of categories, spanning across other categories, to become comparable to one another. This allowed for analysis of what would otherwise be unrelated data or data difficult to interpret and compare with code. The organization of the CSV file and the cleaning of the code also contributed greatly to our success as it simplified the implementation and manipulation of the dataframes.

Given more time, we would consider fine-tuning the current overall scoring system. We would do this by either implementing a weight to the scoring system or continuing the scoring system past the top five from each category. While this would help ranking scores be more differentiated and avoid ties throughout the 149 teams, we had a difficult time implementing these additions due to a variety of issues. With weighting, we could not draw a conclusion as to which of the three analysis categories was considered more important than the others. We also, at the time, did not deem it necessary to award points to all teams as we assumed any team not in the top five of a category would be in the running for the ultimate winner, however, we now recognize that allowing for a higher point distribution would help to avoid ties if the entirety of rankings was desired.

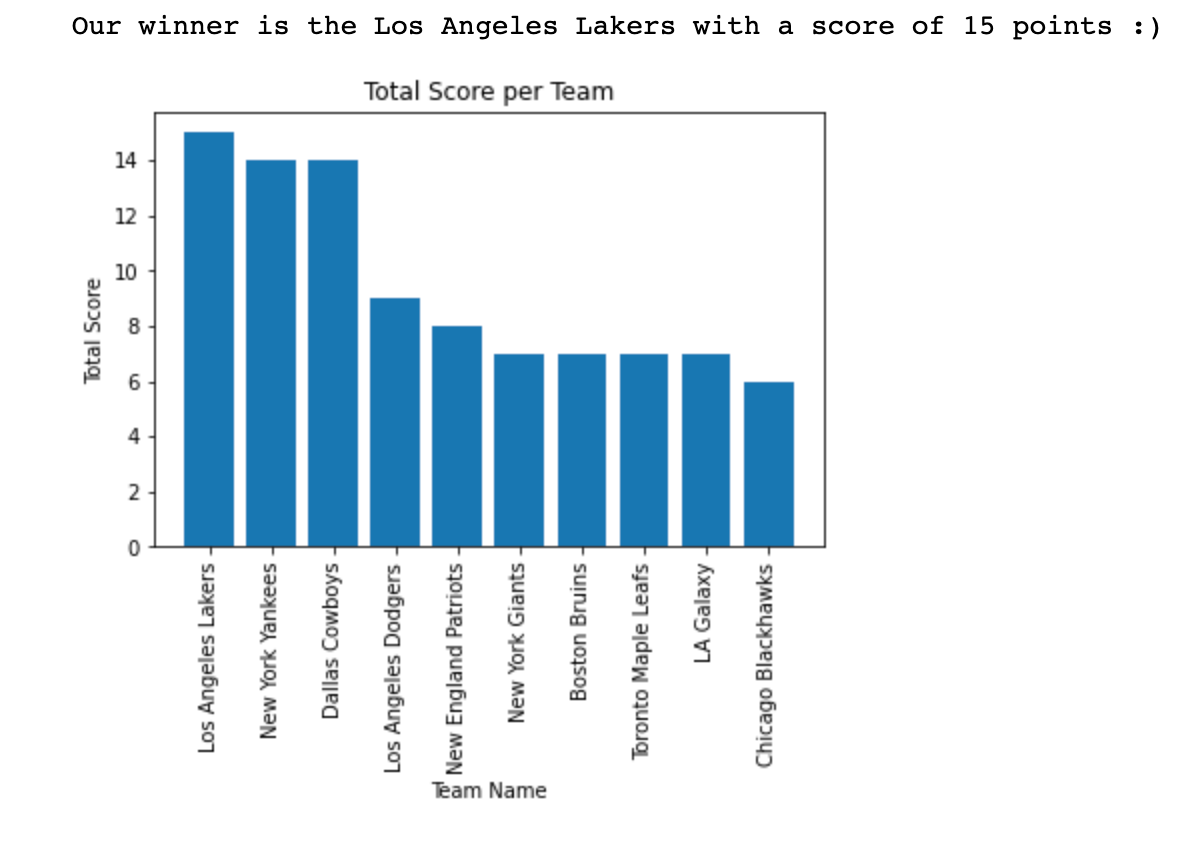
**Exhibits:**

**Figure 1**  **Figure 2**



**Figure 3**

**Figure 4**



**Figure 5 Figure 6**