Kyle Ford

CPSC 5530

Due Date: 4/30/2023

Final Project Report

Introduction:

Professional sports represent one of America's most popular entertainment industries. Throughout time, fans of various sports have shown great interest in their favorite teams and their performance throughout a given season. As such, professional sports have grown not only in popularity, but in monetary value as well. Nowadays, professional sports leagues are enterprises that represent billion-dollar industries. The money that leagues generate is staggering. Thes revenue streams come in the form of memorabilia and merchandise sales, ticket sales, game concessions, TV and streaming revenue, and legal sports betting operations.

Specifically, Major League Baseball (MLB) represents one of the four most popular professional sports leagues in America. Baseball has long been one of the most popular sports in American culture, often colloquially referred to as "America's pastime." As such, there is great interest in quantifying the sport's popularity and ability to generate revenue. The MLB season culminates in a championship series called the World Series, in which a champion for the season is declared. The ultimate goal of a team's season is to win the World Series.

This paper will explore a few key visualization domains specific to Major League Baseball. First, we will explore the prevalence of the league nationally by exploring stadium location throughout the country. Theoretically, the geographic distribution of each stadium will help indicate areas of the country where fandom is more concentrated. It stands to reason that areas featuring their own MLB team will be more invested in the sport than areas that do not feature their own team or stadium. Additionally, identifying geographic disparity through stadium distribution can aid the league in identifying opportunities for league expansion. MLB and other professional sports teams occasionally review opportunities for expansion through the addition of new franchises that would increase the competitive level of the league and provide opportunities for novel revenue streams.

Second, we will explore descriptive statistics specifically related to the 2022 Philadelphia Phillies. While the team ultimately came up short of winning the World Series, the 2022 Phillies represented an underdog team that barely made the playoffs. They qualified for the Wild Card round and had to win a do-or-die playoff game on the road against the Saint Louis Cardinals just to qualify for the postseason. From there, they went on to beat defending World Series champions Atlanta Braves and the highly favored San Diego Padres to qualify for the World Series. Ultimately, they lost the World Series to the Houston Astros and came up short of achieving the ultimate goal. This topic is of particular interest to me, as the Philadelphia Phillies are my favorite team. Through visualizing various player statistics from the 2022 season, we can identify a few things. First, we will analyze the true ability of the 2022 team. Did this team overachieve in the postseason, or did they underachieve in the regular season? Was this team lucky to get as far as they did, or were they miscast and discounted by mistake? Finally, is there any reason to believe that the results of the 2022 can be replicated and built upon, so that Philadelphia may be World Series champions in the near future? Through a visual analysis of the team's 2022 statistical profile, we will explore answers to these questions.

Finally, we will explore an aspect of one of the league's key revenue streams: concessions. Concessions at MLB games are romanticized in ways they are not in other professional sports. Famous seventh inning stretch songs make reference to peanuts and cracker jacks, and eating a hot dog at a baseball stadium is synonymous with game attendance. Moreover, many fans choose to enjoy a beer (or a few) while taking in a Major League ballgame. We will explore the cost of a beer at major league stadiums over time in an attempt to explore one of the league's key revenue streams. Particular attention will be paid to how this revenue stream has changed over time. Furthermore, since I am a Phillies fan, we will compare the average price of a beer at Major League stadiums to the cost of a beer at the Phillies stadium, Citizens Bank Park.

Datasets and Data Types:

Dataset 1: MLB Stadiums Dataset

The first goal of this project was to explore the distribution of Major League ballparks geographically. The dataset explored to achieve this goal was obtained from the following URL:

https://www.seamheads.com/ballparks/about.php

This dataset is organized as a simple table. It contains thirteen columns and 255 rows. As this dataset is organized as a simple table, each column represents a distinct dimension, or attribute, of the data. Similarly, each row represents its own entry, and each cell represents the attribute value for that given entry. Semantically, each row represents an individual Major League ballpark. Of course, there are not 255 active MLB ballparks, so we know we are dealing with a historical element. In terms of attributes of interest, we are specifically interested in the latitude and longitude columns. These columns are geographical values that will be used to map individual active ballparks. Additional dimensions present in this data include nominal values such as stadium name, city the stadium is located in, state the stadium is located in, and a note relevant to the individual stadium. Additionally, since some of these ballparks are not active, the dataset includes date values representing the date the stadium opened, and, if applicable, the date the stadium closed.

For the purposes of this project, specific interest is given to the name of the stadium and the latitude and longitude coordinates of the stadium, thus implying we will be dealing with three dimensions of the data.

Dataset 2-3: MLB Batting and Pitching Player Statistics

The second and third datasets explored for this project are similar in nature, so they will be discussed in tandem. Both datasets are simple tables that describe player statistics for the 2022 MLB season. One of the datasets describes batting statistics for offensive players, and the other describes pitching statistics for pitchers. Special attention will be paid to players who play for the Philadelphia Phillies as we attempt to tell the story of their run to the World Series in 2022. Both the batting and pitching datasets are obtained from the following URL:

https://www.kaggle.com/datasets/vivovinco/2022-mlb-player-stats

First, we will discuss the 2022 batting dataset. As mentioned, this dataset is organized as a simple table with rows representing different dimensions of the data and rows representing different entries in the table. In this dataset, each row represents a different MLB player who took an at bat during the 2022 season.

The majority of the data in this dataset is numerical. Specifically, this data describes descriptive statistics describing how individual MLB players performed in their given at bats during the season. Specific numeric data dimensions include hits (H), home runs (HR), and RBIs. These numerical data elements represent cumulative statistics in that they represent the total amount accumulated during the 2022 season. This dataset also features efficiency-based metrics such as on-base percent (OBP), slugging percent (SLG), and OPS. Even though both categories are numeric in nature, it is important to note that they measure different aspects of player performance.

In addition to purely numerical data elements, this dataset features other data types as well. These data types include nominal data types such as player name and team and ordinal data types such as player rank. For the purposes of this study, we pay specific attention to player name, rank, and team. These dimensions are used to quantify the descriptive statistics described by the numerical data types.

The dataset containing the data for MLB pitchers is similar in its organization to the dataset containing the batting data. The dataset is organized as a simple table where each column represents a different data dimension and each row represents an individual entry in the table. Specifically, each row represents an individual pitcher.

The individual data types contained in the columns of the table are also similar to those contained in the batting database. The majority of the attributes encoded in the table are numeric. These data types represent various descriptive statistics for MLB pitchers. Also like the batting dataset, there is a mix of aggregate, cumulative numeric data types, and efficiency metrics quantified in terms of per nine innings. Since pitchers rarely pitch an entire baseball game, we will focus on statistics that are normalized over nine innings. These statistics include earned runs allowed (ERA), WHIP, hits/9 innings, home runs/9 innings, walks/9 innings, and strike outs/9 innings. This dataset also contains the same nominal and ordinal data types that the batting dataset did, including player name, team, and rank.

One final note regarding the organization of these two datasets. While each data type can be said to be represented by its own column in the table, in actuality the .csv files are organized such that all of the data is encoded within a single column. Individual attributes are delimited by a semicolon. Organizationally, this does not impact the overall data encoded in the table. However, this will play a role once we begin the data preprocessing stage.

Dataset 4: The Price of a Beer at a Baseball Game

The final dataset explored for this project charts the average price of a beer at a MLB game at each stadium from the years 2013-2018. This dataset can be found at the following URL:

https://data.world/makeovermonday/2018w43-what-will-a-beer-cost-you-at-every-major-league-ba

Organizationally, this is a fairly simple dataset. Like the other datasets in this project, it is organized in a simple table, however; it contains far fewer dimensions than the other datasets. Each column represents its own dimension, of which there are seven. Of these seven dimensions, three are numeric. These data types include price, size, and price per ounce. For the purposes of this project, we pay special attention to the price dimension. Of the remaining four dimensions, three are nominal. These data types include team, nickname, and city. Semantically, the team dimension is a concatenation of the city and nickname column and is the dimension that will be used for visualization.

The remaining dimension of this dataset is the dimension that differentiates it from the other datasets. This dimension is year, thus introducing a time series element not present in the other datasets. Thus, in terms of visualization, we will be able to chart the price of beer over time and measure change.

Software Packages and Libraries

The visualizations for this project are compiled using the Python programming language. While I considered using R for this project, I feel more comfortable programming in Python and wanted to continue mastering the language. Specifically, I used Anaconda Navigator's Jupyter Notebook programming interface to compile the source code for the project. I find the notebook environment to be immensely intuitive and like being able to divide my code into separately runnable segments.

The first, and perhaps most fundamental, library that I imported for this project was the pandas library. This library was imported using the traditional alias pd. All of the datasets that I used were formatted using the .csv format, therefore, when importing my data to Jupyter Notebook I made heavy use of the .read_csv() method to create data frames out of my data.

Aside from simply importing the data as a data frame, I made heavy use of pandas methods in the data preprocessing stage of the project. Specifically, I used pandas to select certain rows and reformat the data frames, select individual rows for plotting, and filter data needed for visualization.

In addition to pandas, the first visualization that I created was created using the folium library. This library allowed me to create an interactive map using real time GPS data. In this way, I was able to plot geographic information using simple latitude and longitude data types. Folium was ideal for this specific visualization because the library allows the user to zoom, pan, and further manipulate the visual layout of the plot depending on their individual needs.

Finally, I imported the Plotly library to create visualizations for datasets 2, 3, and 4. Plotly allowed me to create interactive visualizations for individual player statistics and the price of a beer at a ballgame over time. Specifically, I was able to create horizontal bar charts to show cumulative statistics. Furthermore, I was able to create parallel coordinates plots to compare and contrast player performance across a variety of efficiency statistics. Finally, I was able to create interactive line graphs to visually show how the average price of a beer has changed over time.

Plotly was an especially useful library because of the variety of interactive features that are built into it. One of the most important interactive features that these plots make use of is hoverability. The user is able to hover over a visual aspect on the plot in order to obtain more data about that variable. Additionally, Plotly allows for the ability for the user to click on various data elements in order to obtain more information. Finally, I was able to create visualizations where the user is able to filter data in the visualization by simply clicking on certain data elements. The combination of all of these interactive features makes the plots more immersive and greatly improves the user experience when interacting with the visualizations.

Data Preprocessing and Code Preparation

Visualization 1: Folium Map of MLB Stadiums

The first visualization created for this project is a Folium map depicting every active MLB stadium. Recall that the dataset for this visualization contains information of 255 MLB stadiums, most of which are no longer active. Therefore, the first step in the preprocessing stage is to filter out all inactive MLB stadiums. This is accomplished by dropping the rows containing a value in the end date column. If the stadium is active, then there is no value in the end date column, and the value shows as NaN. We drop the rows containing this value and are left with a data frame containing only rows that contain no end date, therefore implying that they are currently active. We can confirm our data is correct by measuring the length of the column. We know there are 30 teams currently in the MLB, and that none of them share a stadium. Therefore, we print the length of the column to confirm we have 30 rows in our column not including the header.

Once our data has been prepared, we can begin creating the plot. We want a map that shows the United States as a base layer. This is accomplished by creating a folium map with the location set to the value [40, -95]. A simple Google search gives us these values to depict the United States. We want the map to be adequately zoomed in such that the entire United States takes up all of the space in the plot. This is accomplished by setting the zoom start argument to 4. We print the map to ensure an accurate base plot.

Now that we have our base map object, we can proceed with plotting the stadium locations on the map. This is accomplished using a for loop to iterate over the rows in the data frame. The position of the marker is determined by iterating over the latitude and longitude columns. We add a popup marker based on the name column such that when the user clicks on the popup marker it displays the name of the stadium. The net result is an interactive map of all of the active stadiums in the MLB. Interactive features include the ability to zoom, pan, and click on a popup marker to display the name of the stadium.

Visualizations 2-5: Horizontal Bar Charts of Philadelphia Batters and Pitchers

Visualizations two through five are created in a similar fashion and depict similar data. Specifically, we create two separate horizontal bar charts for batting statistics, batting average and home runs, and two more horizontal bar charts for pitching statistics, ERA, and WHIP. These plots are designed to tell the story of the Philadelphia Phillies run to the World Series in 2022, so we limit the plot to display players that play for the Phillies.

The first step in the preprocessing stage is to import both the batting and pitching data from our two datasets. This is accomplished using the pandas read .csv method. Once our data frames have been defined, the next step is to filter them down to only players that played for the Phillies in 2022. This is accomplished by creating a new data frame where the value in the 'Tm' column is equal to the value 'PHI'. The net result is a new data frame containing only Phillies players.

At this point, the data is prepared, and we can begin plotting. We will create the plots using the Plotly library. We create a Plotly express object using the. bar method with the data being drawn from the Philadelphia players data frame we just created, the name of the player on the y-axis, and the statistic we are visualizing on the x-axis. In order to make the bar chart horizontal instead of vertical we set the orientation argument to 'h'. Finally, we give each plot an appropriate title and display the plot. Upon immediate review, we notice that the plot is too small along the y-axis to display all of the player names. In order to remedy this, we set the height argument to 1000 and show the plot again. The net effect is four effective, interactive plots that display the batting or pitching statistic they are designed to visualize. Since the height argument is changed, the user has the ability to scroll to see additional players. Additionally, hovering the mouse over an individual bar will display the player's name and the numeric value of the statistic that is being visualized.

Visualizations 6-8: Parallel Coordinates Plots Comparing Player Statistics

In addition to quantifying raw descriptive statistics using a bar chart, we create Plotly parallel coordinates plots to compare how individual players performed over a variety of statistics using the same plot. Specifically, for batting we create two such plots, one comparing hits, RBIs, and home runs, and another comparing OBP, SLG, and OPS. For pitching, we create one plot comparing H/9, HR/9, BB/9, and SO/9.

In order to limit the number of coordinate lines on our plot, we will further filter our data frame to select particular players of interest. For batters, I selected some of Philadelphia's best hitters. These players were Alex Bohm, Bryce Harper, Kyle Schwarber, Nick Castellanos, Rhys Hoskins, and Brandon Marsh. The pitchers selected Jose Alvarado, SirAnthony Dominguez, Aaron Nola, Noah Syndergaard, and Zach Wheeler. These data frames are further filtered by creating individual data frames containing a single row representing each player. Once each data frame is created, we recombine them using the concat method in the pandas library. The net result is a data frame containing the data for only the players we want to plot.

Each parallel coordinates plot is created using Plotly express. We create a Plotly express parallel_coordinates object, and define the filtered data frame we just created, set the color argument to the rank dimension, the dimensions argument to the list of statistics we are going to compare, and set the color scale to teal rose. Finally, we give each plot an appropriate title and display the plot. The net effect is multiple interactive parallel coordinate plots that compare multiple descriptive statistics. The interactive elements of this plot involve the ability to filter data by clicking on individual dimensions displayed on the plot.

Visualization 9: Line Graph Showing the Price of a Beer at a Baseball Game Over Time

The final visualization created for this project is an interactive line graph showing the price of a beer at each MLB stadium over time. It is created using Plotly express. Due to the simplicity of the dataset, the only preprocessing step that needs to be taken is importing the data to a pandas data frame.

Once the data has been imported, we define a Plotly express object using the line method, define the data frame we are pulling the data from, set the x-axis equal to the year column in the data frame, and the y-axis equal to the price column in the data frame, the color argument equal to the team column, and give the plot an appropriate title. Finally, we give the plot markers by updating the trace such that mode equals “markers+lines” and display the plot. The net effect is an effective, interactive line plot showing the price of a beer at an MLB game during the years 2013-2018. The interactive elements of this plot include hoverability, where hovering over a point on the line will display the team, the year, and the price of a beer. Additionally, there is a built-in legend that the user is able to filter the data by clicking on individual team names displayed in the legend.

Visualizations and Discussion

Visualization 1 Folium Map:

This visualization is an interactive folium map depicting every active MLB stadium. In its default view, the visualization appears as is shown below.

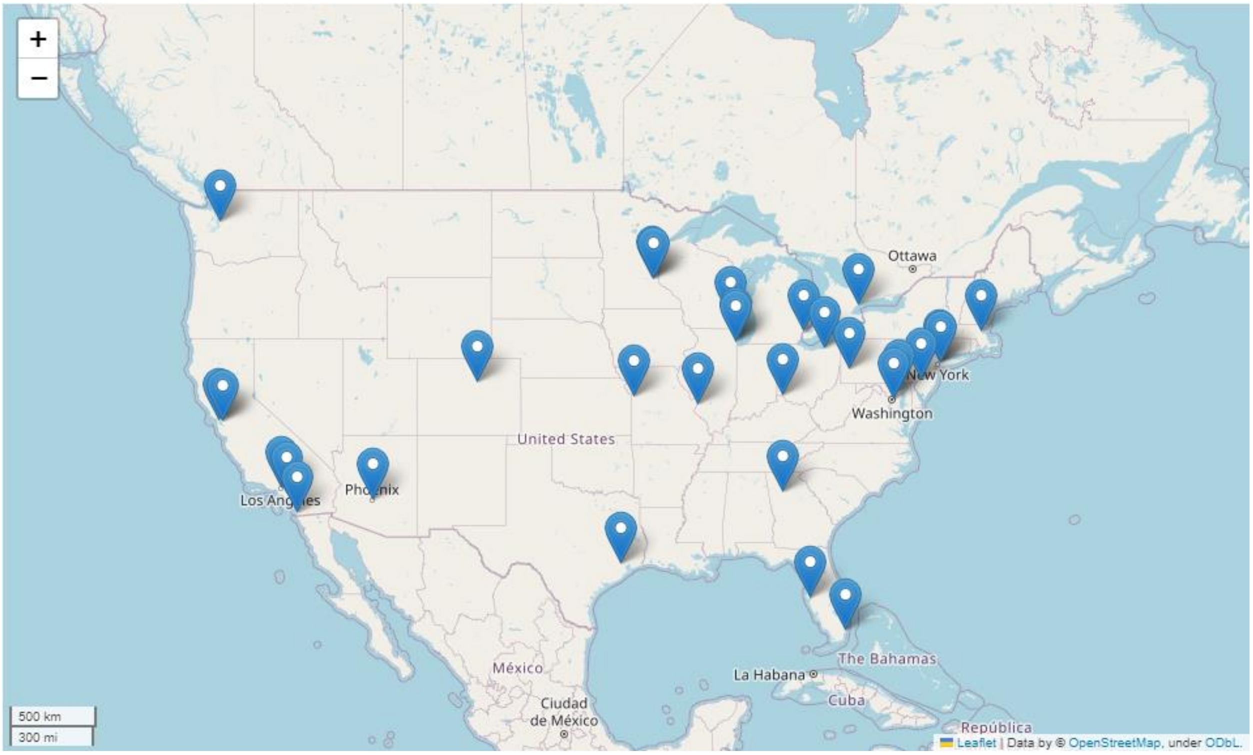

Figure 1: Folium Map of Every Active MLB Stadium

What is immediately noticeable about this plot is the concentration of MLB stadiums in the northeast region of the United States. The majority of the active stadiums are concentrated in the northeast and in the Midwest regions of the country. Additionally, there is a dearth of MLB stadiums in the pacific northwest regions of the country.

One of the goals of this project was to identify areas for possible expansion for the MLB. What this plot illustrates is a lack of stadiums in some major metropolitan areas including Charlotte and Nashville, specifically. It also shows a potentially untapped market in that the only stadium in the Pacific Northwest is located in Seattle.

Also note, this plot enables the user to click on the popup marker and zoom in on a stadium. When a user does this, the plot looks is it does below.

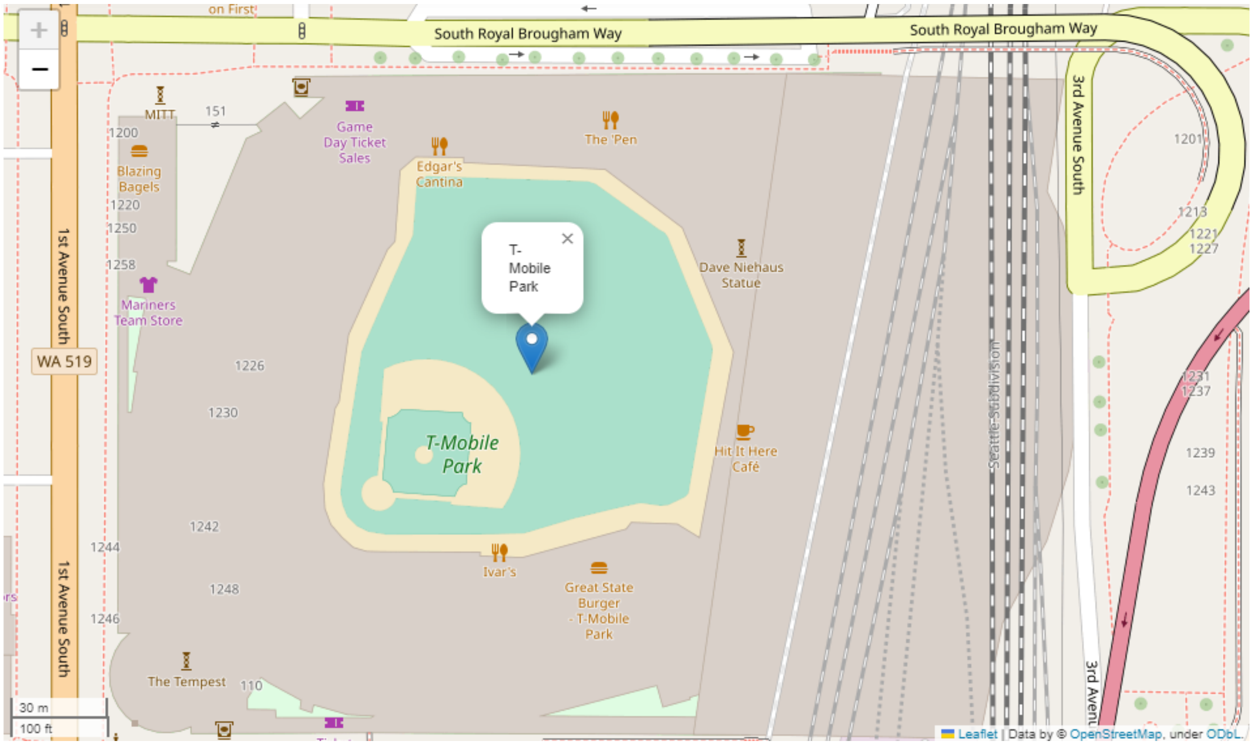

Figure 2: Zoomed in View of Folium Map

Visualizations 2-5 Horizontal Bar Charts:

Visualizations 2-5 are the Plotly express horizontal bar charts that were created to visual individual player statistics for the 2022 Philadelphia Phillies players. First, we will consider the two visualizations for batting statistics, and then the two visualizations for pitching statistics. Below are visualizations 2 and 3, representing Phillies players batting averages and home runs, respectively.

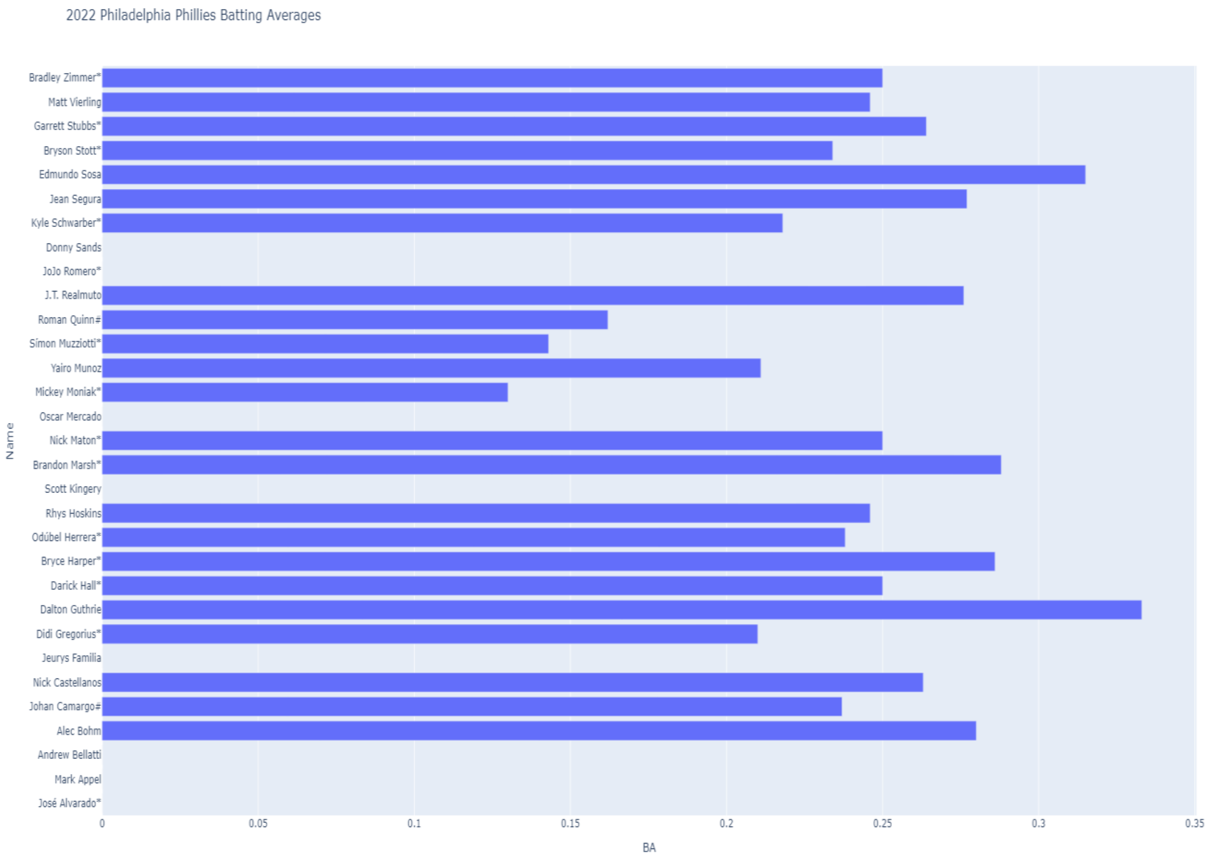

Figure 3: 2022 Philadelphia Phillies Batting Averages

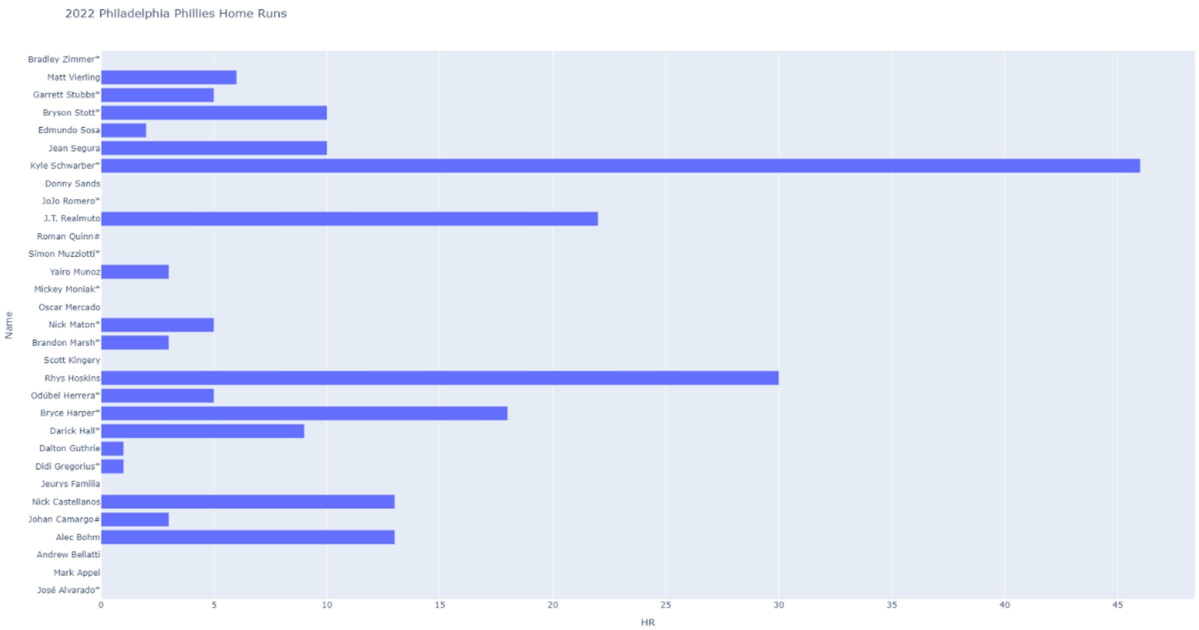

Figure 4: 2022 Philadelphia Phillies Home Runs

First, with regard to the batting average plot, notice that certain players have a .000 batting average. This certainly isn't good, as it means that a player never yields a hit in any of their at bats throughout the season. However, upon further inspection, we see that these players are pitchers. Generally, pitchers are not known for their hitting, so this is a palatable and understandable result. However, it is worth noting that it represents an outlier value in our visualization. On the other end of the spectrum, we see that the players with the two highest batting averages are Dalton Guthrie and Edmundo Sosa, respectively. These players are generally not thought of as some of Philly's most prolific, so this result is fairly unexpected. It points to the idea that some of the Phillies best hitters excel in other areas of batting that are not reflected through batting average. The remaining values tend to be clustered in between these two extremes, and there is not much separation in terms of batting average.

When examining the home run plot, we see a different result than was observed with the batting average plot. First, there are far fewer values for players that hit zero home runs. Again, a good portion of these players are pitchers who took at bats for the Phillies, but there are also some position players who took at bats and failed to register at least one home run. On the other end of the spectrum, we see more extreme values as well. The players with the most home runs include Kyle Schwarber, Rhys Hoskins, and Bryce Harper. These are some of Philadelphia's best players. Kyle Schwarber has won the Home Run Derby in the past, and Bryce Harper has won the NL MVP award. It is worth noting that, while these are Philly's best players, they did not appear at the top of the batting average plot. However, they did show up at the top of the home run plot. This is indicative of the fact that Philly has a roster built around power batters. This means that they might be less likely to register a hit than other players, but, when they register a hit, it is much more likely to be a home run. This helped fuel Philly's run to the World Series in 2022. Power hitting is an unpredictable metric, and teams can get hot at the right time. Philly's proclivity for power hitting leaves it vulnerable to suboptimal results over the course of a long season but leaves open the possibility of positive results over a short-term postseason series.

Now, consider the pitching plots. They are presented in their default form below. Figure 5 depicts a horizontal bar chart for ERA and Figure 6 depicts a horizontal bar chart for WHIP.

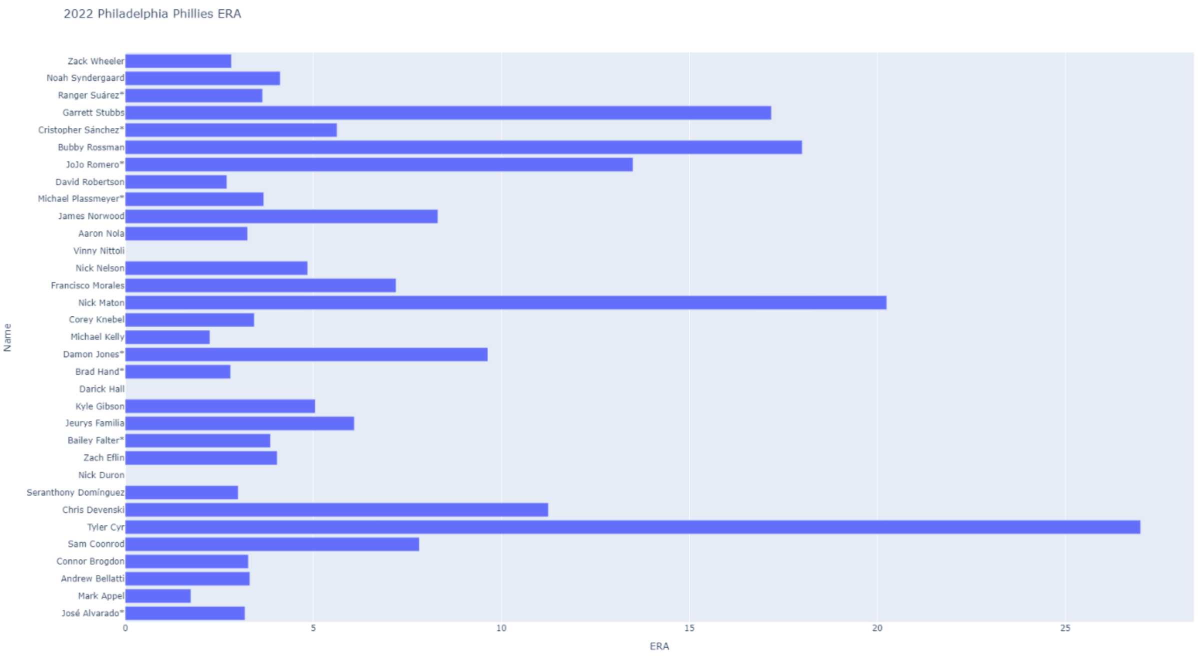

Figure 5: 2022 Philadelphia Phillies ERA

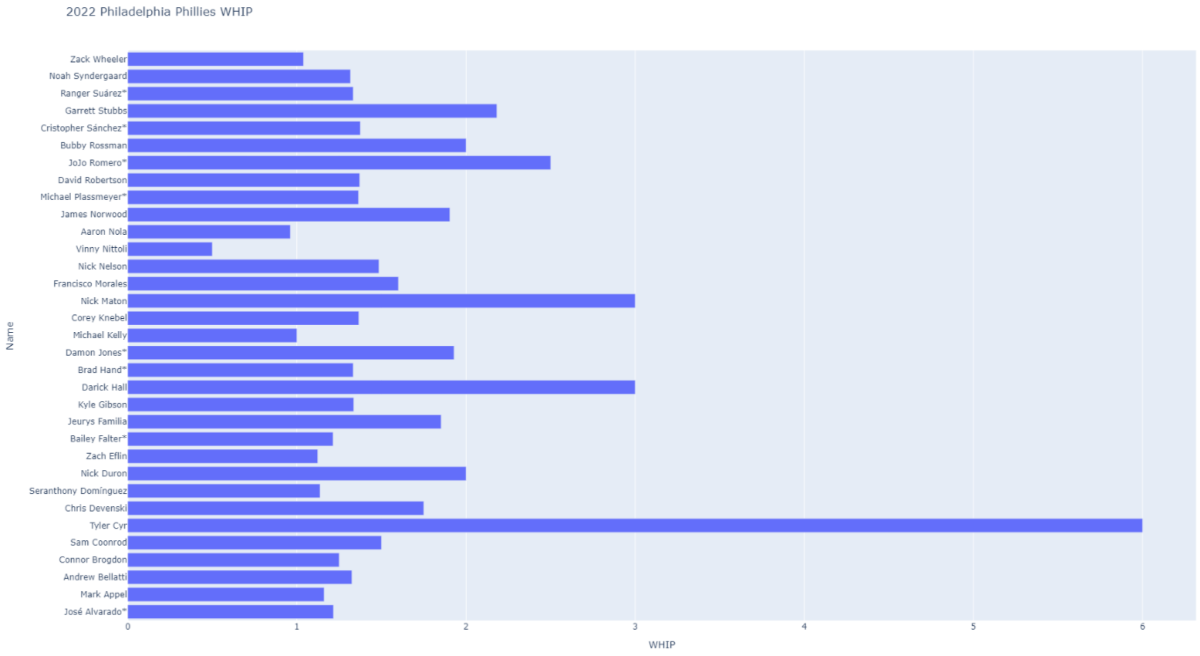

Figure 6: 2022 Philadelphia Phillies Whip

First, it is worth noting that, in terms of general consensus, Philadelphia features two world class pitchers, Aaron Nola and Zach Wheeler. This is reflected in these charts and Nola and Wheeler feature some of the lowest ERA's and WHIP's. This works inversely to most hitting statistics, as a pitcher's goal is to concede as few runs as possible while they are in the game. Thus, the plots we have generated are able to quantify what is considered common knowledge. This idea bore itself out during the 2022 postseason, as Nola and Wheeler started the majority of the Phillies' postseason games, which translated to increased success. Generally, in the postseason teams reduce the size of their pitching rotation such that their best pitchers pitch more often. This contributed to Philadelphia's success in the postseason, as these visuals indicate.

Furthermore, we see a few outlier values indicating poor pitching. Specifically, these values are related to Garrett Stubbs, Bubby Rossman, Nick Maton, and Tyler Cyr. Given these statistics, one might wonder why their opportunity to pitch would not be reduced. In other words, if these pitchers perform poorly, why continue letting them pitch? Simply put, the MLB regular season is a marathon featuring many games spread over the course of a calendar year. Depth in the pitching staff is necessary to preserve the state of the roster and make it to the postseason.

However, once the postseason begins, the rotations shorten and only the best players play. Again, this is what happened in 2022 with the Phillies. While they had a fairly mediocre regular season, once the postseason started only their best players played, resulting in increased success.

In terms of interactivity, each of the four plots discussed above features a hover ability. By hovering the mouse icon over an individual bar, the user is able to see the player's name and the value associated with the statistic being plotted.

Visualizations 7-9: Parallel Coordinates

Next, consider the Plotly express parallel coordinates plots that were developed for this project. We develop two plots for batting statistics and one for pitching statistics. With these plots, instead of including the statistics for every player, we limit the plot to display only a handful. For batters, we include Alec Bohm, Bryce Harper, Kyle Schwarber, Nick Castellanos, Rhys Hoskins, and Brandon Marsh. For pitchers, we include Jose Alvarado, SerAnthony Dominguez, Aaron Nola, Noah Syndergaard, and Zach Wheeler. The plots are displayed below in their default form.

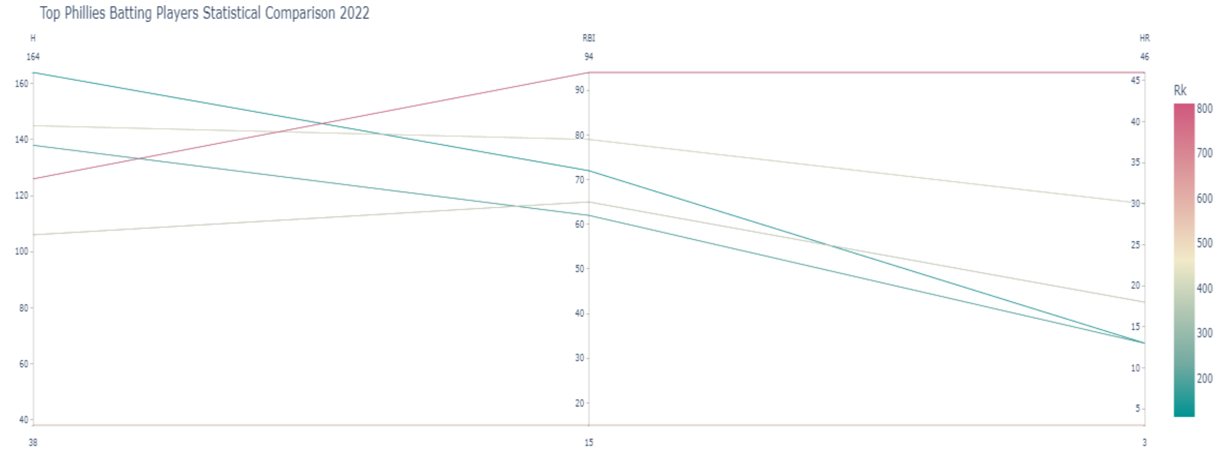

Figure 7: Comparison of H, RBI, and HR

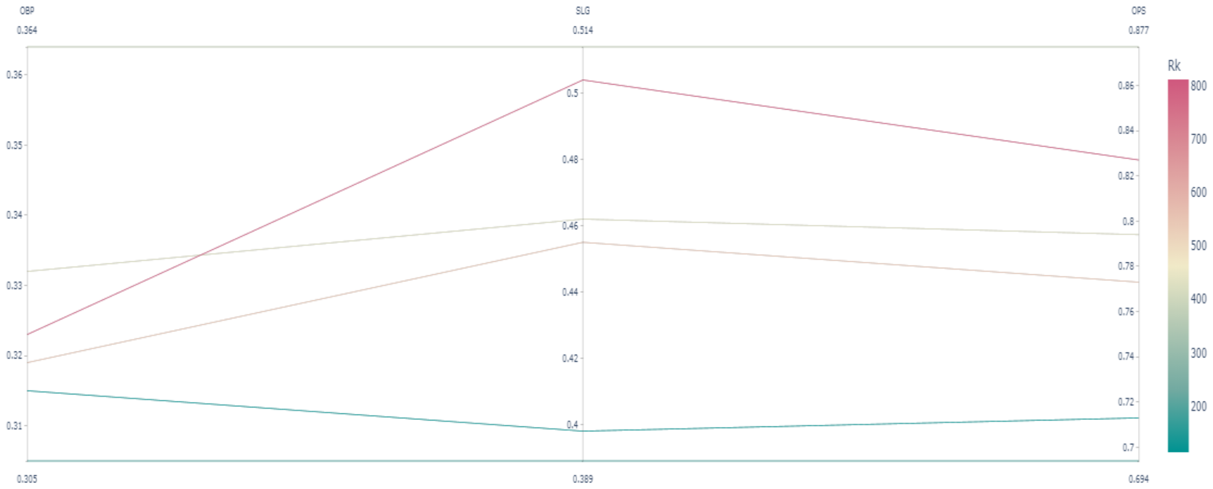

Figure 8: Comparison of OBP, SLG, and OPS

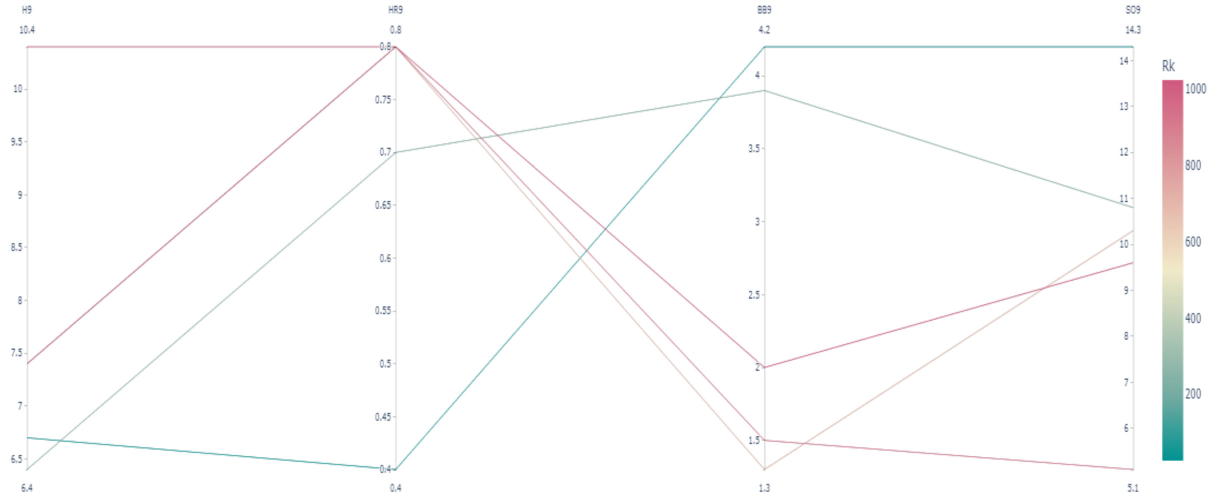

Figure 9: Comparison of H/9, HR/9, BB/9, and SO/9.

These results display a wide range of outcomes. First, with regard to Figure 7, note how the plot seems to point to a few correlations between the statistics plotted. First, note how the players with more RBIs also hit more home runs. This points to the idea that RBIs can be indicative of home run success. Note however, that this result is anecdotal and not causational. Further study would be needed to prove causation. However, the visualization points to the idea that players who hit more RBIs also hit more home runs. However, this correlation does not hold up for hits. Players who obtain more hits oftentimes hit fewer home runs, and sometimes more RBIs as well. The same correlation does not exist for this statistic.

Now, referring to Figure 8, note that there does seem to be a correlation between SLG and OPS. Players who have higher SLG values also have higher OPS values. This is consistent for all players plotted. Again, this result does not prove causation, but the correlation is clearly displayed by the visualization. However, OBP does not correlate well to any of the statistics.

Visualization 10: Line Graph

The final visualization for this project involves charting the price of a beer at an MLB game from the years 2013-2018. Like many of the other plots, this is a Plotly express plot. The plot, in its default form, is presented below.

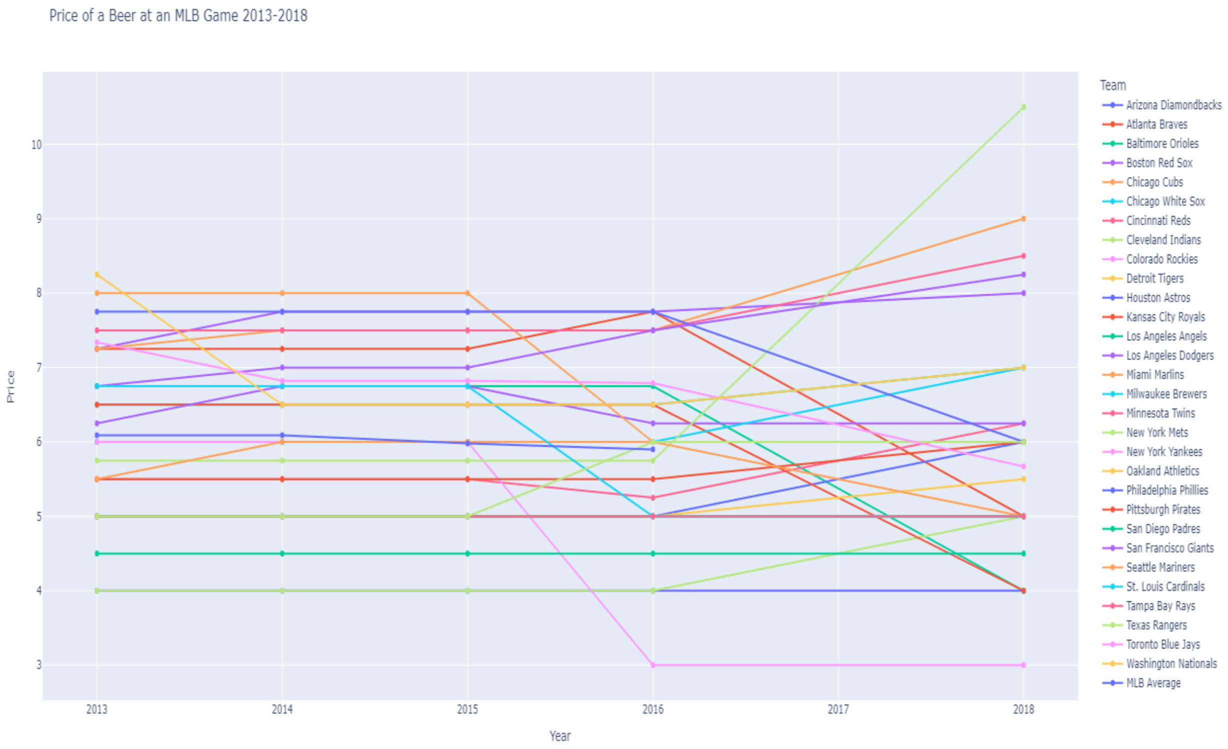

Figure 10: The Price of a Beer at an MLB Game 2013-2018

This plot shows a few key things. First, note the relative stability of the price of a beer at a game from 2013-2015. During these years, there is not much movement on the line chart, indicating that the price of a beer did not change, or changed very little. We don’t see significant change in the price of a beer until 2016. At this point in time, we see larger increases and decreases in the price of a beer. As of the last year of data plotted, the most expensive game to buy a beer at was the New York Mets at $10.50/beer, and the cheapest game to buy a beer at was the Colorado Rockies at $3/beer.

Interactivity is an integral part of this visualization. By clicking on the team names displayed in the legend, we can add and remove specific lines. This allows the user to filter the results that are displayed to better parse out comparisons of interest. For the sake of comparing how expensive the price of beer at a Phillies game is, we illustrate the utility of this feature by comparing the Phillies to the other teams in their division. The Phillies are a member of the NL East, which contains the Atlanta Braves, New York Mets, Miami Marlins, and Washington Nationals. When we apply the appropriate filters to display just the NL East teams, we achieve the following plot.

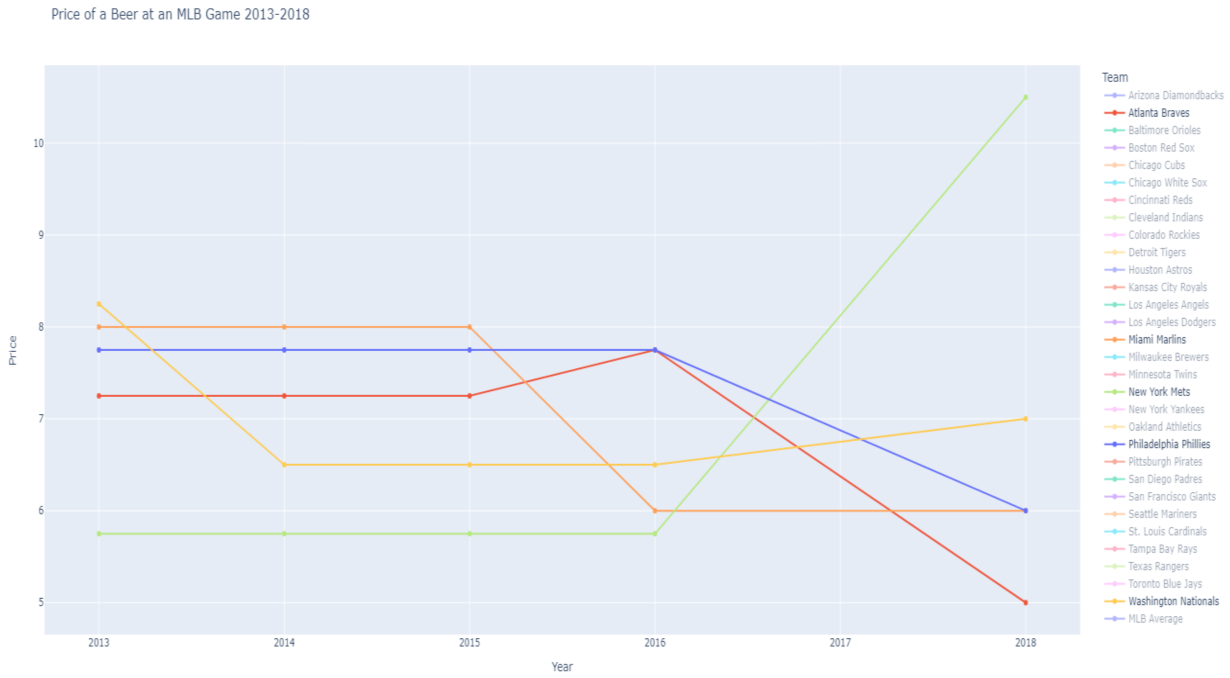

Figure 11: Price of a Beer from 2013-2018 at NL East Games

From this plot, we see that the cheapest game to buy a beer at was the Atlanta Braves, and that the most expensive was the New York Mets. Combining this result with our previous results, we see that the most expensive game to buy a beer at is actually a member of the NL

East. In relation to the Phillies, we see that the Phillies are tied with the Miami Marlins at $6/beer as of 2018, ranking them as tied for third in terms of the most expensive beer in the NL East.