Greg McNeil
EM-622 Final Project
8/18/18

I Pledge my honor that I have abided by the Stevens Honor System.

Table of Contents

Section 1 - Introduction

Every year, as part of the Major League Baseball (MLB) season, teams compete in the World Series of baseball. One of the oldest and most renowned of championship titles, teams have long looked for an edge to determine how to win this once a year tournament. Accordingly, there is a significant amount of easily available data about baseball. This coupled with my innate interest of baseball made it the perfect topic to do my final project of EM-622. The goal of this project was to analyze and visualize different types of baseball data and determine if there is a correlation between these factors and the teams that win world championships. Four different graphs from five datasets were created, each analyzing different aspect of the game of baseball.

The first dataset was to determine if there was a correlation between the amount of money spent per team and the team's chance of winning the World Series. The second graph is an analysis of various offensive and defensive stats compared to league averages. The third graph looks at the total wins of a franchise across the 15 year time period that the data is from and also reviews the amount of playoff wins each team has. The final graph uses a Chernoff Faces to observe the characteristics between each championship winning team's best player. Ultimately several conclusions have been drawn and discussed in the conclusion section of this report.

Section 2 - Data Collection and Preparation

As stated above, there is an enormous amount of datasets about baseball available to dissect. That said, the main two resources used for this case study were "MLB.com" and "baseball-reference.com" which are both cited below in the

“References” section and are both as official as baseball statistics come. The data was downloaded into Microsoft Excel and was left mostly untouched from there. The only time a change was made was when there was a significant error in the data in which it did not make sense to utilize “R” to edit. From the Excel files, all the datasets were saved as “CSVs” and have been included in the submitted zip file.

Section 3 - Discussion

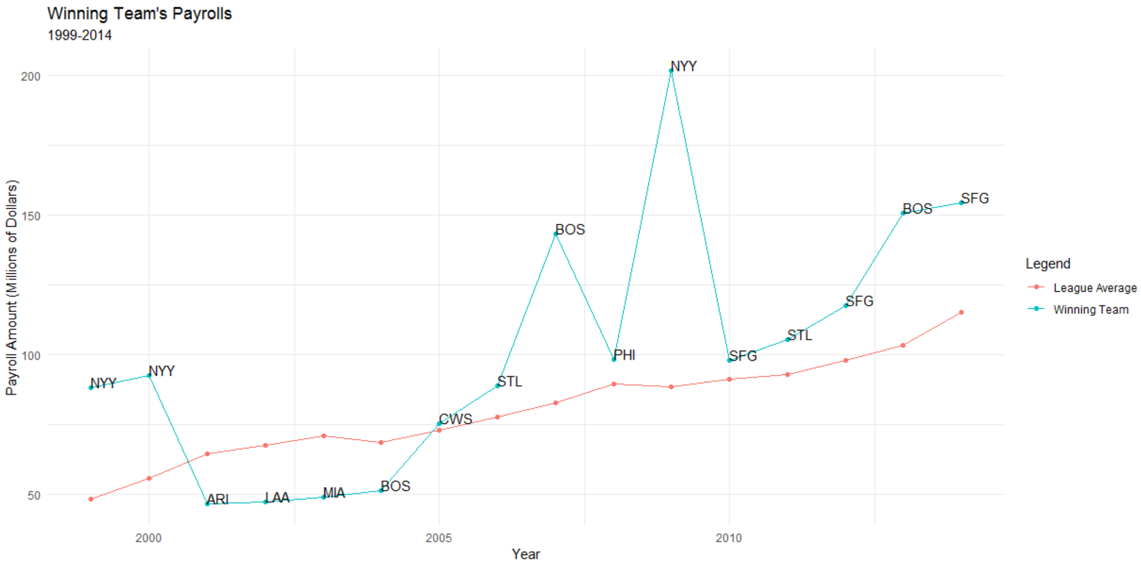

Figure 3.1- Payroll of World Series Winning Teams

The first graph that was completed was *Figure 3.1* which features the total spending of the championship winning team against the league average. In this project, there were 15 teams analyzed over 15 years. This chart shows something immediately interesting in that there are four teams in a four year span that spent under the leagues average on payroll to win world championships. These teams were the 2001 Arizona Diamondbacks, the 2002 Los Angeles Angels, the 2003 Miami Marlins, and the 2004

Boston Red Sox. This can mostly be attributed to the introduction of sabermetrics, which is the study of baseball analytics. Data science in baseball was becoming more prevalent in these years and the few teams that adopted statistics minded analytic methodology first were these four on consecutive years. Put simply, these four teams were famous for finding extremely talented but inexpensive players. In 2005 to 2014, it can be surmised that all teams received access to these data analytics which negated the advantage only several teams held. Another important factor to note is to observe the league average and how much it has increased over the past 15 years. In 1999, the league average was under 50 million dollars. In 2015, the average had more than doubled to around 125 million dollars. Thus, it can be concluded that in the modern day, money plays a factor into winning the World Series.

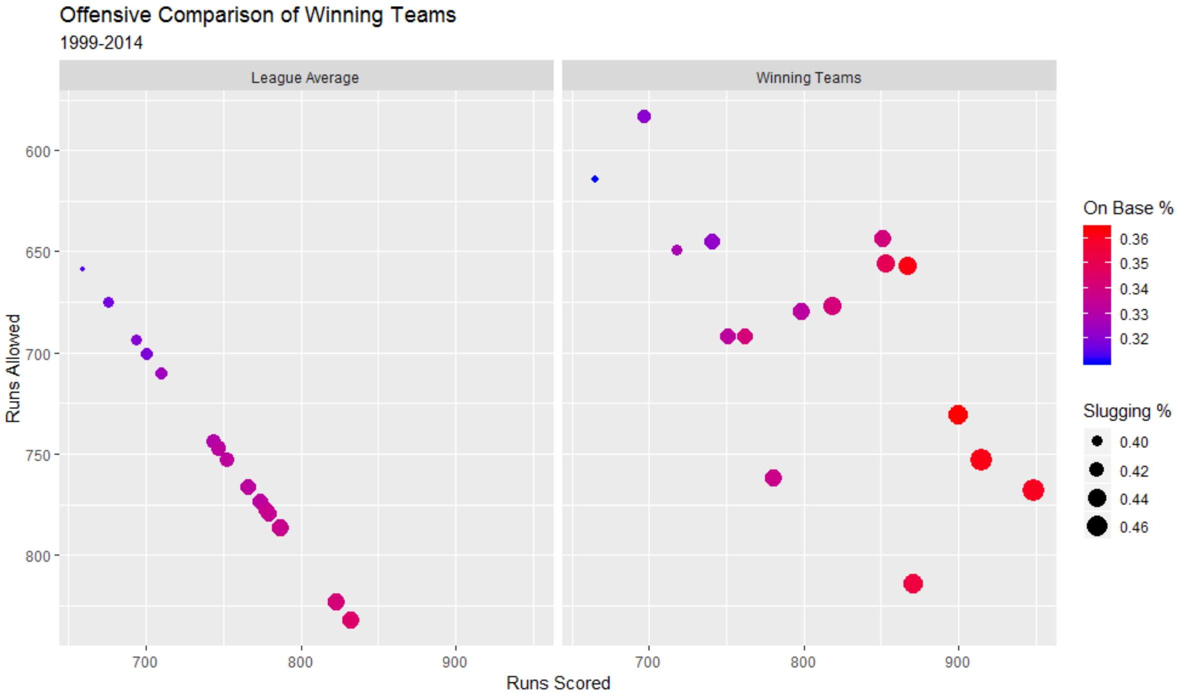

Figure 3.2 - Offensive Comparison of World Series winning teams

In the second graph, shown in *Figure 3.2*, several offensive statistics are plotted along with the league average of each offensive statistic. The four statistics that I chose to focus on were “Runs Allowed”, “Runs Scored”, On Base Percentage” and “Slugging Percentage. This was made in an effort to determine the characteristics of a winning team. On Base Percentage is a good metric for determining a hitter’s talent for not only hitting the ball, but for drawing walks, an equally important skill. On the contrary, Slugging Percentage looks at the total bases a hitter has. For example, when a hitter hits a home run, that would be considered four total bases which is worth more than a single which would only be one total base. It is interesting to compare On Base Percentage and Slugging Percentage because in On Base Percentage, a single and a home run are considered the same but in Slugging Percentage, they are considered vastly different. Runs Allowed is the primary defensive statistic that teams are judged on, which makes it very important. On the contrary, Runs Scored are the primary offensive statistic teams are judged on.

Looking at the graph above provides interesting insights into each statistic. The most interesting fact is that each World Series winning team had well above the league average in Runs Scored and Runs Allowed. From the perspective of On Base Percentage and Slugging Percentage, the championship team generally had above the league average in both of these categories but not exclusively. For example, the smallest point in the top left corner has below the league average in both of these categories. However, it also has the least amount of Runs Scored and Runs Allowed which indicates a highly defensive and pitching focused team. Overall, it can be surmised that World Series winning teams generally score more runs and allow less runs than the league average but how they do it often varies.

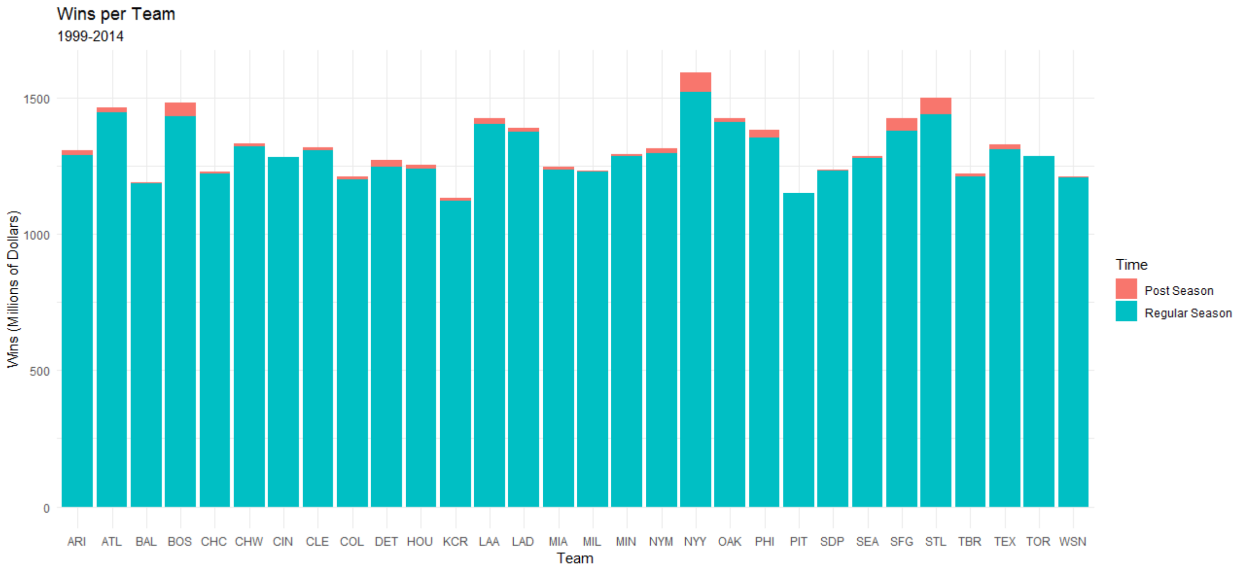

Figure 3.3 - Total Wins per Team

In the above third figure, the wins per team across the 15 year dataset can be shown. In addition, in red the postseason wins are displayed. This graph is useful in showing how wins can be proportional to postseason wins. For example, the Yankees have both the most wins and most postseason wins. That said, it is also useful for showing that there is not necessarily a connection between wins and postseason wins. For example, the Chicago White Sox have very few postseason wins but have many regular season wins.

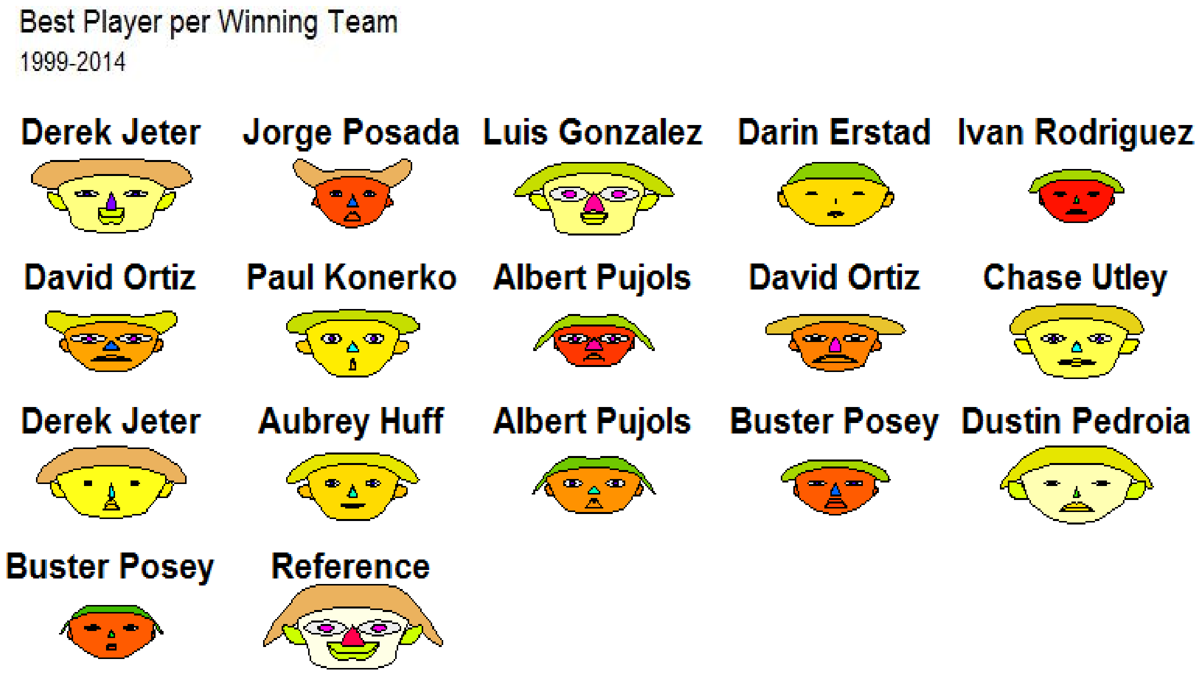

Figure 3.4- Chernoff Faces of each World Series Team (By WAR)

The final graph simply shows the best player on each team and a Chernoff Face to represent that player. The player is considered, by baseball-reference.com, to be the best player on each team using a statistic known as WAR or Wins Above Replacement. Also included in the player list is a reference in the bottom row who uses the maximum of each stat from each player to generate his features. Therefore, it should be noted that the more a player looks like the reference, the better that player is. Below in Figure 3.5 I have included a list of what each factor means.

Modified Item	Variable	Unabbreviated
height of face	G	Games Played
width of face	AB	At Bats

structure of face	R	Runs Scored
height of mouth	H	Hits
width of mouth	2B	Doubles
smiling	3B	Triples
height of eyes	HR	Home Runs
width of eyes	RBI	Runs Batted In
height of hair	SB	Stolen Bases
width of hair	BB	Base on Balls
style of hair	SO	Strikeouts
height of nose	OBP	On Base Percentage
width of nose	SLG	Slugging Percentage
width of ear	G	Games Played
height of ear	AB	At Bats

Table 3.1 - Feature Key for Chernoff Faces Plot

Section 4 - Conclusion

In conclusion, the graphs above helped the end users, who could be someone from a baseball novice to a baseball executive, to better determine how to build their World Series team. To recap the conclusions, it was shown in graph one that teams who spend more money generally have a better chance at winning the world series. There have only been four teams or 27% of teams that have spent below the league average and have won the World Series. In the second graph, it was shown that in order to be a World Series winning team, it does not matter how you score runs, whether it be with home runs or singles but a team must have more runs scored than the league average. A team must also have less runs allowed than the league average. In the third graph, it was shown that wins do not necessarily mean teams perform well in the postseason and that all the wins in the world do not mean anything when a team cannot win the world series. Finally, it was shown that the best player on each team can be a power hitter, good runner, or contact hitter but the team often has a clear best hitter. Overall, there are many factors in baseball that contribute to how well a team performs and all the statistics in the world do not necessarily point to the winning team every team.

References

“MLB Stats, Scores, History, & Records.” Baseball-Reference.com, www.baseball-reference.com/.

“Sortable Player Stats.” Major League Baseball, mlb.mlb.com/stats/sortable.jsp.