Tucker Atwood

WGU MSDA

D214 Task 2

6/14/24

**A1. Research Question**

Using data from the Lahman Baseball Database, I will answer the research question, **“Are there significant statistical differences between World Series-winning Major League Baseball teams and other playoff teams?”**

**A2. Justification and Context**

It is no secret that successful Major League Baseball (MLB) teams are built on some combination of quality **hitting, pitching, fielding, running, strategy, and other valuable skills**. With a successful group of players, teams can often put together a squad capable of consistently making it to the playoffs year after year. However, a common adage suggests that once the playoffs begin, the results are merely a “crapshoot.” The built-in randomness of baseball, combined with series lengths that may not be sufficient in clearly determining the “better” team, has led to **one shocking underdog victory after another**.

Owners and general managers have varied their means of building teams with several different types of strategies: power hitting, small ball, dominating pitchers, tough-to-crack defenses, and so on. Throughout baseball history, debates have raged over **what exactly is the best approach** to assembling a championship-caliber team. **Hundreds of millions of dollars are spent by every team every season** in never-ending attempts to bring home the World Series title. With players, coaches, executives, and endless fans’ wellbeings on the line, **there is significant interest** in determining whether there is a real difference between teams who hoist the championship trophy at the end of October and teams who are left sulking in their dugouts.

**A3. Hypothesis**

I will analyze baseball team factors to determine if some factors describe playoff success, specifically winning the World Series, more than others. I will compare year-adjusted statistics of World Series-winning teams with other playoff teams and will utilize two-sample hypothesis tests to determine key differences between champions and non-champions.

My **null hypothesis** is “There is no significant difference, in any factors, between Major League Baseball teams that win the World Series and other playoff teams who do not.”

My **alternative hypothesis** is “There is a significant difference in at least one factor between Major League Baseball teams that win the World Series and other playoff teams who do not.”

**B1. Data Collection**

The **Lahman Baseball Database** from <seanlahman.com> will be utilized to conduct this analysis.

This database is described as containing “complete batting and pitching statistics from 1871 to 2023, plus fielding statistics, standings, team stats, managerial records, post-season data, and more.” It was created by Sean Lahman, who released his first version of the database in 1994 and has since been a leading figure in ensuring that baseball statistics are free and available to the general public. It is regarded as “the largest and most accurate source for baseball statistics available anywhere.”

Quotations and information from this section, along with the database itself, have been retrieved from [seanlahman.com](file:///C:\Users\atwoo\OneDrive\Desktop\MSDA%20PAs\seanlahman.com).

**B2. Data Collection Advantage**

An **advantage** of using this data to answer the research question is that most team statistics for each individual season are included in an **accurate, straightforward, easy-to-access** CSV file titled “Teams.” Even for statistics that have not been included in the Teams file, several more relevant factors to determining team success are available at an incredibly detailed level for all players who have ever played in an MLB game. These statistics are included in CSV files titled “Batting,” “Pitching,” and “Fielding.”

All files and their data are **clearly marked and organized well**, allowing this analysis to combine the Teams file with the separate player files mentioned to easily aggregate all relevant statistics by team and year. All data on playoff series are also included in the Teams CSV file, allowing a practical comparison between teams that won the World Series and other playoff teams that did not.

**B3. Data Collection Disadvantage**

A **disadvantage** of using this data to answer the research question is that there were **significant variations in MLB rules** in the 19th century and early 20th century that would make comparisons between early baseball and today’s game somewhat problematic. However, the topic of this analysis, the World Series, was not incorporated until 1903, by which time many of the most impactful rule changes had already taken place.

**Changes in game strategies throughout history** also reflect the issue of difficult comparisons between eras; for example, there are many more strikeouts and home runs in recent years than there were in the earlier version of baseball. However, this will be accounted for by adjusting factors by year; each team statistic to be used in this analysis will be presented on a scale comparing its value to the league’s average for that particular season.

**B4. Data Collection Challenge**

A **challenge** in this data collection is that some statistics, such as sacrifice flies and hit-by-pitches, were not properly tracked at earlier points in MLB history and show up as missing data in the Teams file. However, these missing statistics do not play a large role in calculating and comparing the relevant factors to team success that will be used in this analysis. **Imputing the overall averages** of these statistics from years where the data is available will be an acceptable method of dealing with the missing data. Transforming all factors into year-adjusted values will also help in ensuring the missing data does not have a significant impact on the analysis and its results.

**C1. Data Extraction and Preparation**

The language **R** in the **R Studio** environment will be used to conduct this analysis.

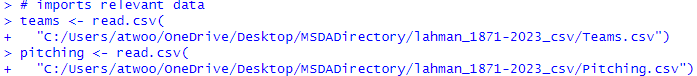
To **extract and prepare the data** for analysis, the following steps are performed as follows:

* The necessary **packages** are loaded using the library() function.

A close-up of a white background

Description automatically generated

* The necessary **data sets** are imported into the R script using the read.csv() function.



* An initial look at the **teams** and **pitching data frames** is provided using the head() function.

A screenshot of a computer program

Description automatically generated

* Because some statistics to be used in the analysis were not included in the initial Teams CSV file, they are **retrieved from the Pitching CSV file**. This file includes all pitching data from all individual pitchers, so in order to express this data on a team-based level, the following steps are taken:
  + All pitching statistics are **grouped by team and year** using the group\_by() function.
  + The required relevant statistics are **hit-by-pitches (HBP), sacrifice flies (SF), and batters faced (BF)**. These are required for the overall analysis because they are necessary in calculating other statistics described later in this section. A sum of each of these statistics is calculated using the mutate() function. Note: hit-by-pitches is denoted “pHBP” and sacrifice flies is denoted “pSF” in order to differentiate these statistics from batter-based versions of the same factors.
  + The pitching data frame is transformed to provide **only these statistics for each team and each year** using the distinct() function.

A close-up of a white background

Description automatically generated

* + Another look at the **transformed pitching data frame** is provided using the head() function.

A screenshot of a computer code

Description automatically generated

* This transformed pitching data frame is **joined with the teams data frame** using the inner\_join() function.



* The columns to be used for the analysis are **specifically selected** from the teams data frame using the select() function.

A close-up of a math equation

Description automatically generated

* A look at the **teams data frame with relevant columns and the transformed pitching data** joined to it is provided using the head() function.

A table of numbers with letters

Description automatically generated with medium confidence

* **Certain years are filtered out of the data** using the filter() function for the following reasons:
  + The World Series was established in 1903, so **all years prior to 1903** are not relevant to the analysis.
  + **No World Series was held in 1904** due to a disagreement between the leagues, so this year is not relevant to the analysis.
  + **No World Series was held in 1994** due to a players’ strike, so this year is not relevant to the analysis.
  + A World Series was held in 2020; however, the season was shortened to a 60-game schedule due to the COVID pandemic, so statistics from 2020 are not considered indicative of other MLB seasons and **2020 will not be included in the analysis**.



* Some teams with **no value in the “WSWin” column** are found to still exist in the data. Upon further research, these teams belong to a defunct “Federal League” which was only in operation for the years 1914 and 1915. **These teams and their statistics are removed** from the data using the anit\_join() function.

A screenshot of a computer screen

Description automatically generated

A close up of a text

Description automatically generated

* Only teams relevant to calculating accurate league averages remain in the data set. Missing data is now **imputed** using overall averages when necessary and using year averages when possible. This is performed with the mutate(), group\_by(), and ungroup() functions.

A computer screen shot of a computer code

Description automatically generated

* **League average statistics** are calculated for each season using the group\_by(), mutate(), and ungroup() functions, so that year-adjusted statistics may be calculated later. Other relevant statistics are calculated prior to this, as they are used in the ensuing calculations.

A screenshot of a computer program

Description automatically generated

* Relevant statistics to team success are **calculated and transformed** so that they are **expressed in relation to that season’s league averages**. This is performed using the mutate() function. A further breakdown of these calculations is provided:
  + The given statistic is **divided by the league average, then multiplied by 100**.
  + Multiplying by 100 is a common method for expressing adjusted baseball statistics; this provides an easy comparison because values **above 100** show that the team performed **better than league average**, while values **below 100** show that the team performed **worse than league average**.
* Once the calculations are performed, only the **relevant columns are selected** using the select() function, and **sorted descending by year** using the arrange() function.

A screenshot of a computer program

Description automatically generated

* Now that league averages and year-adjusted statistics for each team have been calculated, a **new data frame containing only playoff teams** is created using the filter() and select() functions. A look at the resulting data frame is provided using the head() function.

A screenshot of a computer

Description automatically generated

* This data frame is separated into two other new data frames, **World Series winners** and **other playoff teams who did not win the World Series,** using the filter() function. A look at each of these resulting data frames is provided using the head() function.

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

* These are the **final, cleaned data sets** to be used in the analysis. The summary statistics for each data set is provided using the summary() function.

A screenshot of a computer screen

Description automatically generated

A **description of all statistics** included in the data sets to be used for analysis is provided below. For all values except Year, Team, and WSWin, they are **expressed in relation to yearly league averages**, where values above 100 imply the team’s original statistic was higher than league average, and values below 100 imply the team’s original statistic was lower than league average. For most statistics, a higher value implies better team performance; however, in the following statistics, a **lower value is preferable**: **Kr, RAA, WHIP, pBBr, and pHRr**. Also note that this data represents **regular season statistics** **only**, which means any playoff statistics that could inherently tip the scales toward World Series winners are not included.

* **Year**: the year for which the given statistics occurred.
* **Team**: the team for which the given statistics were accrued for.
* **WSWin**: whether or not the given team won the World Series that season.
* **AVG**: batting average, which measures how often a player gets a hit as a percentage of his at-bats. This is calculated by the following formula, in which H represents the number of **hits** a batter achieved and AB represent the number of **at-bats** the batter accumulated.
* **OBP**: on-base percentage, which measures how often a player reaches base safely as a percentage of his plate appearances. This is calculated by the following formula, in which BB represents the number of **walks** the batter drew, HBP represents the number of **hit-by-pitches** the batter drew, and PA represents the number of **plate appearances** the batter accumulated. Plate appearances include every time a batter goes to the plate, in contrast to at-bats, which do not include walks, hit-by-pitches, and sacrifice hits.
* **SLG**: slugging percentage, which measures the average number of total bases a player achieves per at-bat. In other words, this is a version of batting average that gives the appropriate weights to singles, doubles, triples, and home runs. This is calculated by the following formula, in which 1B represents the number of **singles** a batter achieved, 2B represents the number of **doubles** a batter achieved, 3B represents the number of **triples** a batter achieved, and HR represents the number of **home runs** a batter achieved.
* **OPS**: on-base plus slugging percentage, which simply adds the previous values of OBP and SLG. This typically encapsulates batters’ contributions to team success better than OBP or SLG can do individually. This is calculated by the following formula.
* **RPG**: runs per game, which measures the average number of runs a team has scored per game. This is the most basic way to describe offensive team success. This is calculated using the following formula, in which R represents the number of **runs** a team has scored and G represents the number of **games** a team has played.
* **BBr**: walk rate, which measures the percentage of plate appearances in which a player draws a walk.
  + Note: the denotation “BB” comes from the phrase “Base on Balls,” meaning a player has reached base by means of 4 balls (non-strikes) thrown. This is calculated using the following formula.
* **Kr**: strikeout rate, which measures the percentage of plate appearances in which a player strikes out.
  + Note: when these statistics were first developed in the 1860’s, sacrifices were more common than strikeouts, so “S” was set aside to denote sacrifices. This left “K” as the denotation for strikeout, as it was the last letter of the more common term at the time: “struck.”
  + In the Lahman database, strikeouts are abbreviated as SO; however, the denotation used in this analysis will be K to use the most common baseball terminology. Kr is calculated using the following formula, in which SO represents **strikeouts**.
* **HRr**: home run rate, which measures the percentage of plate appearances in which a player hits a home run. This is calculated using the following formula.
* **SBg**: stolen bases per game, which measures the average number of bases a player has stolen per game. This is calculated using the following formula, in which SB represents the number of **stolen bases** a player has achieved.
* **RAA**: runs allowed average, which measures the average number of runs a team has given up to its opposing teams per 9 innings.
  + Note: this statistic is being used instead of the more commonly used Earned Run Average (ERA) because this analysis aims to focus on team performance over individual player performance. An individual pitcher may be charged with an earned or unearned run, depending on whether his defense committed an error which allowed a run to score that would not have scored otherwise. ERA provides an arguably better indicator of individual pitcher success, but RAA provides more information about team success. Both are expressed as an average of “per 9 innings” to represent the length of a non-extra inning regulation baseball game.
  + The Lahman data lists “IPouts,” the total number of outs a pitcher has achieved, rather than innings pitched (IP) as a result of partial innings completed. Since it takes 3 outs to complete an inning, partial innings are often denoted with decimal values 0.1 and 0.2 to represent 1 out and 2 outs, respectively. This is problematic when dealing with data; for example, 0.1 IP actually represents 1/3 of an inning, rather than 1/10. Because IPouts, the number of outs recorded by the pitcher, is provided in this data instead, the RAA formula will multiply by 27 in contrast to the typical formula which multiplies the total innings by 9.
  + RAA is calculated using the following formula, in which RA represents **runs allowed** and IPouts represents **outs** recorded by the pitcher.
* **WHIP**: walks and hits per inning pitched, which measures the average number of walks and hits a pitcher allows per inning he pitches. This is sometimes a better indicator of pitcher success than ERA (and RAA) because it breaks down the number of baserunners a pitcher allows into individual events, whereas runs may be the result of an unlucky accumulation of events.
  + Note: hit-by-pitches are not included in this formula because the creator of the formula simply forgot to include them. Despite this flaw, the statistic will be used without hit-by-pitches for ease of accessibility to its most common usage. WHIP is calculated using the following formula, where BBA represents **walks allowed**.
* **pBBr:** pitcher walk rate, which measures the percentage of batters faced that a pitcher has walked. This is the pitcher version of BBr and has been preceded by “p” to distinguish the pitcher version from the batter version. This is calculated using the following formula, where BF represents **batters faced**, the pitcher equivalent of plate appearances.
* **pKr:** pitcher strikeout rate, which measures the percentage of batters faced that a pitcher has struck out. This is the pitcher version of Kr. This is calculated using the following formula, where SOA represents **strikeouts allowed**, which is not a common baseball term, but is useful in differentiating this statistic from the batter equivalent SO.
* **pHRr**: pitcher home run rate, which measures the percentage of batters faced in which a pitcher has allowed a home run. This is the pitcher version of HRr. This is calculated using the following formula, where HRA represents **home runs allowed**.
* **DEF**: defensive efficiency, which measures the percentage of hit balls in play that a defense converts into outs. Non-pitcher defensive ability is more difficult to measure than other aspects of baseball. Other statistics would be used to attempt to measure an individual player’s defensive ability, but DEF is the best measure of team defensive ability. It is essentially the inverse of an opponent’s batting average with home runs and strikeouts taken out of the equation, as these are not “balls in play” in which the defense can have an effect on the outcome. DEF is calculated using the following formula.

**C2. Data Preparation Advantage**

An **advantage** of the tools and techniques used to extract and prepare the data is that R’s accessible data manipulation packages and functions allowed the raw data to be completely transformed from its basic initial state into a collection of commonly used statistics adjusted for yearly league averages. The original data would not have provided as much insight in the ensuing analysis if not for the tools and techniques available in R.

**C3. Data Preparation Disadvantage**

The tools and techniques used in section C1 were successfully used to extract and prepare the data in the exact manner as intended for this analysis. If a **disadvantage** must be chosen, it is that there was some missing data in the original data set that needed to be replaced via imputation.

**D1. Exploratory Data Analysis Visualizations**

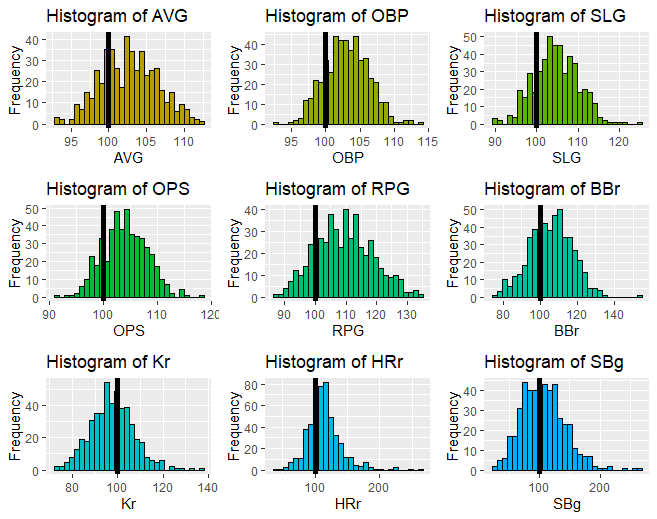
The following **visualizations** provide necessary insight to help understand the structure and spread of the remaining statistics for all playoff teams in the cleaned and prepared data set.

* **Correlation heatmap plots** are generated to show the connections between the various statistics to each other. A positive correlation indicates that, as one factor increases, the other also increases. A negative correlation indicates that, as one factor increases, the other decreases. Values closer to 1 or -1 indicate stronger correlations.
* **Correlation heatmaps, separated into offensive and defensive statistics:**

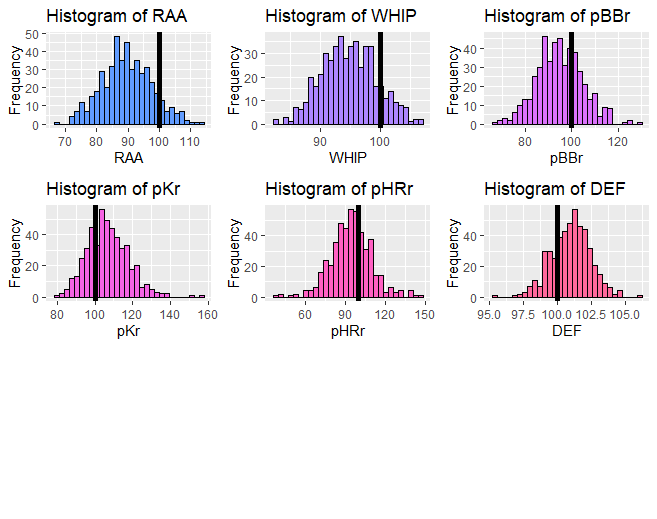
A close-up of a graph

Description automatically generated

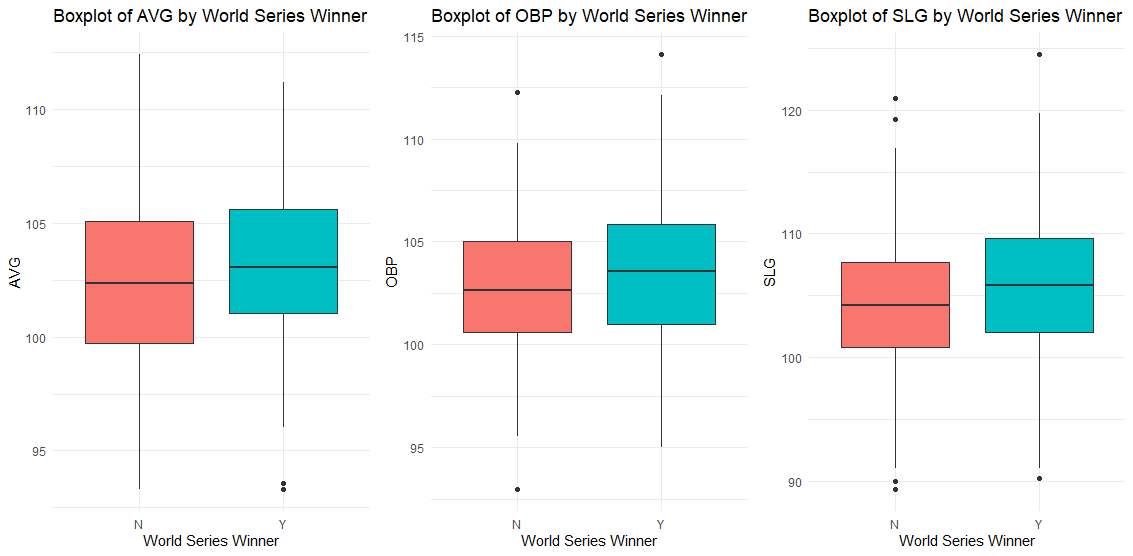
* + These heatmaps confirm many connections between statistics that are commonly understood and/or inherent to the way the statistics are calculated, such as the presence of OBP and SLG within the OPS formula.
  + Some correlations are slightly unexpected or interesting, such as the weak to zero (-0.04) correlation between batter strikeouts and slugging percentage. There is a common thought that “sluggers” have higher strikeout rates than average, since they often sacrifice contact rate for power. There is a moderate (0.32) correlation between strikeout rate and home run rate, which supports this theory, so the much lower correlation between slugging percentage and strikeout rate must come from hitters who have the power to hit doubles and triples, but not as many home runs.
  + There is also a surprisingly low (0.07) correlation between pitcher strikeout rate and pitcher walk rate. It is commonly thought that pitchers who strike out many batters do so with a slight sacrifice in their control and/or willingness to throw in the middle of the strike zone. In other words, they risk walking batters in order to get them to swing at pitches that wouldn’t otherwise be strikes. The low correlation between the two suggests there is less of a connection to pitchers’ strikeouts and walks as is commonly believed.
* **Univariate histograms** are generated for all included statistics to show the spread of the factors for all playoff teams.
  + **Offensive statistics histograms:**

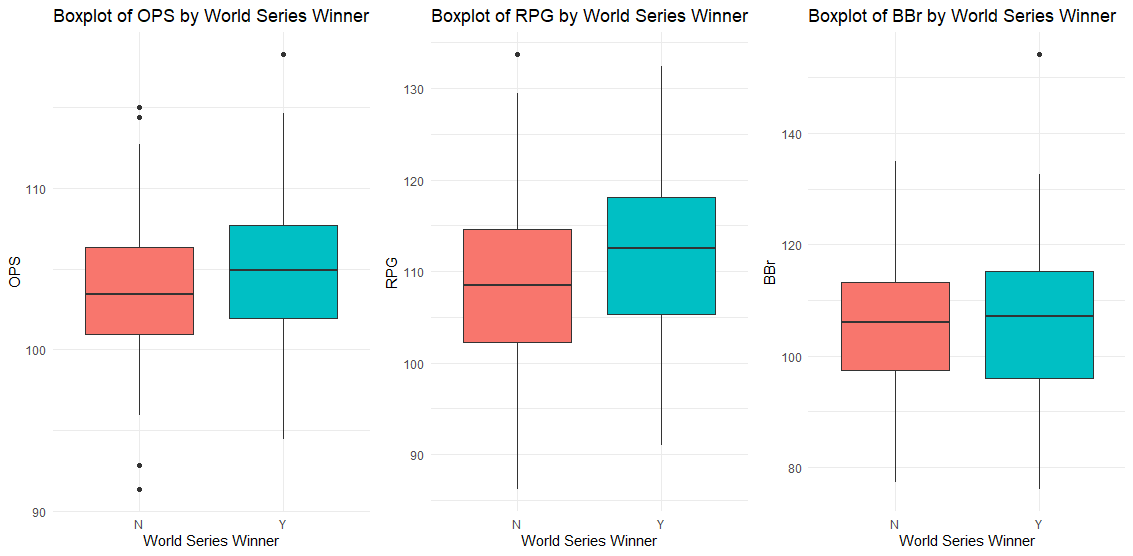
****

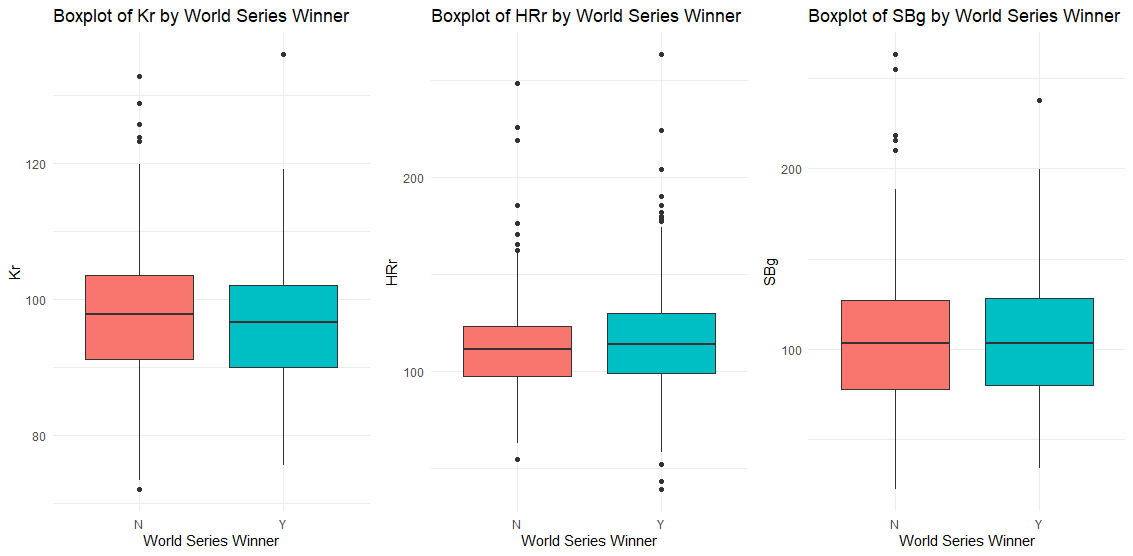
* + **Defensive statistics histograms:**

****

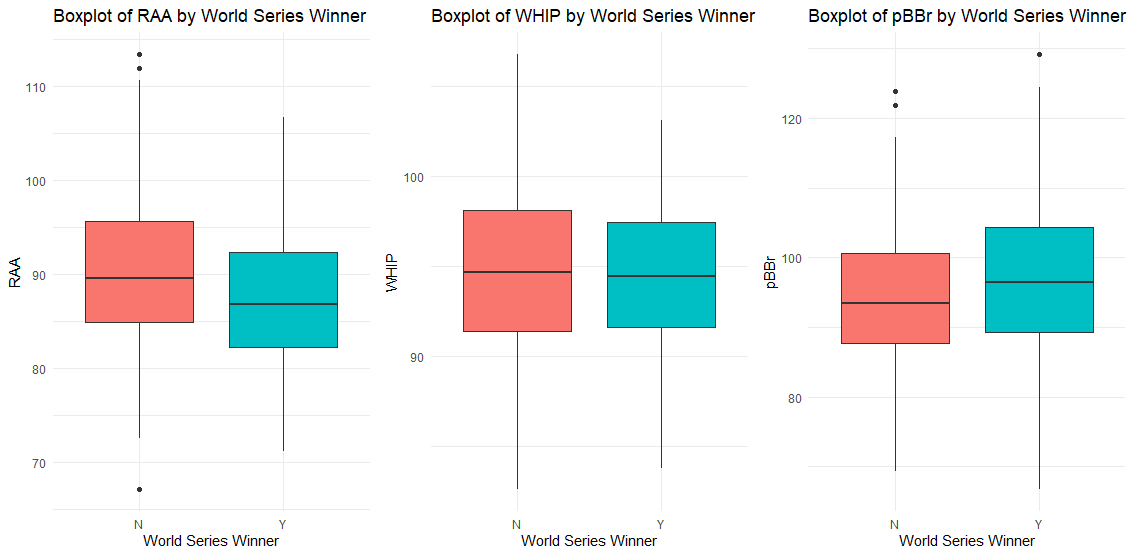
* + A **vertical line at 100** has been placed on each of the histograms to indicate where “league average” exists in relation to the spread of the data. In many cases, this confirms that playoff teams are generally better than league average. For example, in the histogram of OBP, it is clear that many more values are above the league average than below. Reminder: it is preferable for the values Kr, RAA, WHIP, pKr, and pHRr to be lower than league average.
  + Interestingly, the spread of stolen bases per game suggests there have been about as many teams in the playoffs who are below-average in stolen base rate as there have been above-average. Though many teams throughout history have employed specific strategies to boost their stolen base totals, there have also been plenty who believe stealing bases is not worth the risk.
* **Side-by-side boxplots** are generated for all included statistics, separated by the WSWin factor, to show the differences in each factor for World Series winning teams and other playoff teams.
  + **Offensive statistics boxplots:**

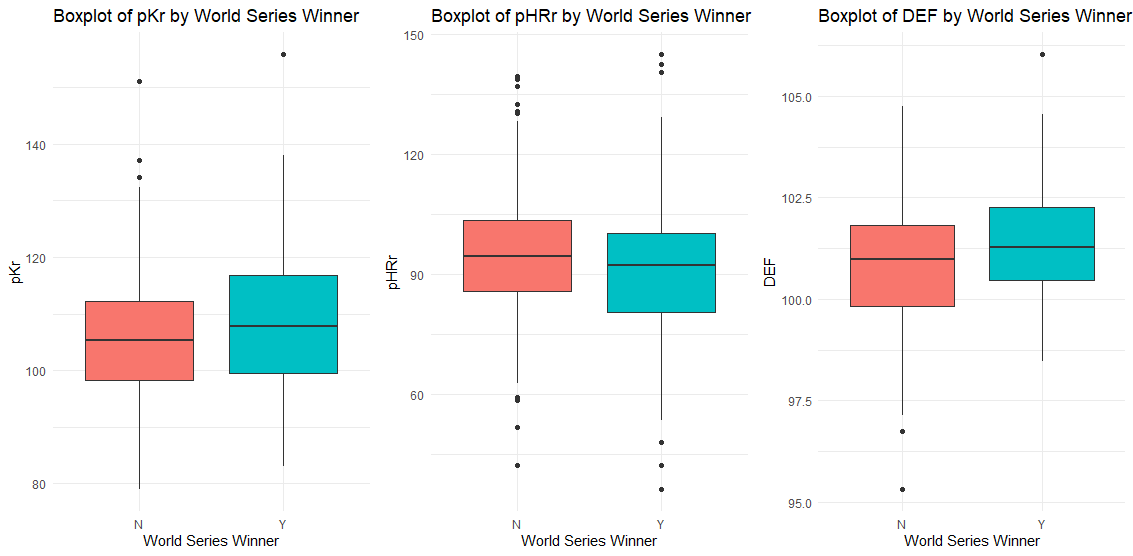
****

****



* **Defensive statistics boxplots:**

****

****

* + These boxplots show the minimum, first quartile, median, third quartile, and maximum values for each factor, separated by playoff teams that won the World Series and playoff teams that did not. Once again, it is clear that in most of these factors, teams in both categories perform better than league average, with the median values consistently above 100 for stats where it is preferable and below 100 for stats where it is not.
  + In many of these boxplots, a slight advantage may be seen for teams who have won the World Series. This is interesting to note, as it may hint that there is a true difference between champions and non-champions. Reminder: this data includes regular season statistics only. While it is obvious teams that win in the playoffs would have better playoff statistics than their counterparts, it is not obvious that they would have better statistics prior to the playoffs.

**D2. Data Analysis Technique**

To answer the research question, a **series of two-sample one-tailed hypothesis tests** will be run on the data to determine if there are statistically significant differences between the means of all mentioned factors for the World Series winners and all other playoff teams. The following null and alternative hypotheses represent an example of the hypotheses for all two-sample tests to be run for all factors (only the factor name changes).

* **Null hypothesis:** The average team slugging percentage for MLB teams that win the World Series is not significantly greater than the average team slugging percentage for other playoff teams that do not win the World Series.
* **Alternative hypothesis**: The average team slugging percentage for MLB teams that win the World Series is significantly greater than the average team slugging percentage for other playoff teams that do not win the World Series.
* The **alpha level** for each hypothesis test will be **0.05**. All p-values less than this value will be considered statistically significant.

Note: since one-tailed tests will be used, the tests will be run with a **null hypothesis of** and an **alternative hypothesis of** , where represents the mean of the given statistic for World Series-winning teams and represents the mean of the given statistic for other playoff teams. Since there are some factors for which it is **preferable to have a lower value**, the calculated **p-values for these hypothesis tests will be subtracted from 1** to adjust for the direction considered to be a “better” performance. The relevant statistics with p-values to be adjusted are Kr, RAA, WHIP, pBBr, and pHRr.

**D3. Technique Justification and Advantage**

**The decision to use two-sample one-tailed hypothesis tests for this analysis was made because** they can be used to identify a significantly higher mean for one group over another. The groups to be compared here are teams who win the World Series and playoff teams who do not.

Though many statistics included in the cleaned data set have long been known as clear indicators of teams’ success in making it to the playoffs, there is significant interest in the baseball community to determine whether the ultimate winner of the playoff tournament is largely a matter of luck, or if there are clear indications that some statistics may determine playoff success over others.

An **advantage** of using hypothesis tests for this analysis is that they will help explain whether a World Series-winning team is significantly more successful than another playoff team in any of the relevant factors, especially if the results show more or less importance in certain areas than is commonly believed. For example, a hypothesis test will provide insight into whether or not teams that have won the World Series have significantly higher stolen base rates than other playoff teams. If so, this will be an interesting discovery, as baserunning ability is commonly believed to be less important than batting and pitching ability.

**D4. Technique Disadvantage**

A **disadvantage** of using hypothesis tests for this analysis is that predictions of future team success cannot necessarily be made from the results of a hypothesis test. Even if the tests show statistically significant results, they are merely describing past events, and with many minor changes being made to baseball rules and strategies every season, even significant results from this analysis cannot be used as predictors of future outcomes.

**D5. Calculations and Outputs**

The input and output results of the two-sample one-tailed hypothesis tests are provided as follows. These results and their implications will be discussed in section E1.

A screenshot of a computer program

Description automatically generated

**E1. Results and Implications**

With a designated alpha level of 0.05 for each two-sample one-tailed hypothesis test, there are **statistically significant results for the factors RAA, RPG, DEF, SLG, OPS, AVG, pKr, OBP, HRr, and pHRr**. A brief discussion of the implications of each test, with significant results or not, is provided, listed in order of p-value (lowest to highest).

* **Runs Allowed Average** had a p-value of approximately 0.0001. Since this was one of the factors for which a lower value was preferable, the initial p-value was subtracted from 1 to represent the inverse of the original alternative hypothesis example. So, based on these results, **the mean RAA for World Series-winning teams is significantly lower than the mean RAA for other playoff teams**. This is not surprising, as RAA is the most basic measure of pitching/defensive success included in this data, and confirms that limiting opponent runs during the regular season leads to more playoff success.
* **Runs Per Game** also had a p-value of approximately 0.0001. Based on these results, **the mean RPG for World Series-winning teams is significantly higher than the mean RPG for other playoff teams**. This is also not surprising, as RPG is the most basic measure of offensive success, and confirms that scoring runs during the regular season leads to more playoff success.
* **Defensive Efficiency** had a p-value of approximately 0.0003. Based on these results, **the mean DEF for World Series-winning teams is significantly higher than the mean DEF for other playoff teams**. Though it may not be surprising that good defense is important in winning championships, its placement as the third-most significant result in this analysis is noteworthy, as there have been several teams throughout baseball history to sacrifice defensive value for offensive value. Of course, the results of the RPG hypothesis test were slightly more significant than this one, but it may be surprising to some that defensive performance is just about as important.
* **Slugging Percentage** had a p-value of approximately 0.003. Based on these results, the **mean SLG** **for World Series-winning teams is significantly higher than the mean SLG for other playoff teams**. It makes sense teams that hit for power in the regular season are successful in the playoffs, but these results are noteworthy due to the ongoing debate about which offensive attribute is more important: power-hitting or on-base ability. Though many players are able to do both, it is common for certain player types to sacrifice power hitting for contact hitting and patience, or vice versa. Since these results are more significant than the results for the AVG and OBP, this may be an indication that power hitting is slightly more beneficial to a team than other offensive strategies.
* **On-base Plus Slugging Percentage** also had a p-value of approximately 0.003. Based on these results, **the mean OPS** **for World Series-winning teams is significantly higher than the mean OPS for other playoff teams**. Since SLG is one-half of the formula for OPS, these results are not particularly surprising; however, it is commonly believed that OPS provides more information about a team’s and player’s success than SLG and OBP do on their own. The significance of this result is about the same as the significance for the SLG result, but it may have been expected that this should have been more significant than SLG. This may indicate that OBP is less important than commonly believed.
* **Batting Average** had a p-value of approximately 0.010. Based on these results, **the mean AVG for World Series-winning teams is significantly higher than the mean AVG for other playoff teams**. Batting average is perhaps the oldest, most traditional measure of batter performance in baseball, so it is unsurprising and comforting that regular season batting average may be an indicator of playoff success. In addition to the point made in the previous paragraph relating to SLG, this also may indicate that OBP is not as important as commonly thought. OBP arose as an alternative to AVG and is thought to provide more information than AVG, since it takes into account all ways for batter to reach base, not just hits. However, getting hits moves baserunners around in a more productive manner, and this result may indicate AVG has more of an argument against OBP than considered.
* **Pitcher Strikeout Rate** had a p-value of approximately 0.015. Based on these results, **the mean pKr for World Series-winning teams is significantly higher than the mean pKr for other playoff teams**. Besides RAA, this is the most significant pitcher-specific test result, and that may not be surprising. Since pitchers are often unable to control whether or not a batter’s contacted ball falls into play for a hit or not, it is commonly believed that pitcher strikeout rates are important to success because it indicates the pitcher has taken matters into his own hands and prevented his defense from even having the opportunity to fail him. Having pitchers who keep batters from putting the ball in play during the regular season likely leads to less balls in play during the playoffs as well.
* **On-base Percentage** had a p-value of approximately 0.018. Based on these results, **the mean OBP for World Series-winning teams is significantly higher than the mean OBP for other playoff teams**. While this is unsurprising and confirms the obvious that batters who avoid making outs are preferable to their counterparts, the above arguments relating to the SLG, OPS, and AVG hypothesis tests may indicate that OBP is not quite as important as some believe. This means that batter types who are patient, sacrificing good pitches to hit in order to try drawing a walk, may be sacrificing too much. It may benefit them to try to make more contact and/or try to hit for more power when they do make contact.
* **Home Run Rate** had a p-value of approximately 0.033. Based on these results, **the mean HRr for World Series-winning teams is significantly higher than the mean HRr for other playoff teams**. Hitting a home run is perhaps the most exciting play in baseball, and many players try their best to hit as many home runs as possible. Since a home run counts for at least one automatic run (and up to 3 others) without giving the defense an opportunity to make a play on the ball, it is unsurprising that teams with more home runs in the regular season fare better in the playoffs than their counterparts. However, it is interesting that this result is less significant than the SLG hypothesis test, since home run hitters typically have high SLG values. The disparity in these results may indicate that players who sacrifice too much contact for power may be doing themselves a disservice.
* **Pitcher Home Run Rate** had a p-value of approximately 0.0484. Since this was one of the factors for which a lower value was preferable, the initial p-value was subtracted from 1 to represent the inverse of the original alternative hypothesis example. Based on these results, **the mean pHRr for World Series-winning teams is significantly lower than the mean pHRr for other playoff teams**. For reasons similar to (but in the other direction from) those listed in the previous paragraph, it is unsurprising that allowing less home runs in the regular season indicates more team success in the playoffs. When a team allows less home runs, it gives its defense more opportunities to make plays, and thus more opportunities to get the opposing team out.
* **Walks plus Hits per Inning Pitched** had a p-value of approximately 0.147. Since this was one of the factors for which a lower value was preferable, the initial p-value was subtracted from 1 to represent the inverse of the original alternative hypothesis example. Based on these results, **there is not enough evidence to show that the mean WHIP for World Series-winning teams is significantly lower than the mean WHIP for other playoff teams**. These results are likely related to the surprising results for the pBBr hypothesis test that is discussed below, since walks are included in both formulas.
* **Batter Strikeout Rate** had a p-value of approximately 0.199. Since this was one of the factors for which a lower value was preferable, the initial p-value was subtracted from 1 to represent the inverse of the original alternative hypothesis example. Based on these results, **there is not enough evidence to show that the mean Kr for World Series-winning teams is significantly lower than the mean Kr for other playoff teams**. While it may seem that having batters who strike out often is a bad thing, in recent years there has been more willingness for batters to sacrifice an uptick in strikeout rate in exchange for more power. Striking out also usually removes the possibility of a double play. With a lack of significance in this strikeout rate hypothesis test, the concession of more strikeouts may be a fine decision on batters’ parts.
* **Stolen Bases per Game** had a p-value of approximately 0.213. Based on these results, **there is not enough evidence to show that the mean SBg for World Series-winning teams is significantly higher than the mean SBg for other playoff teams**. This lack of evidence may confirm the contemporary theories that strategies centered around stealing bases and risking outs through baserunning are not worth it. It is also commonly believed that only teams that can steal bases at high success rates should attempt to be aggressive on the basepaths. The lack of significance in this result may suggest this theory is correct.
* **Walk Rate** had a p-value of approximately 0.281. Based on these results, **there is not enough evidence to show that the mean BBr for World Series-winning teams is significantly higher than the mean BBr for other playoff teams**. The lack of evidence here may explain fully why the OBP hypothesis test was less significant than other measures it is often thought to be more descriptive than. As discussed in previous sections, perhaps there is less value in taking a walk than often believed, and more batters should be focusing on making contact and hitting with power.
* **Pitcher Walk Rate** had a p-value of approximately 0.999. Since this was one of the factors for which a lower value was preferable, the initial p-value was subtracted from 1 to represent the inverse of the original alternative hypothesis example. Based on these results, **there is not enough evidence to show that the mean pBBr for World Series-winning teams is significantly lower than the mean pBBr for other playoff teams**. This is perhaps the most surprising result of the analysis. The p-value suggests that there is actually counterintuitive evidence in these results: that teams whose pitchers walk more batters have more success in the playoffs. This is curious, as it is generally understood that putting more batters on base, without even making them hit the ball, is not a good thing for a pitcher to do. The results may initially imply that the correlation between pitcher strikeouts and pitcher walks is to blame: a pitcher who strikes out more batters will also walk more batters, and though walking batters is bad, striking them out is good enough to outweigh the disadvantages. However, the correlation plots in section D1 indicate a weak correlation between pitcher strikeout rates and pitcher walk rates. It may also be the case that pitchers who walk more batters are providing the batters with less pitches in the middle of the strike zone, which in turn leads to less contact and less power from the batters. These results may be the most intriguing to follow up on in future research.

The research question for this analysis was, “Are there significant statistical differences between World Series-winning Major League Baseball teams and other playoff teams?” **Based on the discussed results of these hypothesis tests, there is enough statistical evidence to suggest that World Series-winning MLB teams perform better than other playoff teams in several factors.**

**E2. Analysis Limitation**

A **limitation** of this analysis is that MLB rules and strategies have shifted throughout the course of history, including a gradual expansion in the number of teams included in the playoff bracket from 2 teams in the original playoffs to the current tournament of 12 teams. Since most seasons since 1903 were used for this analysis, some or most of the data used may not be relevant in understanding the contemporary game. Though it is typically better to be able to use as much data as possible to answer a question, it is also important to recognize when data is being used that does not provide meaningful insight into the question at hand. If this analysis was to be repeated, perhaps a separation of data into different eras of baseball would provide different results.

**E3. Recommended Course of Action**

Based on the results of the hypothesis tests conducted in this analysis, some **recommended courses of action** for somebody working within or about Major League Baseball are as follows:

* A further investigation of the difference in public opinion and statistical evidence regarding **batter walk rates** is warranted. It is commonly believed that “a walk is as good as a hit,” a theory that was propagated in the popular book and movie *Moneyball*, which described Oakland Athletics’ General Manager Billy Beane’s strategy to target specific players who were not valued as highly on the free agent market as they should have been. A large focus of the *Moneyball* plot centered around players who were patient and were able to get on base even if that didn’t always mean taking a walk. Due to the results of the OBP and BBr hypothesis tests, further research may need to be conducted to test Beane’s theory. It may be more important for hitters to focus on contact and power than on patience.
* Similarly, **pitcher walk rates** should also be studied further to determine why the results of the pBBr hypothesis test were counterintuitive to popular belief and to general baseball strategy. It does not make sense on the surface that teams whose pitchers walk more batters should have more playoff success, but this is exactly what was implied in the results of this analysis. The theories discussed in section E1 should be tested and researched further.

**E4. Future Study**

Two **directions of future study** for this data set include:

* Specific impacts of **MLB rule and strategy changes**, including the following recent changes to the game:
  + An enforced pitch clock timer. This new rule was instituted to decrease game times. Does this have an effect on any of the pitcher or batter factors studied in this analysis? If so, which group of players does it affect more? Are players who are more capable of speeding up their individual paces now more likely to experience playoff success?
  + Larger bases and limited pitcher pickoffs. These new rules were instituted to increase the number of stolen bases in the game. It was shown in this analysis that stolen bases were not significantly higher for World Series-winning teams than other playoff teams. Will the new rules change this emphasis at all, or will stolen bases simply increase without a corresponding effect on playoff success?
  + Infield shift limitations. This new rule was instituted to increase the number of batted balls that become hits. The results in this analysis implied that walks were less important than commonly believed. Are they going to become more or less important if hits become more frequent?
* The factors studied in this analysis could be further **broken down by time periods**, perhaps by the number of teams allowed into the playoffs. Is playoff success dependent on how many teams must be defeated in order to win the World Series? A team in 2024 must win at least three rounds in the playoffs to be deemed the champion; for the first 65 years of MLB playoffs, there was only a World Series. A team only had to win one round of playoffs to be the champions. Does the increase in playoff rounds change what determines success? Future study could follow the same steps as this analysis, but with the tests broken down by era:
  + 1903-1968: 1 round, 2 playoff teams. Baseball saw many periods of different strategies as teams dealt with the “dead ball” of the early 1900’s, experienced a home run revolution in Babe Ruth’s 1920’s and 30’s, lost many players to military service in the middle of the 1940’s, and began using farm systems which changed the overall development of players.
  + 1969-1993: 2 rounds, 4 playoff teams. The game changed dramatically, with new rules such as free agency and the designated hitter, quick expansion into new markets, and Rickey Henderson-inspired speed merchants shifting the baseball landscape in countless ways -- not to mention the new extra step it suddenly took to win the gold.
  + 1995-2011: 3 rounds, 8 playoff teams. The aftermath of a player’s strike in 1994 left baseball in the heart of the steroid era, where suddenly every hitter was hitting home runs like never before, leading awkwardly into a banning of performance-enhancing drugs which soured offensive totals at the tail end of this second major playoff expansion.
  + 2012-2021: 4 rounds, 10 playoff teams. A Wild Card play-in game that annually left the fates of two teams in each league in the hands of one singular game, a rapid succession of deadening and livening the baseball with the effects of extreme fluctuations in offensive environments, and a worldwide pandemic which proved impossible to plan around were all contributing factors to a strange new game of baseball.
  + 2022-present: 4 rounds, 12 playoff teams. With more than one-third of all teams now making the playoffs, the aforementioned new rules kicking the tires on a game dueling with football and basketball to stay relevant, and analytics thoroughly pronouncing a knockout of several prehistoric baseball strategies, the current game is open for business to anyone who can get their Master’s in Data Analytics.

What’s next for this data set? Just about everything you can imagine.

**F. Sources**

WGU Courseware was used as a resource to learn the methods, concepts, and functions used to create the codes in this project. There is no content in this analysis that has been quoted, paraphrased, summarized, or otherwise requires direct citation.

The data used in this analysis was retrieved from [seanlahman.com](file:///C:\Users\atwoo\OneDrive\Desktop\MSDA%20PAs\seanlahman.com). Sean Lahman owns this database. The work is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License.