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.

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

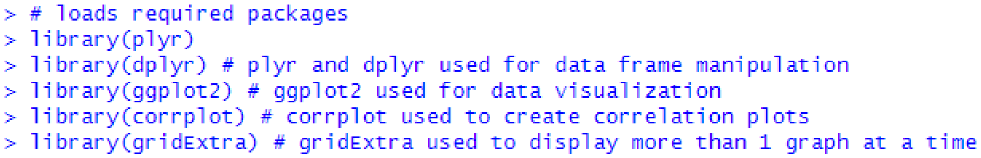
A **challenge** in this data collection is that some statistics, such as sacrifice flies and hit-by-pitches, were not properly tracked at earlier points in MLB history and show up as missing data in the Teams file. However, these missing statistics do not play a large role in calculating and comparing 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.

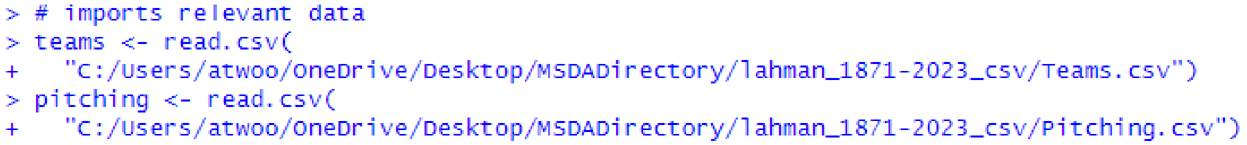
To **extract and prepare the data** for analysis, the following steps are performed as follows:

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



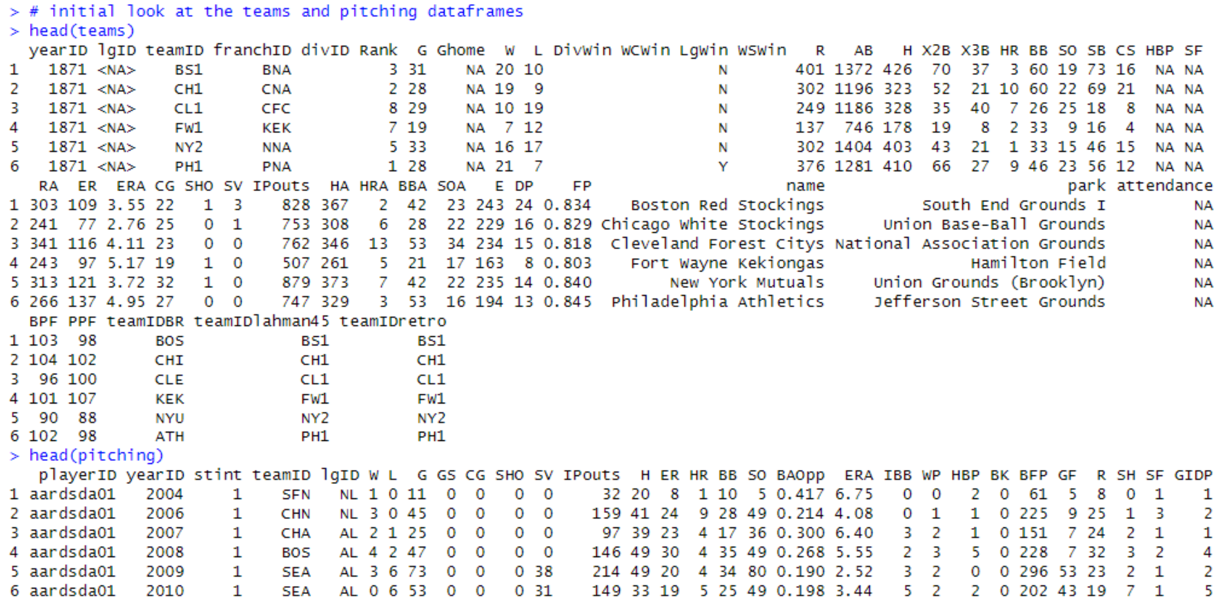
```
> # loads required packages
> library(plyr)
> library(dplyr) # plyr and dplyr used for data frame manipulation
> library(ggplot2) # ggplot2 used for data visualization
> library(corrplot) # corrplot used to create correlation plots
> library(gridExtra) # gridExtra used to display more than 1 graph at a time
```

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



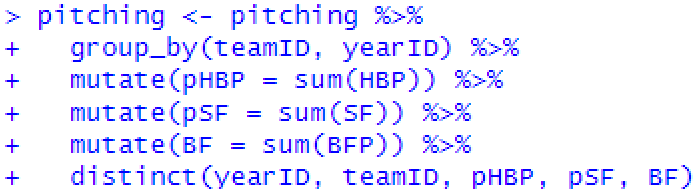
```
> # imports relevant data
> teams <- read.csv(
+   "C:/Users/atwoo/OneDrive/Desktop/MSDADirectory/lahman_1871-2023_csv/Teams.csv")
> pitching <- read.csv(
+   "C:/Users/atwoo/OneDrive/Desktop/MSDADirectory/lahman_1871-2023_csv/Pitching.csv")
```

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



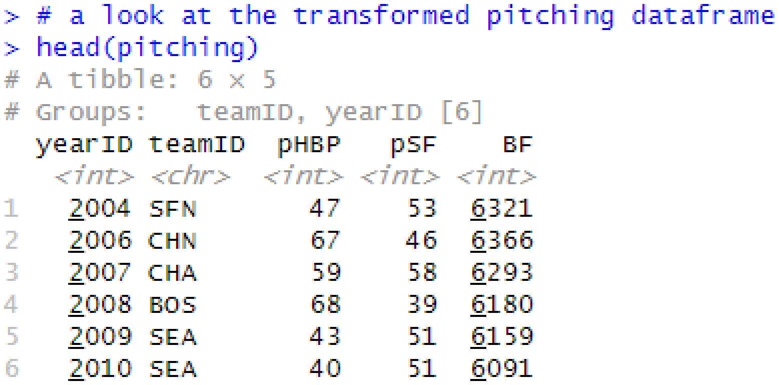
```
> # initial look at the teams and pitching dataframes
> head(teams)
  yearID lgID teamID franchID divID Rank  G Ghome  W  L DivWin WCWin LgWin WSWin   R   AB   H X2B X3B HR BB SO SB CS HBP SF
1   1871 <NA>    BS1      BNA          3 31    NA 20 10                  N       401 1372 426  70  37  3 60 19 73 16  NA NA
2   1871 <NA>    CH1      CNA          2 28    NA 19  9                  N       302 1196 323  52  21 10 60 22 69 21  NA NA
3   1871 <NA>    CL1      CFC          8 29    NA 10 19                  N       249 1186 328  35  40  7 26 25 18  8  NA NA
4   1871 <NA>    FW1      KEK          7 19    NA  7 12                  N       137  746 178  19   8  2 33  9 16  4  NA NA
5   1871 <NA>    NY2      NNA          5 33    NA 16 17                  N       302 1404 403  43  21  1 33 15 46 15  NA NA
6   1871 <NA>    PH1      PNA          1 28    NA 21  7                  Y       376 1281 410  66  27  9 46 23 56 12  NA NA
   RA  ER  ERA CG SHO SV IPouts  HA HRA BBA SOA   E DP    FP                    name                         park attendance
1 303 109 3.55 22   1  3    828 367   2  42  23 243 24 0.834    Boston Red Stockings          South End Grounds I         NA
2 241  77 2.76 25   0  1    753 308   6  28  22 229 16 0.829 Chicago White Stockings      Union Base-Ball Grounds         NA
3 341 116 4.11 23   0  0    762 346  13  53  34 234 15 0.818  Cleveland Forest Citys National Association Grounds         NA
4 243  97 5.17 19   1  0    507 261   5  21  17 163  8 0.803    Fort Wayne Kekiongas               Hamilton Field         NA
5 313 121 3.72 32   1  0    879 373   7  42  22 235 14 0.840        New York Mutuals     Union Grounds (Brooklyn)         NA
6 266 137 4.95 27   0  0    747 329   3  53  16 194 13 0.845  Philadelphia Athletics     Jefferson Street Grounds         NA
  BPF PPF teamIDBR teamIDlahman45 teamIDretro
1 103  98      BOS            BS1         BS1
2 104 102      CHI            CH1         CH1
3  96 100      CLE            CL1         CL1
4 101 107      KEK            FW1         FW1
5  90  88      NYU            NY2         NY2
6 102  98      ATH            PH1         PH1
> head(pitching)
   playerID yearID stint teamID lgID W L  G GS CG SHO SV IPouts  H ER HR BB SO BAOpp  ERA IBB WP HBP BK BFP GF  R SH SF GIDP
1 aardsda01   2004     1    SFN   NL 1 0 11  0  0   0  0     32 20  8  1 10  5 0.417 6.75   0  0   2  0  61  5  8  0  1    1
2 aardsda01   2006     1    CHN   NL 3 0 45  0  0   0  0    159 41 24  9 28 49 0.214 4.08   0  1   1  0 225  9 25  1  3    2
3 aardsda01   2007     1    CHA   AL 2 1 25  0  0   0  0     97 39 23  4 17 36 0.300 6.40   3  2   1  0 151  7 24  2  1    1
4 aardsda01   2008     1    BOS   AL 4 2 47  0  0   0  0    146 49 30  4 35 49 0.268 5.55   2  3   5  0 228  7 32  3  2    4
5 aardsda01   2009     1    SEA   AL 3 6 73  0  0   0 38    214 49 20  4 34 80 0.190 2.52   3  2   0  0 296 53 23  2  1    2
6 aardsda01   2010     1    SEA   AL 0 6 53  0  0   0 31    149 33 19  5 25 49 0.198 3.44   5  2   2  0 202 43 19  7  1    5
```

- 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.



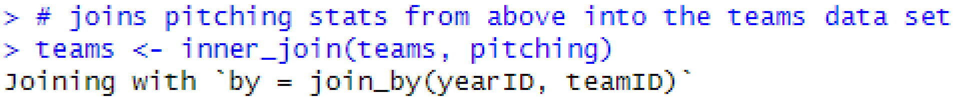
```
> pitching <- pitching %>%
+   group_by(teamID, yearID) %>%
+   mutate(pHBP = sum(HBP)) %>%
+   mutate(pSF = sum(SF)) %>%
+   mutate(BF = sum(BFP)) %>%
+   distinct(yearID, teamID, pHBP, pSF, BF)
```

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



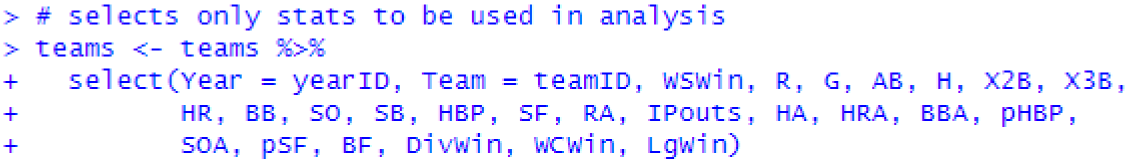
```
> # a look at the transformed pitching dataframe
> head(pitching)
# A tibble: 6 × 5
# Groups:   teamID, yearID [6]
  yearID teamID  pHBP   pSF    BF
   <int> <chr>  <int> <int> <int>
1   2004 SFN       47    53  6321
2   2006 CHN       67    46  6366
3   2007 CHA       59    58  6293
4   2008 BOS       68    39  6180
5   2009 SEA       43    51  6159
6   2010 SEA       40    51  6091
```

- This transformed pitching data frame is **joined with the teams data frame** using the inner_join() function.



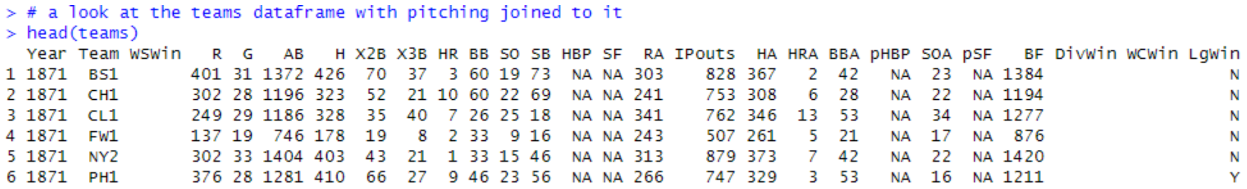
```
> # joins pitching stats from above into the teams data set
> teams <- inner_join(teams, pitching)
Joining with `by = join_by(yearID, teamID)`
```

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



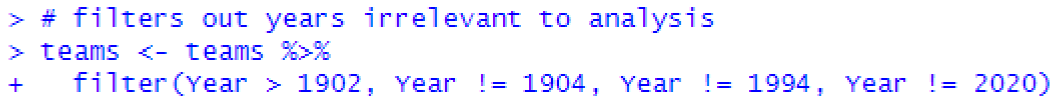
```
> # selects only stats to be used in analysis
> teams <- teams %>%
+   select(Year = yearID, Team = teamID, WSWin, R, G, AB, H, X2B, X3B,
+          HR, BB, SO, SB, HBP, SF, RA, IPouts, HA, HRA, BBA, pHBP,
+          SOA, pSF, BF, DivWin, WCWin, LgWin)
```

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



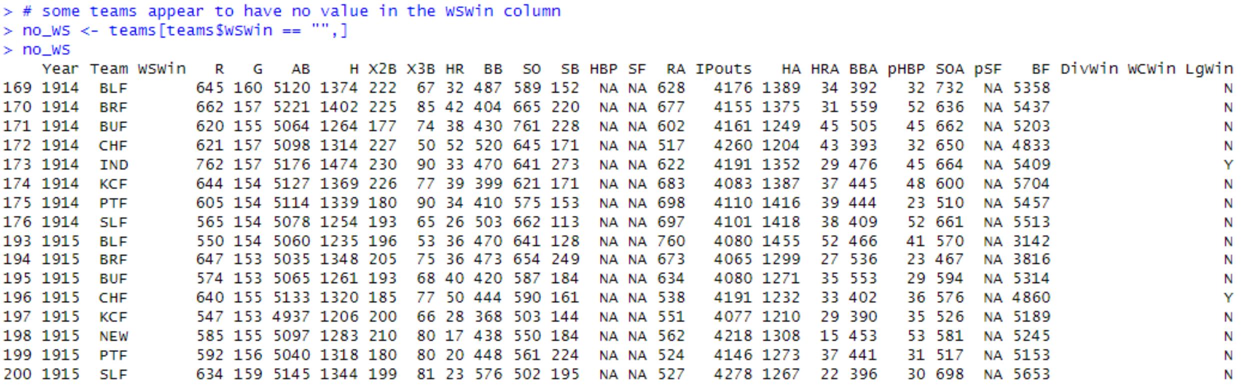
```
> # a look at the teams dataframe with pitching joined to it
> head(teams)
  Year Team WSWin   R  G   AB   H X2B X3B HR BB SO SB HBP SF  RA IPouts  HA HRA BBA pHBP SOA pSF   BF DivWin WCWin LgWin
1 1871  BS1        401 31 1372 426  70  37  3 60 19 73  NA NA 303    828 367   2  42   NA  23  NA 1384                   N
2 1871  CH1        302 28 1196 323  52  21 10 60 22 69  NA NA 241    753 308   6  28   NA  22  NA 1194                   N
3 1871  CL1        249 29 1186 328  35  40  7 26 25 18  NA NA 341    762 346  13  53   NA  34  NA 1277                   N
4 1871  FW1        137 19  746 178  19   8  2 33  9 16  NA NA 243    507 261   5  21   NA  17  NA  876                   N
5 1871  NY2        302 33 1404 403  43  21  1 33 15 46  NA NA 313    879 373   7  42   NA  22  NA 1420                   N
6 1871  PH1        376 28 1281 410  66  27  9 46 23 56  NA NA 266    747 329   3  53   NA  16  NA 1211                   Y
```

- **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**.

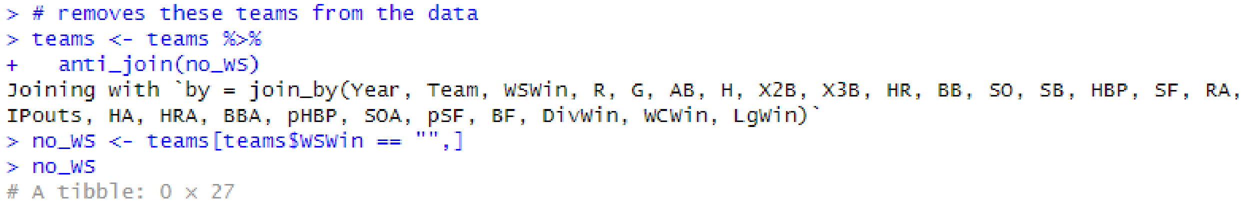


```
> # filters out years irrelevant to analysis
> teams <- teams %>%
+   filter(Year > 1902, Year != 1904, Year != 1994, Year != 2020)
```

- 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.

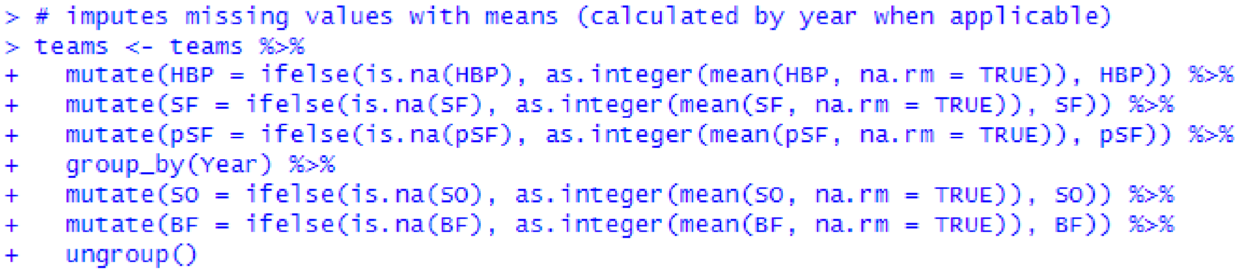


```
> # some teams appear to have no value in the WSWin column
> no_WS <- teams[teams$WSWin == "",]
> no_WS
    Year Team WSWin   R   G   AB    H X2B X3B HR  BB  SO  SB HBP SF  RA IPouts   HA HRA BBA pHBP SOA pSF   BF DivWin WCWin LgWin
169 1914  BLF        645 160 5120 1374 222  67 32 487 589 152  NA NA 628   4176 1389  34 392   32 732  NA 5358                   N
170 1914  BRF        662 157 5221 1402 225  85 42 404 665 220  NA NA 677   4155 1375  31 559   52 636  NA 5437                   N
171 1914  BUF        620 155 5064 1264 177  74 38 430 761 228  NA NA 602   4161 1249  45 505   45 662  NA 5203                   N
172 1914  CHF        621 157 5098 1314 227  50 52 520 645 171  NA NA 517   4260 1204  43 393   32 650  NA 4833                   N
173 1914  IND        762 157 5176 1474 230  90 33 470 641 273  NA NA 622   4191 1352  29 476   45 664  NA 5409                   Y
174 1914  KCF        644 154 5127 1369 226  77 39 399 621 171  NA NA 683   4083 1387  37 445   48 600  NA 5704                   N
175 1914  PTF        605 154 5114 1339 180  90 34 410 575 153  NA NA 698   4110 1416  39 444   23 510  NA 5457                   N
176 1914  SLF        565 154 5078 1254 193  65 26 503 662 113  NA NA 697   4101 1418  38 409   52 661  NA 5513                   N
193 1915  BLF        550 154 5060 1235 196  53 36 470 641 128  NA NA 760   4080 1455  52 466   41 570  NA 3142                   N
194 1915  BRF        647 153 5035 1348 205  75 36 473 654 249  NA NA 673   4065 1299  27 536   23 467  NA 3816                   N
195 1915  BUF        574 153 5065 1261 193  68 40 420 587 184  NA NA 634   4080 1271  35 553   29 594  NA 5314                   N
196 1915  CHF        640 155 5133 1320 185  77 50 444 590 161  NA NA 538   4191 1232  33 402   36 576  NA 4860                   Y
197 1915  KCF        547 153 4937 1206 200  66 28 368 503 144  NA NA 551   4077 1210  29 390   35 526  NA 5189                   N
198 1915  NEW        585 155 5097 1283 210  80 17 438 550 184  NA NA 562   4218 1308  15 453   53 581  NA 5245                   N
199 1915  PTF        592 156 5040 1318 180  80 20 448 561 224  NA NA 524   4146 1273  37 441   31 517  NA 5153                   N
200 1915  SLF        634 159 5145 1344 199  81 23 576 502 195  NA NA 527   4278 1267  22 396   30 698  NA 5653                   N
```



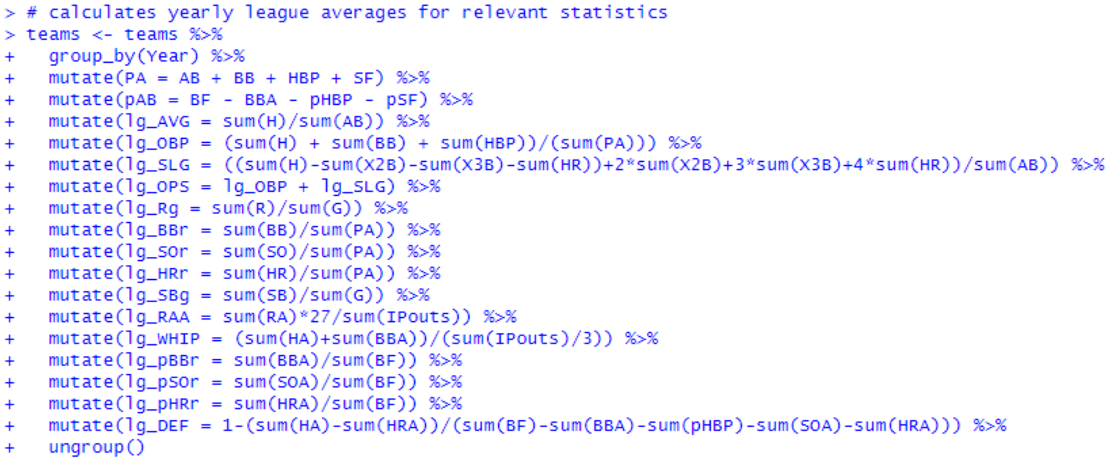
```
> # removes these teams from the data
> teams <- teams %>%
+   anti_join(no_WS)
Joining with `by = join_by(Year, Team, WSWin, R, G, AB, H, X2B, X3B, HR, BB, SO, SB, HBP, SF, RA,
IPouts, HA, HRA, BBA, pHBP, SOA, pSF, BF, DivWin, WCWin, LgWin)`
> no_WS <- teams[teams$WSWin == "",]
> no_WS
# A tibble: 0 × 27
```

- 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.



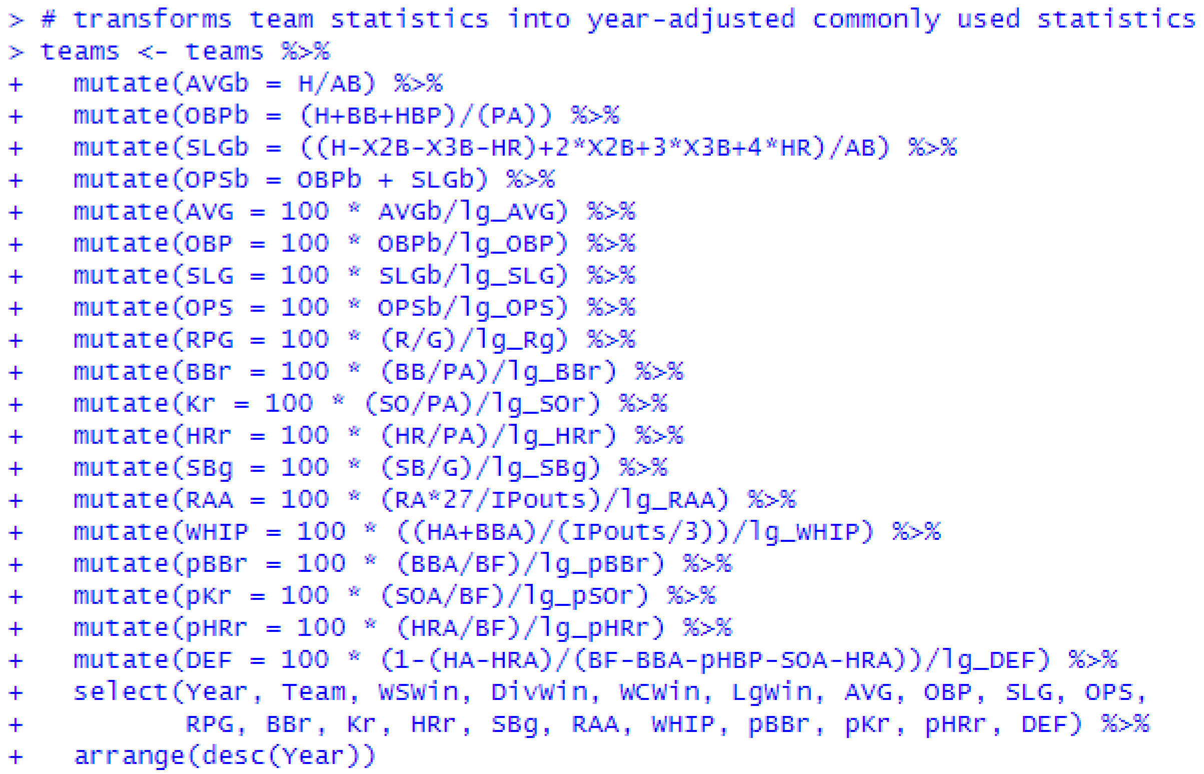
```
> # imputes missing values with means (calculated by year when applicable)
> teams <- teams %>%
+   mutate(HBP = ifelse(is.na(HBP), as.integer(mean(HBP, na.rm = TRUE)), HBP)) %>%
+   mutate(SF = ifelse(is.na(SF), as.integer(mean(SF, na.rm = TRUE)), SF)) %>%
+   mutate(pSF = ifelse(is.na(pSF), as.integer(mean(pSF, na.rm = TRUE)), pSF)) %>%
+   group_by(Year) %>%
+   mutate(SO = ifelse(is.na(SO), as.integer(mean(SO, na.rm = TRUE)), SO)) %>%
+   mutate(BF = ifelse(is.na(BF), as.integer(mean(BF, na.rm = TRUE)), BF)) %>%
+   ungroup()
```

- **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.



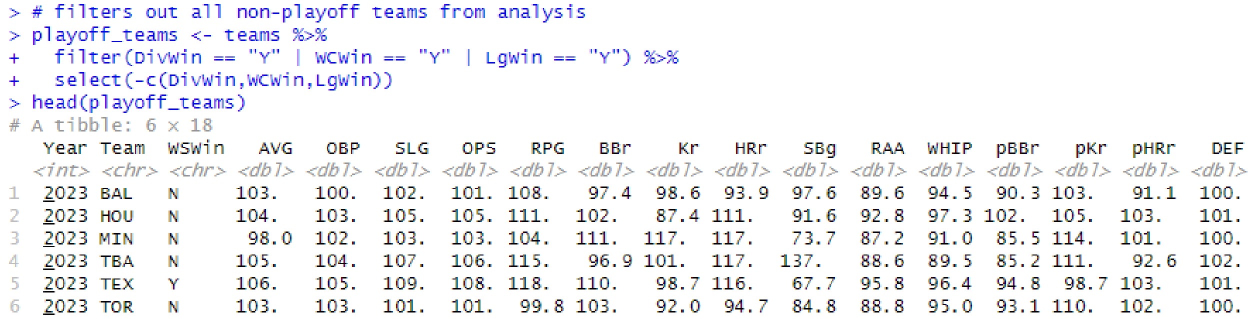
```
> # calculates yearly league averages for relevant statistics
> teams <- teams %>%
+   group_by(Year) %>%
+   mutate(PA = AB + BB + HBP + SF) %>%
+   mutate(pAB = BF - BBA - pHBP - pSF) %>%
+   mutate(lg_AVG = sum(H)/sum(AB)) %>%
+   mutate(lg_OBP = (sum(H) + sum(BB) + sum(HBP))/(sum(PA))) %>%
+   mutate(lg_SLG = ((sum(H)-sum(X2B)-sum(X3B)-sum(HR))+2*sum(X2B)+3*sum(X3B)+4*sum(HR))/sum(AB)) %>%
+   mutate(lg_OPS = lg_OBP + lg_SLG) %>%
+   mutate(lg_Rg = sum(R)/sum(G)) %>%
+   mutate(lg_BBr = sum(BB)/sum(PA)) %>%
+   mutate(lg_SOr = sum(SO)/sum(PA)) %>%
+   mutate(lg_HRr = sum(HR)/sum(PA)) %>%
+   mutate(lg_SBg = sum(SB)/sum(G)) %>%
+   mutate(lg_RAA = sum(RA)*27/sum(IPouts)) %>%
+   mutate(lg_WHIP = (sum(HA)+sum(BBA))/(sum(IPouts)/3)) %>%
+   mutate(lg_pBBr = sum(BBA)/sum(BF)) %>%
+   mutate(lg_pSOr = sum(SOA)/sum(BF)) %>%
+   mutate(lg_pHRr = sum(HRA)/sum(BF)) %>%
+   mutate(lg_DEF = 1-(sum(HA)-sum(HRA))/(sum(BF)-sum(BBA)-sum(pHBP)-sum(SOA)-sum(HRA))) %>%
+   ungroup()
```

- 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.



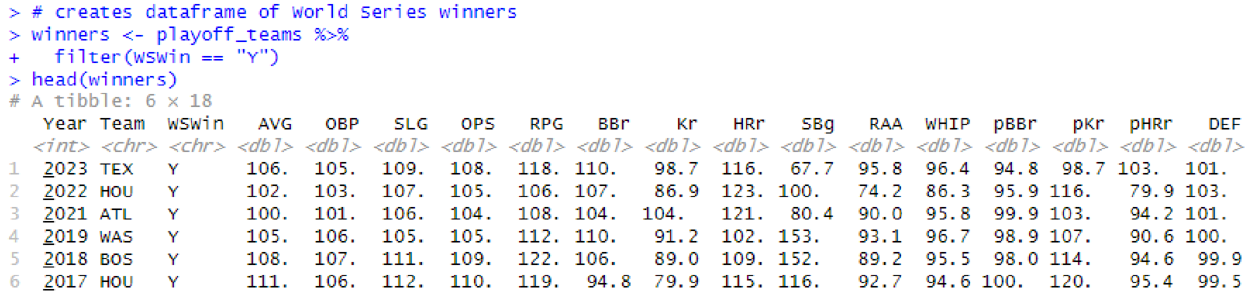
```
> # transforms team statistics into year-adjusted commonly used statistics
> teams <- teams %>%
+   mutate(AVGb = H/AB) %>%
+   mutate(OBPb = (H+BB+HBP)/(PA)) %>%
+   mutate(SLGb = ((H-X2B-X3B-HR)+2*X2B+3*X3B+4*HR)/AB) %>%
+   mutate(OPSb = OBPb + SLGb) %>%
+   mutate(AVG = 100 * AVGb/lg_AVG) %>%
+   mutate(OBP = 100 * OBPb/lg_OBP) %>%
+   mutate(SLG = 100 * SLGb/lg_SLG) %>%
+   mutate(OPS = 100 * OPSb/lg_OPS) %>%
+   mutate(RPG = 100 * (R/G)/lg_Rg) %>%
+   mutate(BBr = 100 * (BB/PA)/lg_BBr) %>%
+   mutate(Kr = 100 * (SO/PA)/lg_SOr) %>%
+   mutate(HRr = 100 * (HR/PA)/lg_HRr) %>%
+   mutate(SBg = 100 * (SB/G)/lg_SBg) %>%
+   mutate(RAA = 100 * (RA*27/IPouts)/lg_RAA) %>%
+   mutate(WHIP = 100 * ((HA+BBA)/(IPouts/3))/lg_WHIP) %>%
+   mutate(pBBr = 100 * (BBA/BF)/lg_pBBr) %>%
+   mutate(pKr = 100 * (SOA/BF)/lg_pSOr) %>%
+   mutate(pHRr = 100 * (HRA/BF)/lg_pHRr) %>%
+   mutate(DEF = 100 * (1-(HA-HRA)/(BF-BBA-pHBP-SOA-HRA))/lg_DEF) %>%
+   select(Year, Team, WSWin, DivWin, WCWin, LgWin, AVG, OBP, SLG, OPS,
+          RPG, BBr, Kr, HRr, SBg, RAA, WHIP, pBBr, pKr, pHRr, DEF) %>%
+   arrange(desc(Year))
```

- 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.

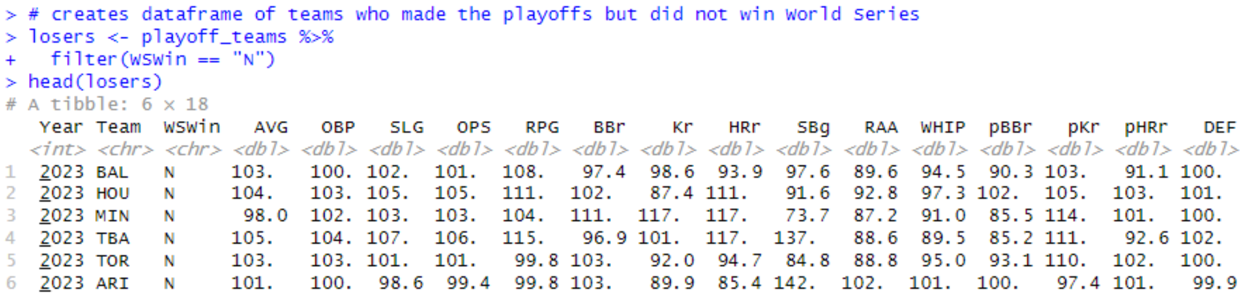


```
> # filters out all non-playoff teams from analysis
> playoff_teams <- teams %>%
+   filter(DivWin == "Y" | WCWin == "Y" | LgWin == "Y") %>%
+   select(-c(DivWin,WCWin,LgWin))
> head(playoff_teams)
# A tibble: 6 × 18
   Year Team  WSWin   AVG   OBP   SLG   OPS   RPG   BBr    Kr   HRr   SBg   RAA  WHIP  pBBr   pKr  pHRr   DEF
  <int> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1  2023 BAL   N     103.  100.  102.  101.  108.   97.4  98.6  93.9  97.6  89.6  94.5  90.3 103.   91.1  100.
2  2023 HOU   N     104.  103.  105.  105.  111.  102.   87.4 111.   91.6  92.8  97.3 102.  105.  103.   101.
3  2023 MIN   N      98.0 102.  103.  103.  104.  111.  117.  117.   73.7  87.2  91.0  85.5 114.  101.   100.
4  2023 TBA   N     105.  104.  107.  106.  115.   96.9 101.  117.  137.   88.6  89.5  85.2 111.   92.6  102.
5  2023 TEX   Y     106.  105.  109.  108.  118.  110.   98.7 116.   67.7  95.8  96.4  94.8  98.7 103.   101.
6  2023 TOR   N     103.  103.  101.  101.   99.8 103.   92.0  94.7  84.8  88.8  95.0  93.1 110.  102.   100.
```

- 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.

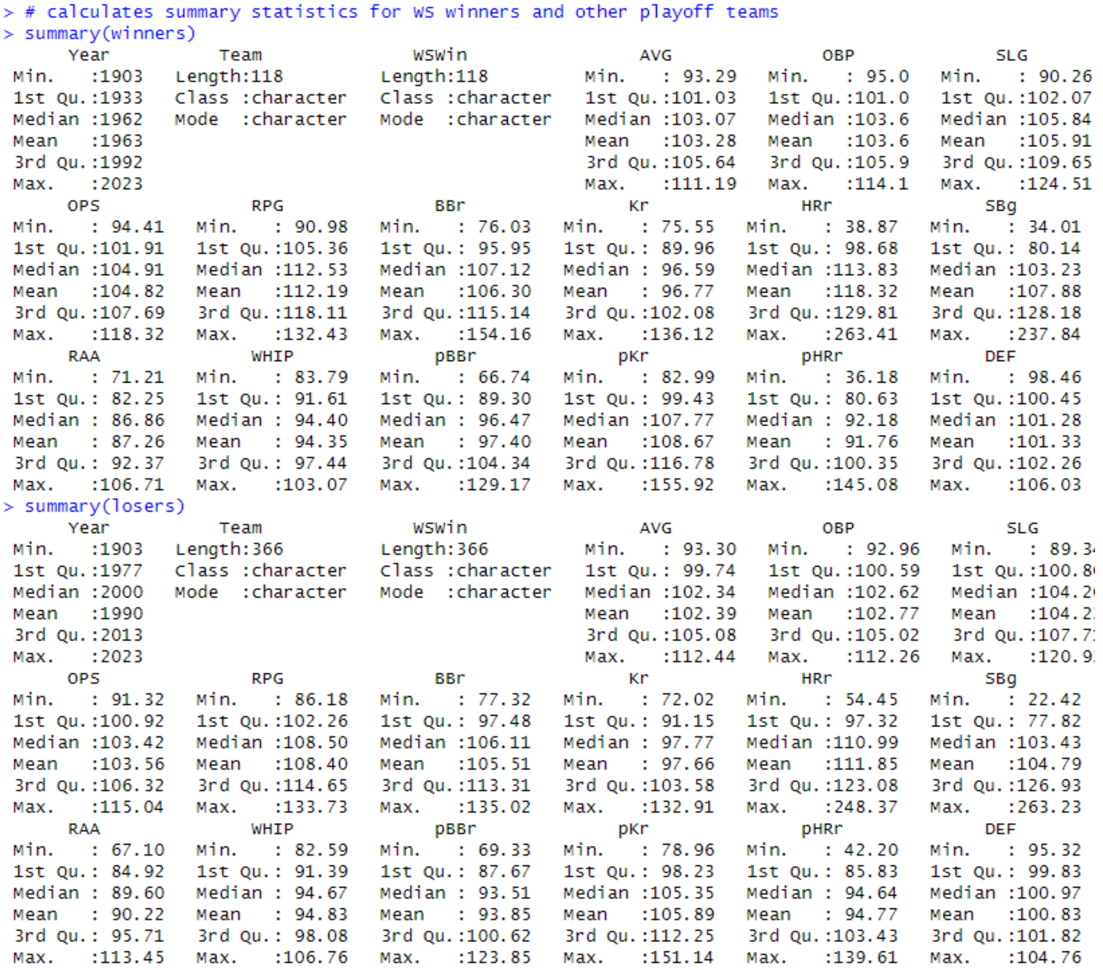


```
> # creates dataframe of World Series winners
> winners <- playoff_teams %>%
+   filter(WSWin == "Y")
> head(winners)
# A tibble: 6 × 18
   Year Team  WSWin   AVG   OBP   SLG   OPS   RPG   BBr    Kr   HRr   SBg   RAA  WHIP  pBBr   pKr  pHRr   DEF
  <int> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1  2023 TEX   Y      106.  105.  109.  108.  118.  110.   98.7 116.   67.7  95.8  96.4  94.8  98.7 103.  101.
2  2022 HOU   Y      102.  103.  107.  105.  106.  107.   86.9 123.  100.   74.2  86.3  95.9 116.   79.9 103.
3  2021 ATL   Y      100.  101.  106.  104.  108.  104.  104.  121.   80.4  90.0  95.8  99.9 103.   94.2 101.
4  2019 WAS   Y      105.  106.  105.  105.  112.  110.   91.2 102.  153.   93.1  96.7  98.9 107.   90.6 100.
5  2018 BOS   Y      108.  107.  111.  109.  122.  106.   89.0 109.  152.   89.2  95.5  98.0 114.   94.6  99.9
6  2017 HOU   Y      111.  106.  112.  110.  119.   94.8  79.9 115.  116.   92.7  94.6 100.  120.   95.4  99.5
```



```
> # creates dataframe of teams who made the playoffs but did not win World Series
> losers <- playoff_teams %>%
+   filter(WSWin == "N")
> head(losers)
# A tibble: 6 × 18
   Year Team  WSWin   AVG   OBP   SLG   OPS   RPG   BBr    Kr   HRr   SBg   RAA  WHIP  pBBr   pKr  pHRr   DEF
  <int> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1  2023 BAL   N     103.  100.  102.  101.  108.   97.4  98.6  93.9  97.6  89.6  94.5  90.3 103.   91.1 100. 
2  2023 HOU   N     104.  103.  105.  105.  111.  102.   87.4 111.   91.6  92.8  97.3 102.  105.  103.  101. 
3  2023 MIN   N      98.0 102.  103.  103.  104.  111.  117.  117.   73.7  87.2  91.0  85.5 114.  101.  100. 
4  2023 TBA   N     105.  104.  107.  106.  115.   96.9 101.  117.  137.   88.6  89.5  85.2 111.   92.6 102. 
5  2023 TOR   N     103.  103.  101.  101.   99.8 103.   92.0  94.7  84.8  88.8  95.0  93.1 110.  102.  100. 
6  2023 ARI   N     101.  100.   98.6  99.4  99.8 103.   89.9  85.4 142.  102.  101.  100.   97.4 101.   99.9
```

- 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.



```
> # calculates summary statistics for WS winners and other playoff teams
> summary(winners)
      Year          Team              WSWin                AVG              OBP              SLG        
 Min.   :1903   Length:118         Length:118         Min.   : 93.29   Min.   : 95.0   Min.   : 90.26  
 1st Qu.:1933   Class :character   Class :character   1st Qu.:101.03   1st Qu.:101.0   1st Qu.:102.07  
 Median :1962   Mode  :character   Mode  :character   Median :103.07   Median :103.6   Median :105.84  
 Mean   :1963                                         Mean   :103.28   Mean   :103.6   Mean   :105.91  
 3rd Qu.:1992                                         3rd Qu.:105.64   3rd Qu.:105.9   3rd Qu.:109.65  
 Max.   :2023                                         Max.   :111.19   Max.   :114.1   Max.   :124.51  
      OPS              RPG              BBr               Kr              HRr              SBg        
 Min.   : 94.41   Min.   : 90.98   Min.   : 76.03   Min.   : 75.55   Min.   : 38.87   Min.   : 34.01  
 1st Qu.:101.91   1st Qu.:105.36   1st Qu.: 95.95   1st Qu.: 89.96   1st Qu.: 98.68   1st Qu.: 80.14  
 Median :104.91   Median :112.53   Median :107.12   Median : 96.59   Median :113.83   Median :103.23  
 Mean   :104.82   Mean   :112.19   Mean   :106.30   Mean   : 96.77   Mean   :118.32   Mean   :107.88  
 3rd Qu.:107.69   3rd Qu.:118.11   3rd Qu.:115.14   3rd Qu.:102.08   3rd Qu.:129.81   3rd Qu.:128.18  
 Max.   :118.32   Max.   :132.43   Max.   :154.16   Max.   :136.12   Max.   :263.41   Max.   :237.84  
      RAA              WHIP             pBBr             pKr              pHRr             DEF        
 Min.   : 71.21   Min.   : 83.79   Min.   : 66.74   Min.   : 82.99   Min.   : 36.18   Min.   : 98.46  
 1st Qu.: 82.25   1st Qu.: 91.61   1st Qu.: 89.30   1st Qu.: 99.43   1st Qu.: 80.63   1st Qu.:100.45  
 Median : 86.86   Median : 94.40   Median : 96.47   Median :107.77   Median : 92.18   Median :101.28  
 Mean   : 87.26   Mean   : 94.35   Mean   : 97.40   Mean   :108.67   Mean   : 91.76   Mean   :101.33  
 3rd Qu.: 92.37   3rd Qu.: 97.44   3rd Qu.:104.34   3rd Qu.:116.78   3rd Qu.:100.35   3rd Qu.:102.26  
 Max.   :106.71   Max.   :103.07   Max.   :129.17   Max.   :155.92   Max.   :145.08   Max.   :106.03  
> summary(losers)
      Year          Team              WSWin                AVG              OBP               SLG        
 Min.   :1903   Length:366         Length:366         Min.   : 93.30   Min.   : 92.96   Min.   : 89.3
 1st Qu.:1977   Class :character   Class :character   1st Qu.: 99.74   1st Qu.:100.59   1st Qu.:100.8
 Median :2000   Mode  :character   Mode  :character   Median :102.34   Median :102.62   Median :104.2
 Mean   :1990                                         Mean   :102.39   Mean   :102.77   Mean   :104.2
 3rd Qu.:2013                                         3rd Qu.:105.08   3rd Qu.:105.02   3rd Qu.:107.7
 Max.   :2023                                         Max.   :112.44   Max.   :112.26   Max.   :120.9
      OPS              RPG              BBr               Kr              HRr              SBg        
 Min.   : 91.32   Min.   : 86.18   Min.   : 77.32   Min.   : 72.02   Min.   : 54.45   Min.   : 22.42  
 1st Qu.:100.92   1st Qu.:102.26   1st Qu.: 97.48   1st Qu.: 91.15   1st Qu.: 97.32   1st Qu.: 77.82  
 Median :103.42   Median :108.50   Median :106.11   Median : 97.77   Median :110.99   Median :103.43  
 Mean   :103.56   Mean   :108.40   Mean   :105.51   Mean   : 97.66   Mean   :111.85   Mean   :104.79  
 3rd Qu.:106.32   3rd Qu.:114.65   3rd Qu.:113.31   3rd Qu.:103.58   3rd Qu.:123.08   3rd Qu.:126.93  
 Max.   :115.04   Max.   :133.73   Max.   :135.02   Max.   :132.91   Max.   :248.37   Max.   :263.23  
      RAA              WHIP             pBBr             pKr              pHRr             DEF        
 Min.   : 67.10   Min.   : 82.59   Min.   : 69.33   Min.   : 78.96   Min.   : 42.20   Min.   : 95.32  
 1st Qu.: 84.92   1st Qu.: 91.39   1st Qu.: 87.67   1st Qu.: 98.23   1st Qu.: 85.83   1st Qu.: 99.83  
 Median : 89.60   Median : 94.67   Median : 93.51   Median :105.35   Median : 94.64   Median :100.97  
 Mean   : 90.22   Mean   : 94.83   Mean   : 93.85   Mean   :105.89   Mean   : 94.77   Mean   :100.83  
 3rd Qu.: 95.71   3rd Qu.: 98.08   3rd Qu.:100.62   3rd Qu.:112.25   3rd Qu.:103.43   3rd Qu.:101.82  
 Max.   :113.45   Max.   :106.76   Max.   :123.85   Max.   :151.14   Max.   :139.61   Max.   :104.76  
```

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.

$$\boldsymbol{AVG = \frac{H}{AB}}$$

- **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.

$$\boldsymbol{OBP = \frac{H + BB + HBP}{PA}}$$

- **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.

$$\boldsymbol{SLG = \frac{1B + 2B * 2 + 3B * 3 + HR * 4}{AB}}$$

- **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.

$$\boldsymbol{OPS = OBP + SLG}$$

- **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.

$$\boldsymbol{RPG = R/G}$$

- **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.

$$\boldsymbol{BBr = BB/PA}$$

- **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**.

$$\boldsymbol{Kr = SO/PA}$$

- **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.

$$\boldsymbol{HRr = HR/PA}$$

- **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.

$$\boldsymbol{SBg = SB/G}$$

- **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.

$$\boldsymbol{RAA = 27 * RA/IPouts}$$

- **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**.

$$\boldsymbol{WHIP = \frac{H + BBA}{\frac{IPouts}{3}}}$$

- **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.

$$\boldsymbol{pBBr = \frac{BBA}{BF}}$$

- **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.

$$\boldsymbol{pKr = \frac{SOA}{BF}}$$

- **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**.

$$\boldsymbol{pHRr = \frac{HRA}{BF}}$$

- **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.

$$\boldsymbol{DEF = 1 - \frac{HA - HRA}{BF - BBA - pHBP - SOA - HRA}}$$

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.

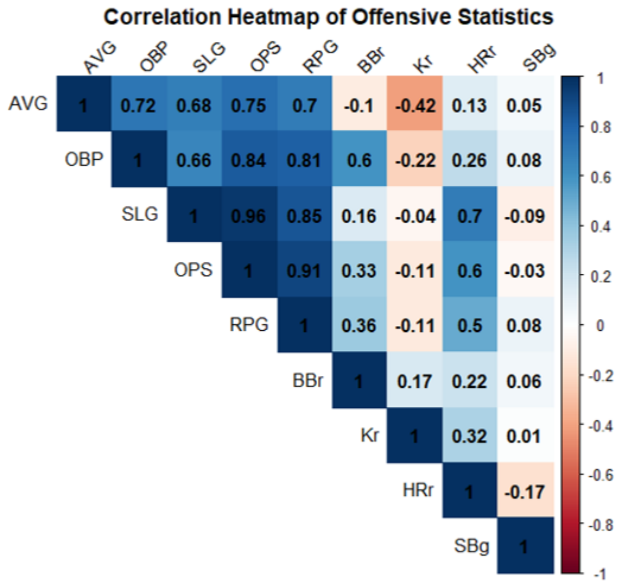
C3. Data Preparation Disadvantage

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

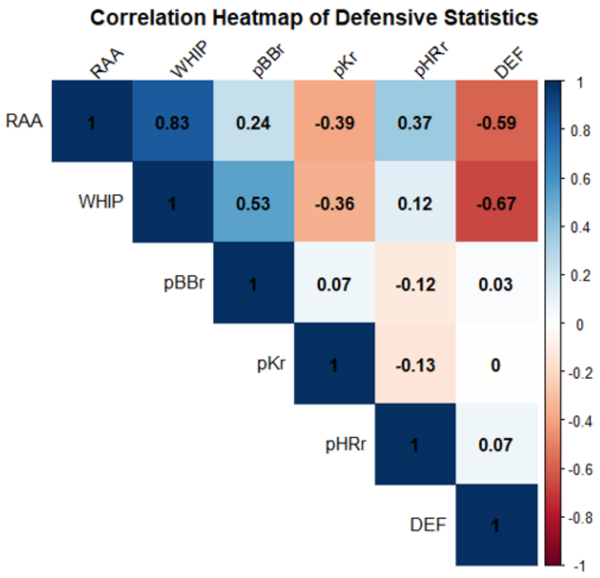
The following **visualizations** provide necessary insight to help understand the structure and spread of the remaining statistics for all playoff teams in the cleaned and prepared data set.

- **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:**



Correlation Heatmap of Offensive Statistics

	AVG	OBP	SLG	OPS	RPG	BBr	Kr	HRr	SBg
AVG	1	0.72	0.68	0.75	0.7	-0.1	-0.42	0.13	0.05
OBP		1	0.66	0.84	0.81	0.6	-0.22	0.26	0.08
SLG			1	0.96	0.85	0.16	-0.04	0.7	-0.09
OPS				1	0.91	0.33	-0.11	0.6	-0.03
RPG					1	0.36	-0.11	0.5	0.08
BBr						1	0.17	0.22	0.06
Kr							1	0.32	0.01
HRr								1	-0.17
SBg									1

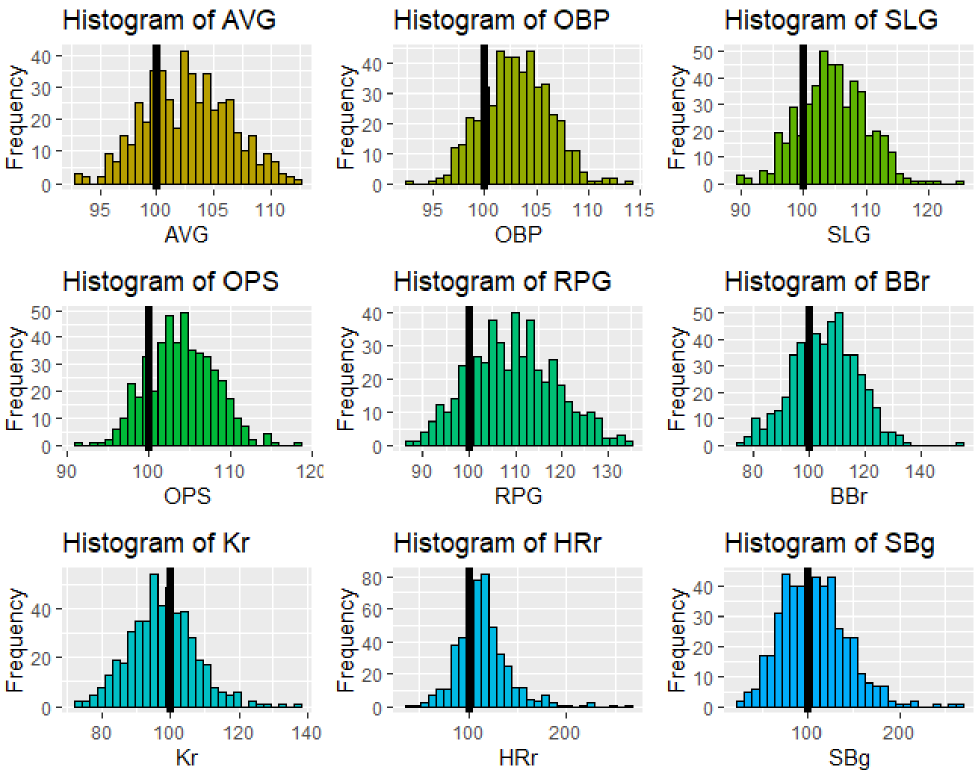


Correlation Heatmap of Defensive Statistics

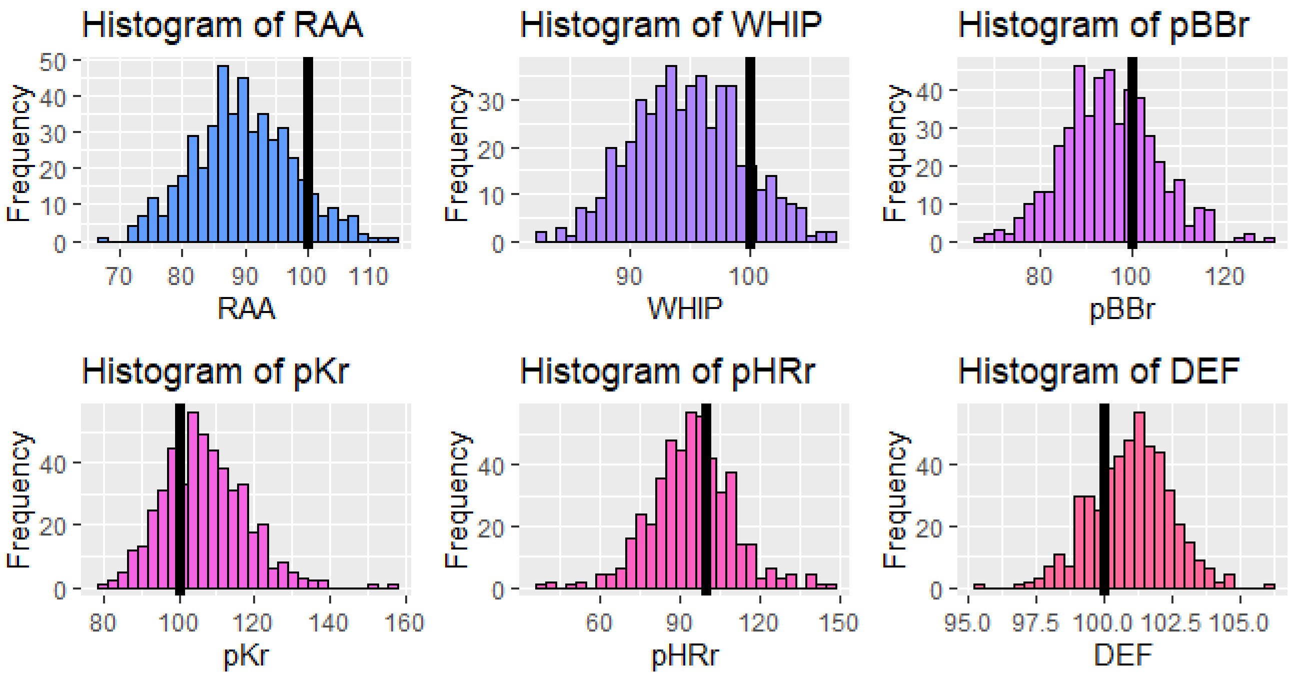
	RAA	WHIP	pBBr	pKr	pHRr	DEF
RAA	1	0.83	0.24	-0.39	0.37	-0.59
WHIP		1	0.53	-0.36	0.12	-0.67
pBBr			1	0.07	-0.12	0.03
pKr				1	-0.13	0
pHRr					1	0.07
DEF						1

  - 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.

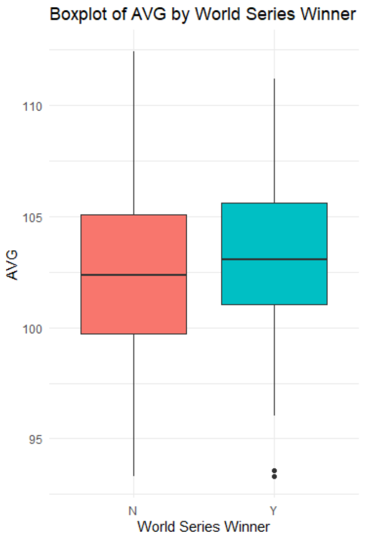
- **Univariate histograms** are generated for all included statistics to show the spread of the factors for all playoff teams.
  - **Offensive statistics histograms:**

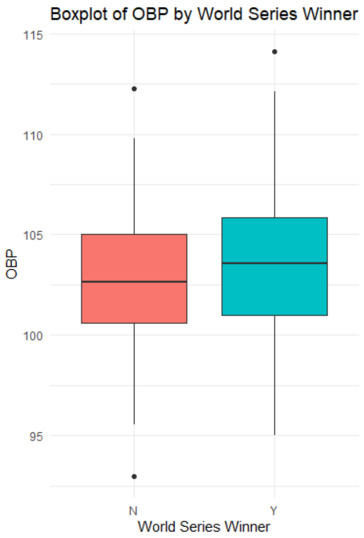


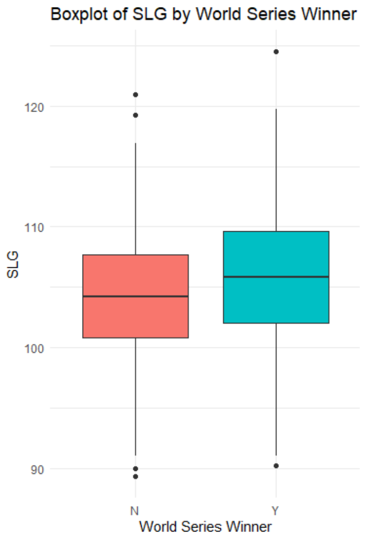
  - **Defensive statistics histograms:**

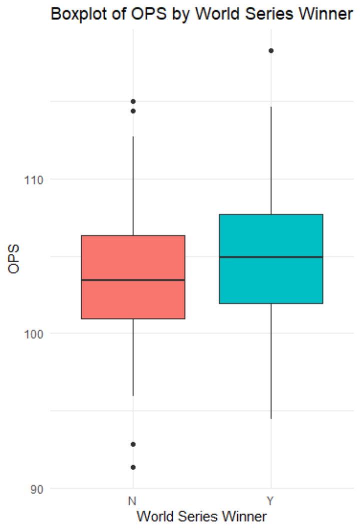


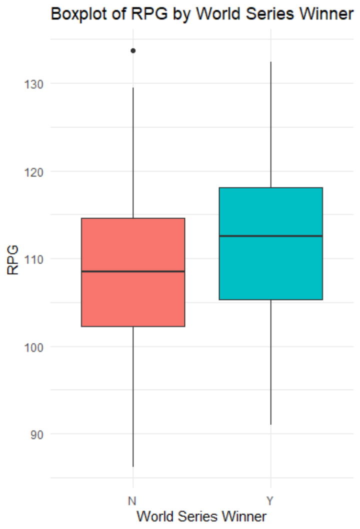
  - A **vertical line at 100** has been placed on each of the histograms to indicate where "league average" exists in relation to the spread of the data. In many cases, this confirms that playoff teams are generally better than league average. For example, in the histogram of OBP, it is clear that many 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:**

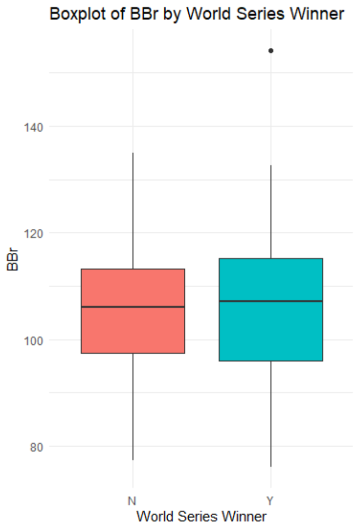


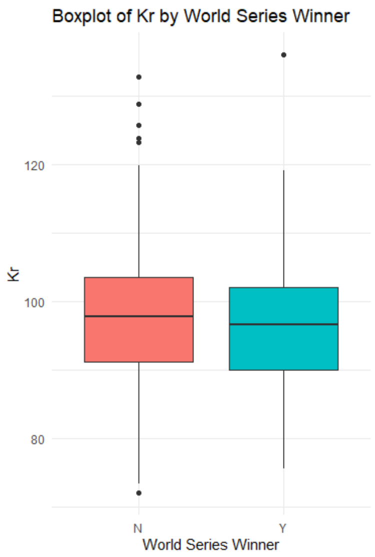


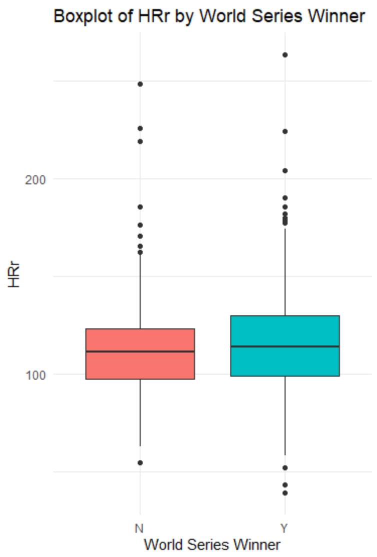


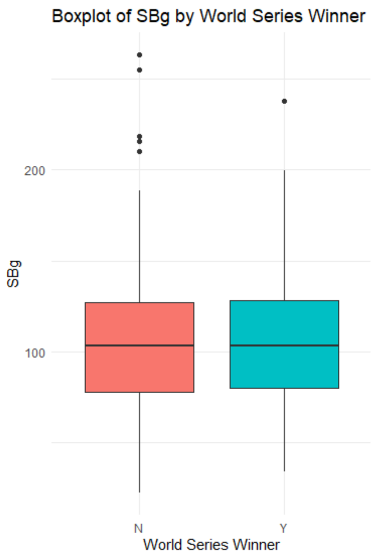




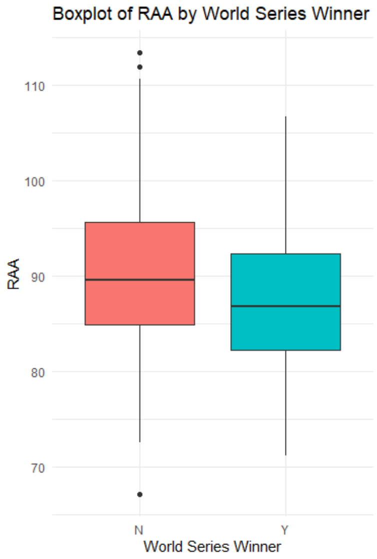


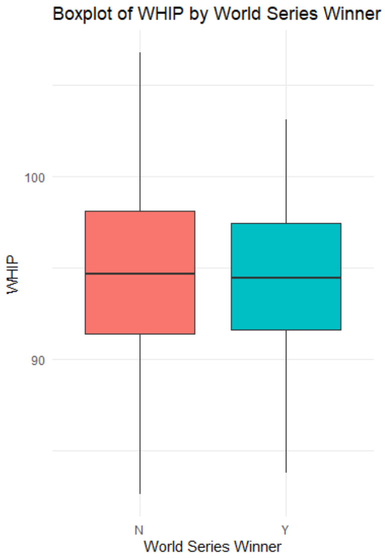


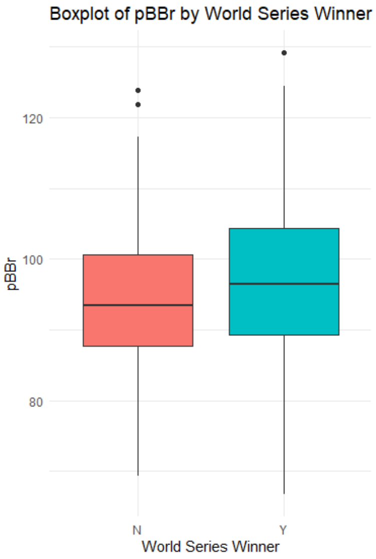


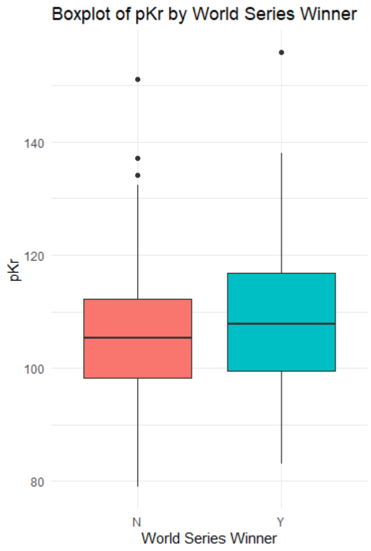


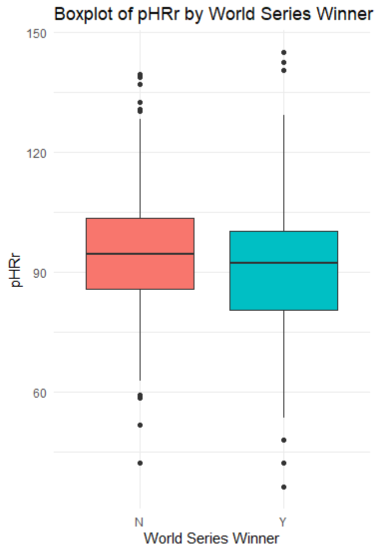
- **Defensive statistics boxplots:**

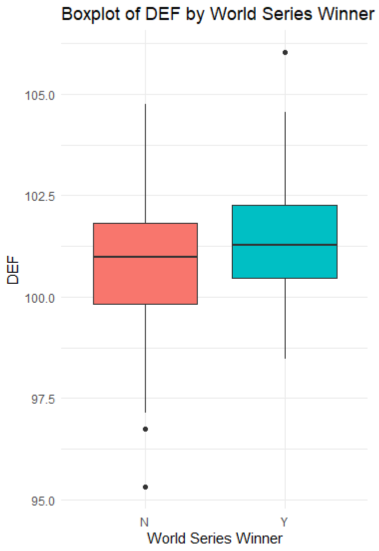












- These boxplots show the minimum, first quartile, median, third quartile, and maximum values for each factor, separated by playoff teams that won the World Series and playoff teams that did not. Once again, it is clear that in most of these factors, teams in both categories perform better than league 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 $\boldsymbol{\mu_1 = \mu_2}$** and an **alternative hypothesis of $\boldsymbol{\mu_1 > \mu_2}$**, where μ_1 represents the mean of the given statistic for World Series-winning teams and μ_2 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.

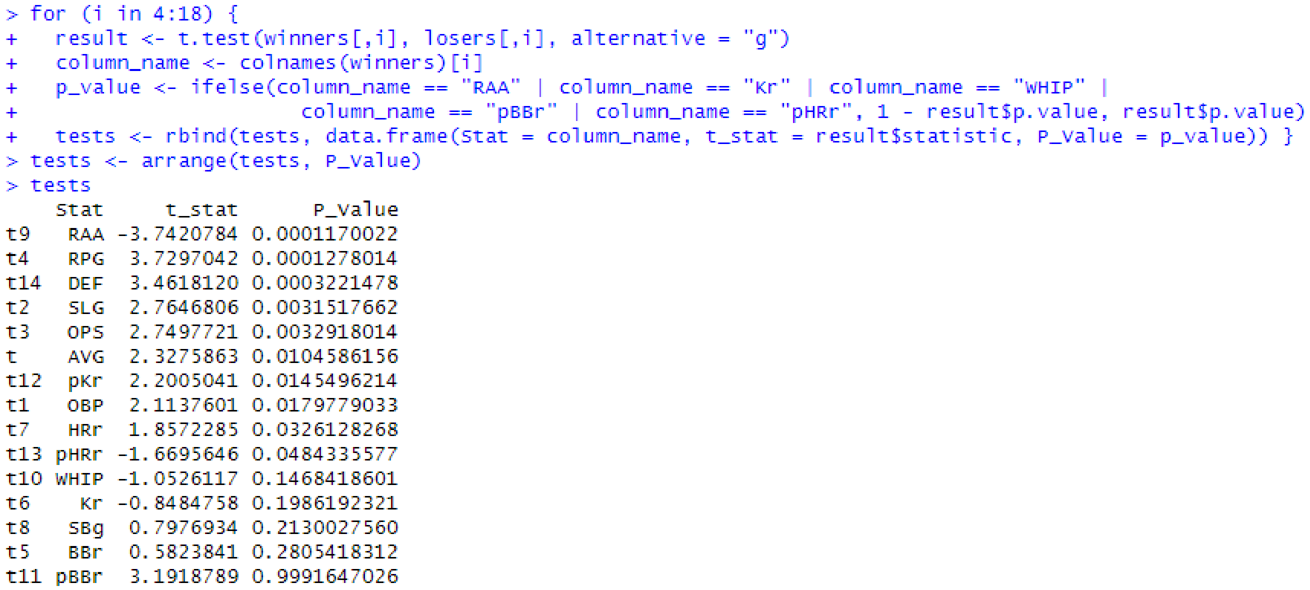
An **advantage** of using hypothesis tests for this analysis is that they will help explain whether a World Series-winning team is significantly more successful than another playoff team in any of the relevant factors, especially if the results show more or less importance in certain areas than is commonly 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.



```
> for (i in 4:18) {
+   result <- t.test(winners[,i], losers[,i], alternative = "g")
+   column_name <- colnames(winners)[i]
+   p_value <- ifelse(column_name == "RAA" | column_name == "Kr" | column_name == "WHIP" |
+                       column_name == "pBBr" | column_name == "pHRr", 1 - result$p.value, result$p.value)
+   tests <- rbind(tests, data.frame(Stat = column_name, t_stat = result$statistic, P_Value = p_value)) }
> tests <- arrange(tests, P_Value)
> tests
    Stat     t_stat      P_Value
t9   RAA -3.7420784 0.0001170022
t4   RPG  3.7297042 0.0001278014
t14  DEF  3.4618120 0.0003221478
t2   SLG  2.7646806 0.0031517662
t3   OPS  2.7497721 0.0032918014
t    AVG  2.3275863 0.0104586156
t12  pKr  2.2005041 0.0145496214
t1   OBP  2.1137601 0.0179779033
t7   HRr  1.8572285 0.0326128268
t13 pHRr -1.6695646 0.0484335577
t10 WHIP -1.0526117 0.1468418601
t6    Kr -0.8484758 0.1986192321
t8   SBg  0.7976934 0.2130027560
t5   BBr  0.5823841 0.2805418312
t11 pBBr  3.1918789 0.9991647026
```

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. Sean Lahman owns this database. The work is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License.