World Series Championship

EB04 - Group 2

Zhanica Arrindell (596427) Vera Gak Anagrova (772713)
Aleksandra Tatko (648925) Ly Le (644801)

2025-17-30

Contents

1 Introduction **1**

2 Task 1 **1**

3 Task 2 **2**

4 Task 3 **3**

1 Introduction

The goal of this report is to scrape, clean, and analyze data on Major League Baseball's World Series results using R. A static table containing the winning and losing teams was extracted from Wikipedia using the rvest package. The dataset was then cleaned to ensure accuracy and consistency. After cleaning, we explored patterns in recent World Series outcomes and visualized the frequency of team wins and losses, excluding wildcard-winning teams. The analysis highlights which teams have been most successful in baseball history and reveals patterns in championship competitiveness through charts comparing team performance over time.

2 Task 1

```
#QUESTION 1

#read wikipedia page
wiki <- "https://en.wikipedia.org/wiki/"
url <- "List_of_World_Series_champions"
champions <- read_html(paste0(wiki,url))

#extract the table we want
champions.table <- html_table(champions)
tab1 <- champions.table[[2]]
```

```r
#inspect my data
head(tab1)
```

```
## # A tibble: 6 x 6
##   Year    `Winning team`             Manager        Series `Losing team` Manager
##   <chr>   <chr>                      <chr>          <chr>  <chr>         <chr>
## 1 1903[a] Boston Americans (1, 1-0)  Jimmy Collins  5-3.m~ Pittsburgh P~ Fred C~
## 2 1904[b] (not played)[c]            (not played)[~ (not ~ (not played)~ (not p~
## 3 1905[d] New York Giants (1, 1-0)   John McGraw    4-1    Philadelphia~ Connie~
## 4 1906    Chicago White Sox (1, 1-0) Fielder Jones  4-2    Chicago Cubs~ Frank ~
## 5 1907    Chicago Cubs (2, 1-1)      Frank Chance   4-0-(~ Detroit Tige~ Hugh J~
## 6 1908    Chicago Cubs (3, 2-1)      Frank Chance   4-1    Detroit Tige~ Hugh J~
```

```r
names(tab1)
```

```
## [1] "Year"         "Winning team" "Manager"      "Series"       "Losing team"
## [6] "Manager"
```

3 Task 2

```r
#QUESTION 2

#handle no world series with NA
tab1[tab1$Year == "No World Series held", ] <- NA

#remove parenthesis and number in team names
# Remove anything inside parentheses
tab1$`Winning team` <- gsub("\\(.*?\\)", "", tab1$`Winning team`)
tab1$`Losing team`  <- gsub("\\(.*?\\)", "", tab1$`Losing team`)

# Remove the [W] wildcard mark (keep this for later filtering)
tab1$`Winning team` <- gsub("\\[W\\]", "", tab1$`Winning team`)

# Trim extra spaces
tab1$`Winning team` <- trimws(tab1$`Winning team`)
tab1$`Losing team`  <- trimws(tab1$`Losing team`)

# Using strsplit for losing team
split_lose <- strsplit(tab1$`Losing team`, "\\(", fixed = FALSE)
tab1$`Losing team` <- sapply(split_lose, function(x) trimws(x[1]))

# Using strsplit for winning team
split_win <- strsplit(tab1$`Winning team`, "\\(", fixed = FALSE)
tab1$`Winning team` <- sapply(split_win, function(x) trimws(x[1]))

colnames(tab1) <- c("Year", "WinningTeam", "WinningManager",
                    "Series", "LosingTeam", "LosingManager")
#show last 20 rows
kable(tail(tab1, 20), caption = "Table 1: Last 20 World Series Results (Cleaned)")
```

Table 1: Table 1: Last 20 World Series Results (Cleaned)

Year	WinningTeam	WinningManager	Series	LosingTeam	LosingManager
2005	Chicago White Sox	Ozzie Guillén	4–0	Houston Astros[W][N]	Phil Garner
2006	St. Louis Cardinals	Tony La Russa	4–1	Detroit Tigers[W]	Jim Leyland
2007	Boston Red Sox	Terry Francona	4–0	Colorado Rockies[W]	Clint Hurdle
2008	Philadelphia Phillies	Charlie Manuel	4–1	Tampa Bay Rays	Joe Maddon
2009	New York Yankees	Joe Girardi	4–2	Philadelphia Phillies	Charlie Manuel
2010	San Francisco Giants	Bruce Bochy	4–1	Texas Rangers	Ron Washington
2011	St. Louis Cardinals	Tony La Russa	4–3	Texas Rangers	Ron Washington
2012	San Francisco Giants	Bruce Bochy	4–0	Detroit Tigers	Jim Leyland
2013	Boston Red Sox	John Farrell	4–2	St. Louis Cardinals	Mike Matheny
2014	San Francisco Giants	Bruce Bochy	4–3	Kansas City Royals[W]	Ned Yost
2015	Kansas City Royals	Ned Yost	4–1	New York Mets	Terry Collins
2016	Chicago Cubs	Joe Maddon	4–3	Cleveland Indians	Terry Francona
2017	Houston Astros	A.J. Hinch	4–3	Los Angeles Dodgers	Dave Roberts
2018	Boston Red Sox	Alex Cora	4–1	Los Angeles Dodgers	Dave Roberts
2019	Washington Nationals	Dave Martinez	4–3	Houston Astros	A. J. Hinch
2020	Los Angeles Dodgers	Dave Roberts	4–2	Tampa Bay Rays	Kevin Cash
2021	Atlanta Braves	Brian Snitker	4–2	Houston Astros	Dusty Baker
2022	Houston Astros	Dusty Baker	4–2	Philadelphia Phillies[W]	Rob Thomson
2023	Texas Rangers	Bruce Bochy	4–1	Arizona Diamondbacks[W]	Torey Lovullo
2024	Los Angeles Dodgers	Dave Roberts	4–1	New York Yankees	Aaron Boone

Over the past 20 World Series championships (2005–2024), 16 different franchises have claimed titles, with only the San Francisco Giants (3) and Boston Red Sox (3) winning multiple times. Eight wildcard teams captured championships, representing 40% of winners, highlighting the impact of postseason upsets. Managerial achievements are notable: Bruce Bochy won four championships across two franchises (three with the Giants, one with the Rangers), while Dave Roberts led the Dodgers to two titles after multiple finals appearances.
- Examining the last 20 entries reveals clear patterns:
- San Francisco Giants (Bruce Bochy) appear three times (2010, 2012, 2014).
- Boston Red Sox and Houston Astros each appear three times as winners.
- Los Angeles Dodgers (Dave Roberts) appear multiple times, both as winners and runners-up, showing consistency.
- Bruce Bochy demonstrates success with different teams (Giants, Rangers).
Game lengths also reveal trends: many recent series ended 4–1 or 4–2, with only 2011, 2016, and 2019 going to a full seven games. There is parity between American League (AL) and National League (NL) teams, and from 2015 onward, different teams have won most years, reflecting competitive balance and likely exciting matchups across MLB.

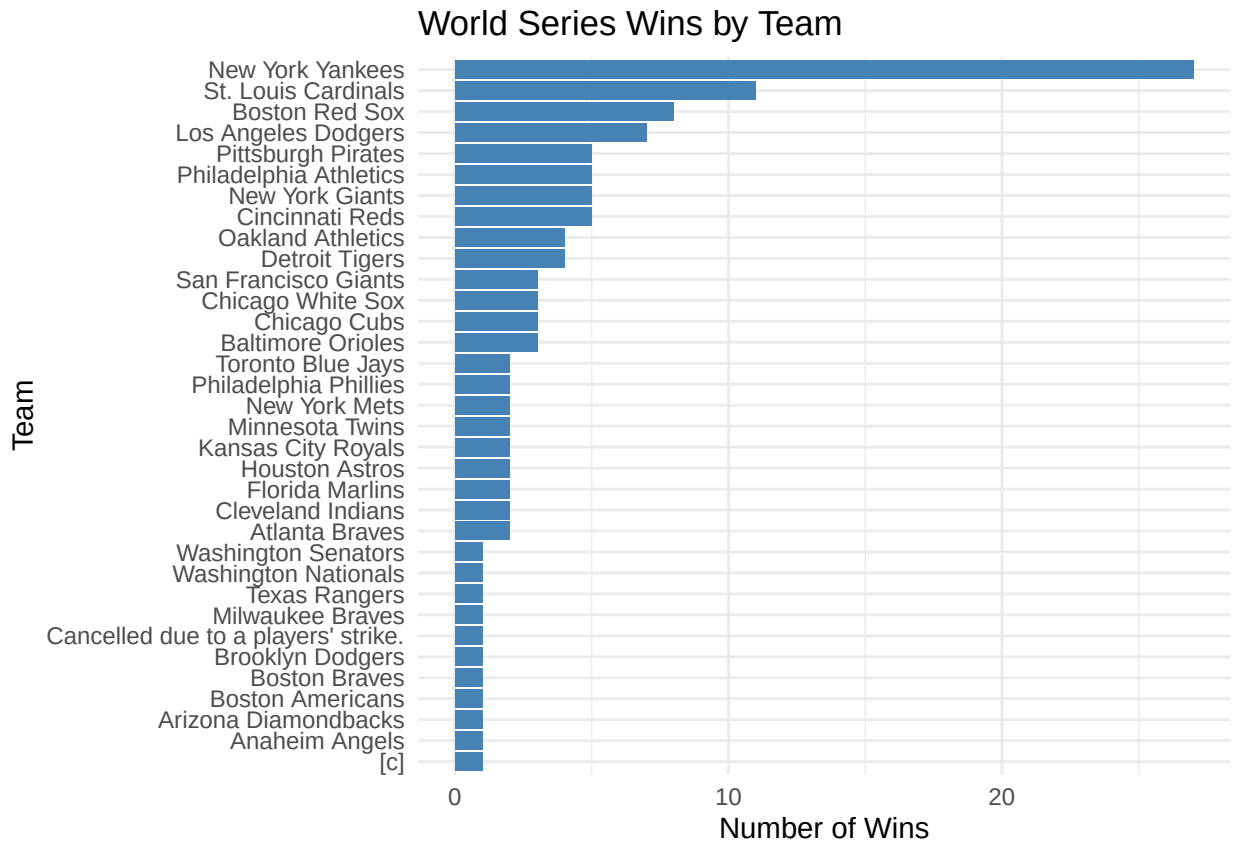
4 Task 3

```
# QUESTION 3
# Count wins (excluding wildcard winners)
wins <- tab1 %>%
  filter(!grepl("\\[W\\]", WinningTeam)) %>%
```

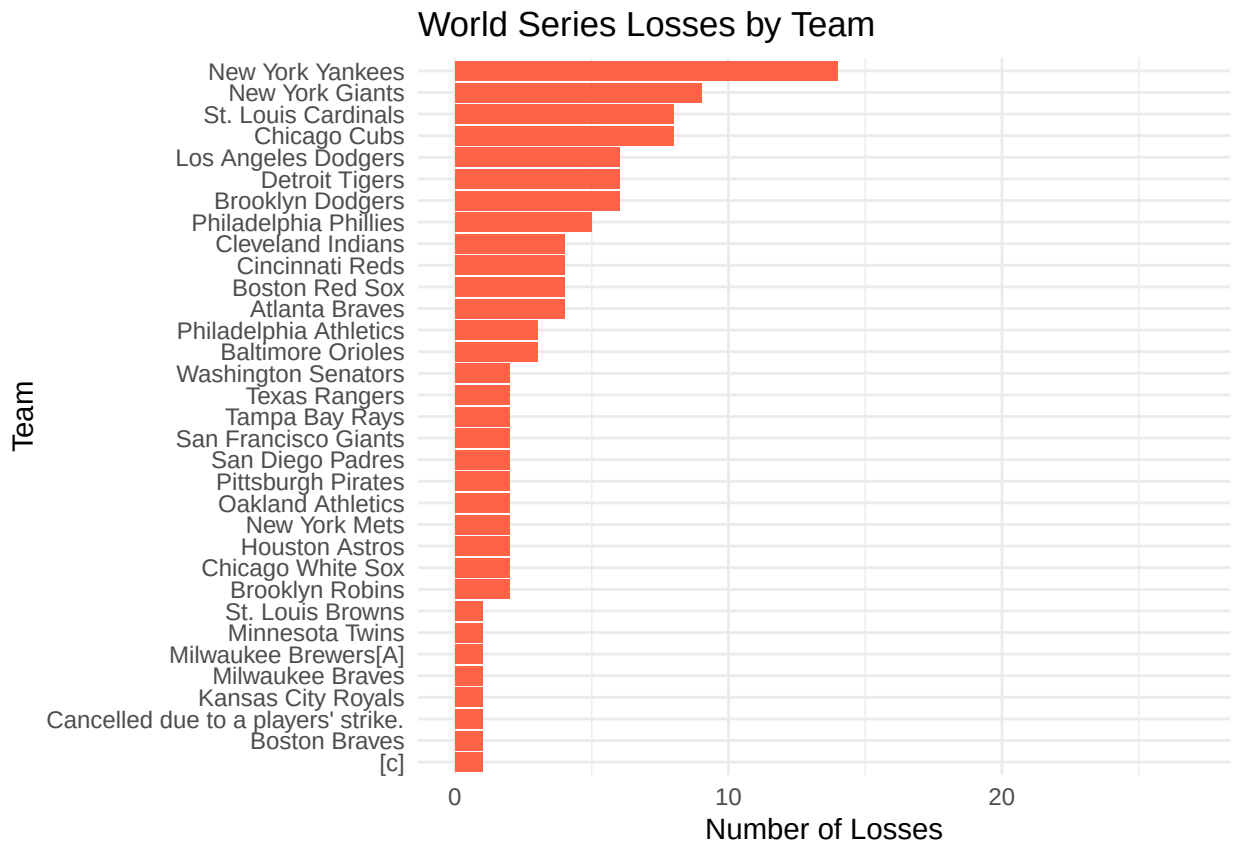
```
  group_by(WinningTeam) %>%
  summarise(Wins = n())

names(tab1)
```

```
## [1] "Year"           "WinningTeam"    "WinningManager" "Series"
## [5] "LosingTeam"     "LosingManager"
```

```
# Count losses
losses <- tab1 %>%
  filter(!grepl("\\[W\\]", LosingTeam)) %>%
  group_by(LosingTeam) %>%
  summarise(Losses = n())

max_val <- max(wins$Wins, losses$Losses)

# Wins barplot
ggplot(wins, aes(x = reorder(WinningTeam, Wins), y = Wins)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  ylim(0, max_val) +
  labs(title = "World Series Wins by Team",
       x = "Team", y = "Number of Wins") +
  theme_minimal()
```



```r
# Losses barplot
ggplot(losses, aes(x = reorder(LosingTeam, Losses), y = Losses)) +
  geom_bar(stat = "identity", fill = "tomato") +
  coord_flip() +
  ylim(0, max_val) +
  labs(title = "World Series Losses by Team",
       x = "Team", y = "Number of Losses") +
  theme_minimal()
```



The two bar charts below highlight teams' performance in World Series final games. The New York Yankees lead both lists with 27 wins and 13 losses, making them the most active and dominant franchise in championship history. Some teams, such as the Cardinals, Dodgers, and Red Sox, appear on both charts, reflecting sustained success over many years. Others, like the Cubs and Giants, appear more frequently on the losses chart, indicating multiple finals appearances but fewer championships.

Overall, the charts show that frequent participation increases both wins and losses: teams that reach the finals often are statistically more likely to appear on both lists. Teams with fewer losses either participate in fewer series or maintain higher winning efficiency during their appearances, highlighting long-term consistency in championship performance.