TermPaper - Tennis

Group 3

Table of contents

Introduction

Describing the Data

Structure of the Data Set

Audience and Research Questions (optional)

Key findings

One of the primary objectives in analyzing this dataset was to explore the question: *What makes a great tennis player?* To address this, we adopted two distinct perspectives. The first focused on match outcomes, aiming to identify which players had the highest number of match victories. We include a section describing the data manipulation process as an example of our coding workflow.

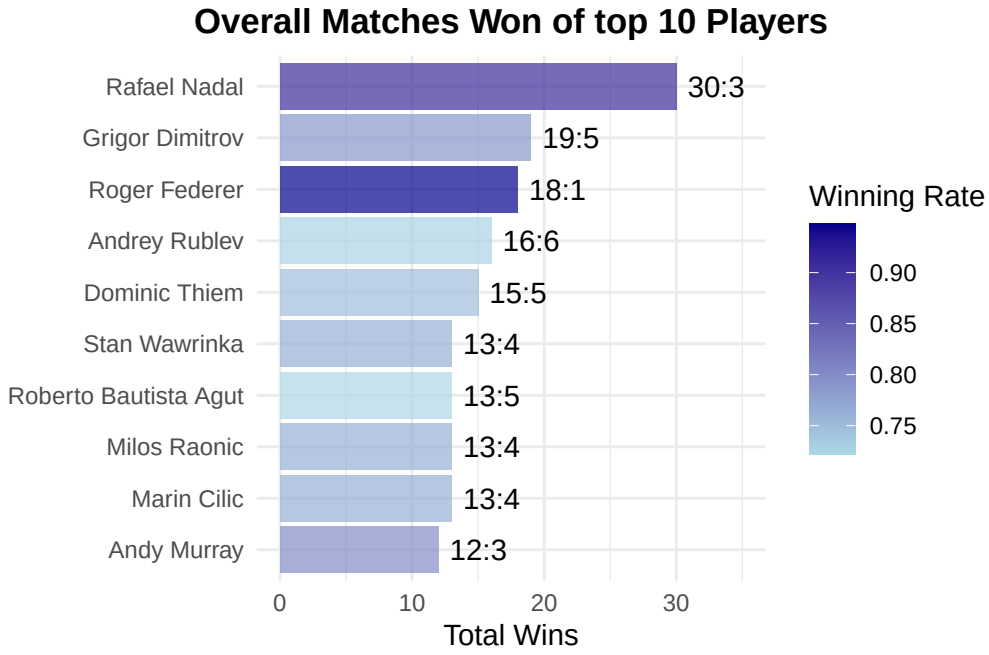
```
# import dataset
tennis <- read.csv2("Dataset3_Tennis.csv")

# calculate total wins using dplyr
totalWins <- tennis %>%
  select(tourney_slug, winner_name) %>%
  count(winner_name) %>%
  rename(player = winner_name) %>%
  arrange(desc(n))

# calculate total losses using dplyr
totalLosses <- tennis %>%
  select(tourney_slug, loser_name) %>%
  count(loser_name) %>%
  rename(player = loser_name) %>%
  arrange(desc(n))
```

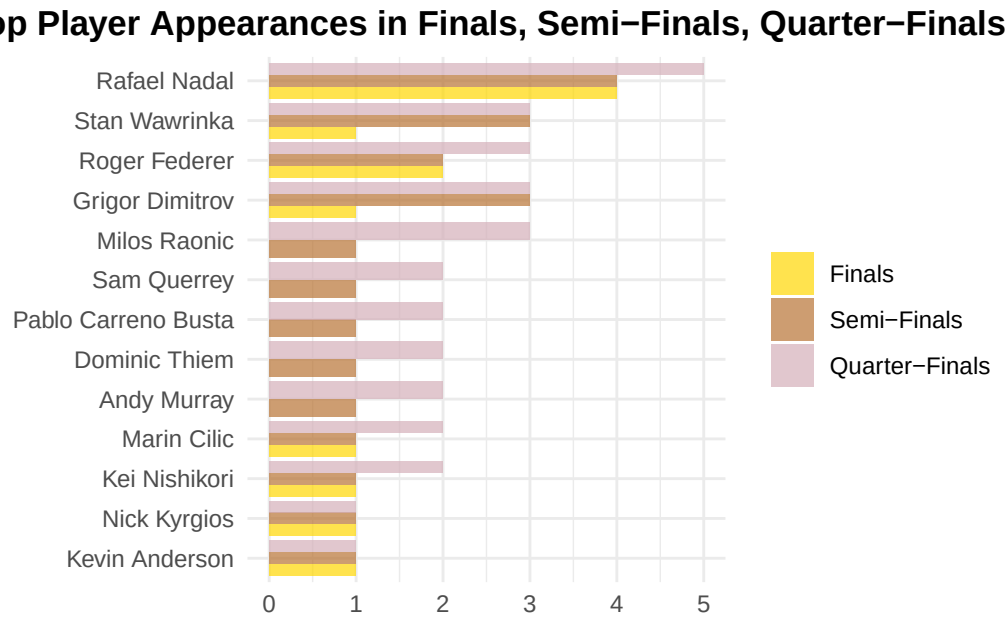
```
# create subset using dplyr
winnerTable <- left_join(totalWins, totalLosses, by = "player") %>%
  mutate(win_loss_score = paste0(n.x, ":", n.y)) %>%
  mutate(win_loss_ratio = n.x/(n.x+n.y)) %>%
  arrange(desc(n.x)) %>%
  head(n = 10)
```

```r
# plot data using ggplot
  ggplot(data = winnerTable,
         aes(x = n.x, y = reorder(player, n.x),
         fill = win_loss_ratio)) +
    geom_bar(stat = "identity", alpha = 0.7) +
    geom_text(aes(label = win_loss_score),hjust = -0.2, vjust = 0.5,size = 4) +
    theme_minimal() +
    labs(title = "Overall Matches Won of top 10 Players",
         x = "Total Wins", y = "") +
    theme(plot.title = element_text(face="bold", hjust = 0.5),
          legend.position = "right") +
    scale_x_continuous(limits = c(0, 35)) +
    scale_fill_gradient(name = "Winning Rate",
                        low = "lightblue",
                        high = "darkblue")
```



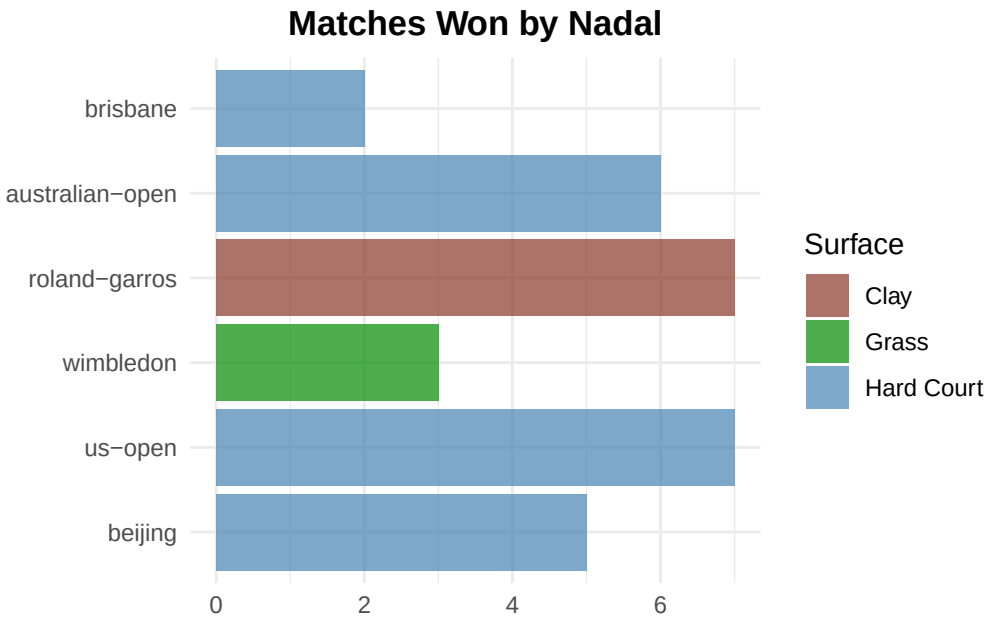
This plot displays the top 10 tennis players with the most match wins. Each bar represents a player's total number of wins, with the bars ordered from most to least wins. The fill color of each bar reflects the player's win-loss ratio—darker blue indicates a higher winning rate. A label on each bar shows the win-loss record (e.g., 30:5). This visualization helps highlight not only who won the most matches but also how dominant they were in terms of win percentage. It is immediately apparent that players such as Nadal and Federer demonstrated a high level of consistency across all recorded matches in the dataset.

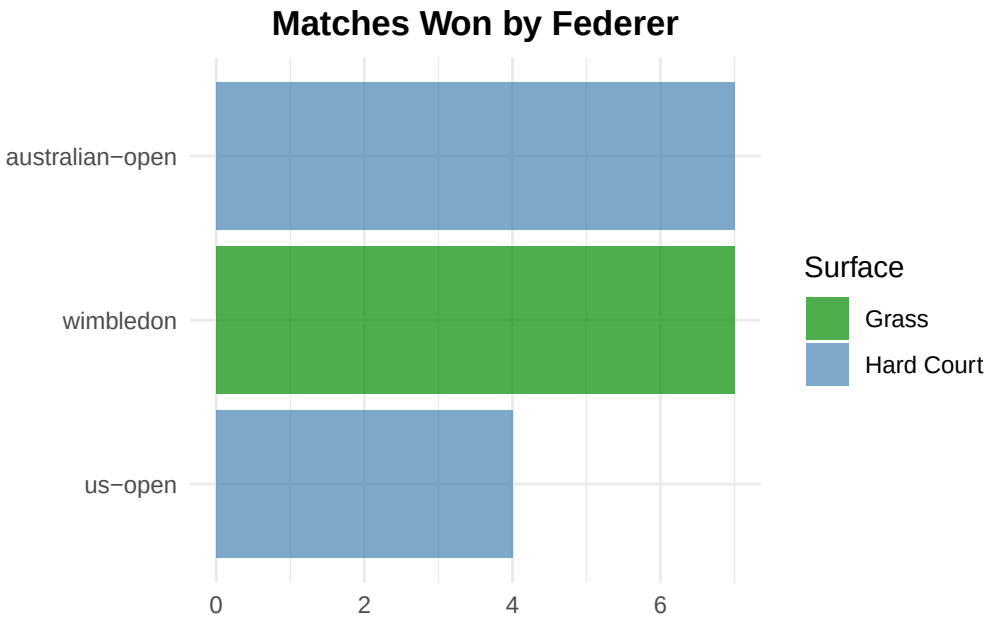
Another way to identify the top-performing players is by examining their appearances in the final rounds of the six tournaments under analysis. The following plot demonstrates that, by this measure as well, the same names consistently emerge as the leading players.



Again, Nadal and Federer are among the best performing players, having reached the most finals.

We also explored whether different court surfaces may have influenced match outcomes, considering that some players perform better on clay, while others excel on grass or hard courts. To illustrate this, we focused again on Nadal and Federer, two of the top-performing players in our dataset. The analysis shows that Nadal dominated on clay courts, whereas Federer secured many of his wins on grass. Both players performed strongly on hard courts. However, while these patterns highlight their individual strengths, they do not offer a clear explanation for *why* these players were so consistently successful overall.

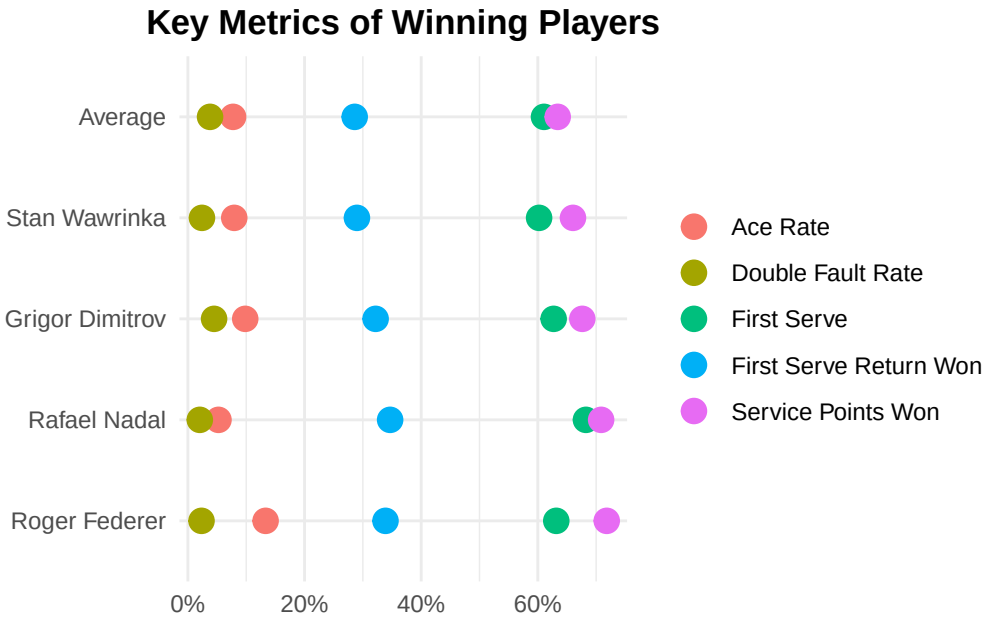




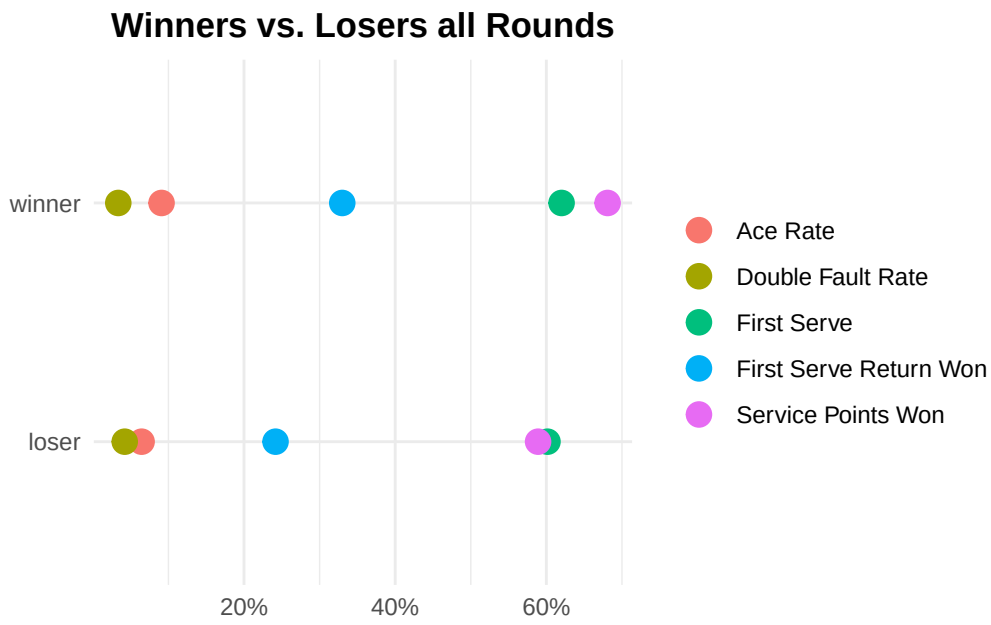
The second perspective was more analytical, involving a comparison of performance metrics between players. The goal was to determine which factors distinguish successful players from those who are less successful. To achieve this, we calculated and compared several key performance indicators, including ace rate, double fault rate, first serve percentage, first serve return

points won, and total service points won. Each of these metrics was expressed as a percentage, allowing for a clear and effective visual comparison across the two groups.

The second perspective can be applied to compare single players or to compare groups of players, e.g. winners and losers. If we are interested in the top player performance again, we can plot the elite players against the average player.



If we divide the field by winners and losers of matches we get the following metrics:



From this analysis, we conclude that metrics such as first serve percentage, service points won, and first serve return success are more indicative of a winning player than ace rate or double fault rate. This suggests that strong serving ability is a fundamental factor in overall player success. This outcome aligns with general tennis knowledge - serving is the only aspect of the game entirely within a player's control, unaffected by the opponent's actions.

It is worth acknowledging that these findings may appear somewhat self-evident. Our analysis ultimately confirms that players with effective serves and solid return games are more likely to winan - insight already familiar to most tennis players. However, the scope of our conclusions was limited by the dataset, which lacked additional informative variables such as player age or break point statistics. As a result, we were constrained to focus on the available performance metrics.

A detailied summary of all attempted plots can be found here in this shinyapp.

Statistical Analysis

Correlation

Regression Model

Hypothesis testing

End

Summary

Winning players demonstrated consistently strong performance across all aspects of their game. Both their serving and returning abilities were above average when compared to losing players. In particular, winners secured significantly more points on their first serves as well as on returns against their opponents' first serves. These two metrics (service points won and first serve return points) emerged as reliable predictors of match success. Contrary to common intuition, the ace rate did not prove to be a particularly strong indicator of whether a player would win.

Limitations

This analysis is subject to several important limitations. First, the dataset includes only match-level data from a single year (2017) and covers just six ATP tournaments. Given that the ATP World Tour comprises over 60 tournaments annually, this selection represents only a small fraction of the full tour calendar. As a result, the findings may not be representative of player performance across the broader tour or over time.

Second, the limited time frame restricts the possibility of longitudinal analysis or year-over-year comparisons. Certain top-ranked players, such as Novak Djokovic, appear to be absent from the dataset, possibly due to injury or non-participation in the selected tournaments. Including data from all four Grand Slam tournaments over multiple years would likely allow for more comprehensive and generalizable insights.

Third, the dataset lacks several key performance metrics commonly used in tennis analysis, such as the number of winners, unforced errors, and break points won. While tiebreak information is included, it is less informative than break point data, which is often considered a critical indicator of match dynamics and player resilience under pressure.

Additional Remarks