Analysis of Tennis Match Data

An In-Depth Look at Player Performance and Match Statistics

Group 5

July 2024

Introduction

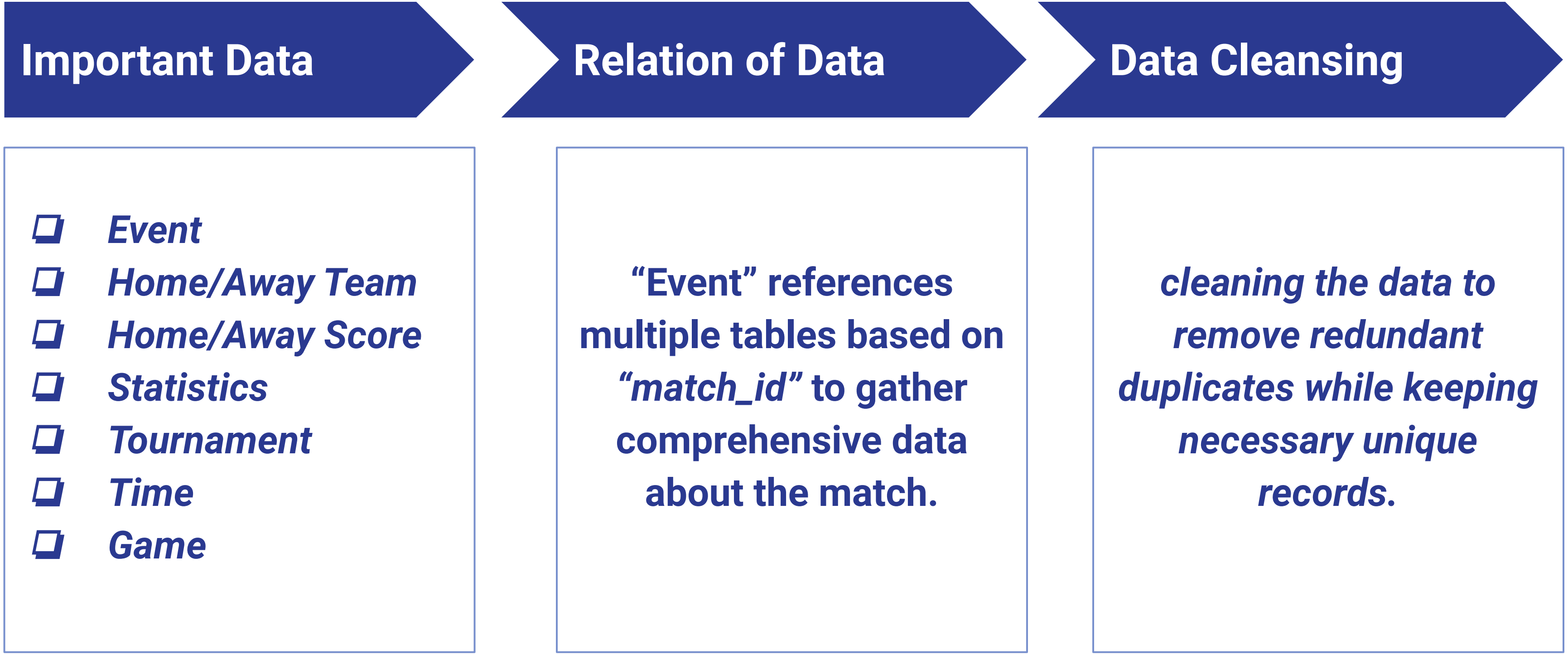
Objective

Analyzing tennis match data to uncover patterns and insights related to player performance and match outcomes.

Data sources

The dataset includes player statistics, match results, and various match-specific metrics from the Parquet file for professional tennis tournaments in May 2024.

Data Overview



Important Data

- ❑ ***Event***
- ❑ ***Home/Away Team***
- ❑ ***Home/Away Score***
- ❑ ***Statistics***
- ❑ ***Tournament***
- ❑ ***Time***
- ❑ ***Game***

Relation of Data

“Event” references multiple tables based on *“match_id”* to gather comprehensive data about the match.

Data Cleansing

cleaning the data to remove redundant duplicates while keeping necessary unique records.

Data Preparation Challenges

- Extract entire files without the risk of data loss.

- Extract all the files related to a schema , concat them together, save them in a new Parquet file and use it as a dataframe.

- Invalid data (related to winner code of event schema)

- Existence of duplicate records based on match_id and non-duplicate in other fields.

Solution

- Removed duplicates records.
- Add a date column to all datasets, sort data by date field, use the latest data, and ignore other records.
- Implement methods to extract valid result.

Analysis Result

Player Analysis

1. Number of unique tennis players : 2,352

2. Average height of the players : 1.82

13. Distribution of left-handed versus right-handed players

10. Correlation between a player's height and their ranking

S : Home/Away team

drop_duplicates(keep='last')

dropna()

Player Analysis

3. Player with the highest number of win: Sun Fajing with 15 wins.

Event (winner_code) ✘

S : Method based on pbp ✔

O : Winners

Player Analysis

9. Player with the highest number of tournament wins : Uchijima Moyuka with 15 wins.

14. The most common tournament surface is red clay.

S : Winners

S : Tournament

drop_duplicates(keep='last')

Player Analysis

16. Player with the highest winning percentage against top 10 ranked opponents : Grigor Dimitrov.

S : Winners

S : Home/Away team

Match Analysis

4. The match with the longest duration lasting 47.54 hours.

5. The typical number of sets is 2.

11. The average duration of matches is 124.25 minutes.

S : time

drop_duplicates(keep='last')

skipna=True

mode

Match Analysis

7. The average number of aces per match is 4.07.

8. The average double fault based on gender

17. The average number of breaks of serve per match is 7.29.

S : statistics

S : Home/Away team

drop_duplicates(keep='last')

Match Analysis

12. The average number of game per set based on gender

S : pbp

S : Home/Away team

drop_duplicates(keep='last')

Country Analysis

15. Number of unique countries in dataset is 99.

6. Country with most successful tennis players

S : Home/Away team

S : winners

drop_duplicates()

Count the number of wins for each country

Find top 100 ranked players with their country

Additional Findings

Statistics related to players

1. ***Only one Iranian player : Moghimi, Sina has participated in this competition in May 2024.***
2. ***He has won 3 of the 6 matches he has participated in.***
3. ***Players have a career spanning more than 15 years.***
4. ***Distribution of players' weights***

S : Home/Away team

S : Winners

drop_duplicates(keep='last')

dropna()

Statistics related to tournament

5. Tournament with the highest number of participating players is French Open, Paris, France with 130 players.

S : Tournament

```
drop_duplicates(keep='last')
```

Statistics related to matches

6. The number of matches decided by a tie-break is: 1413

7. Average number of games per match

8. Player with the most double faults in a single match is Braynin, Aleksandr with 48 fault.

9. Cities which hosted the most tennis matches

S : pbp

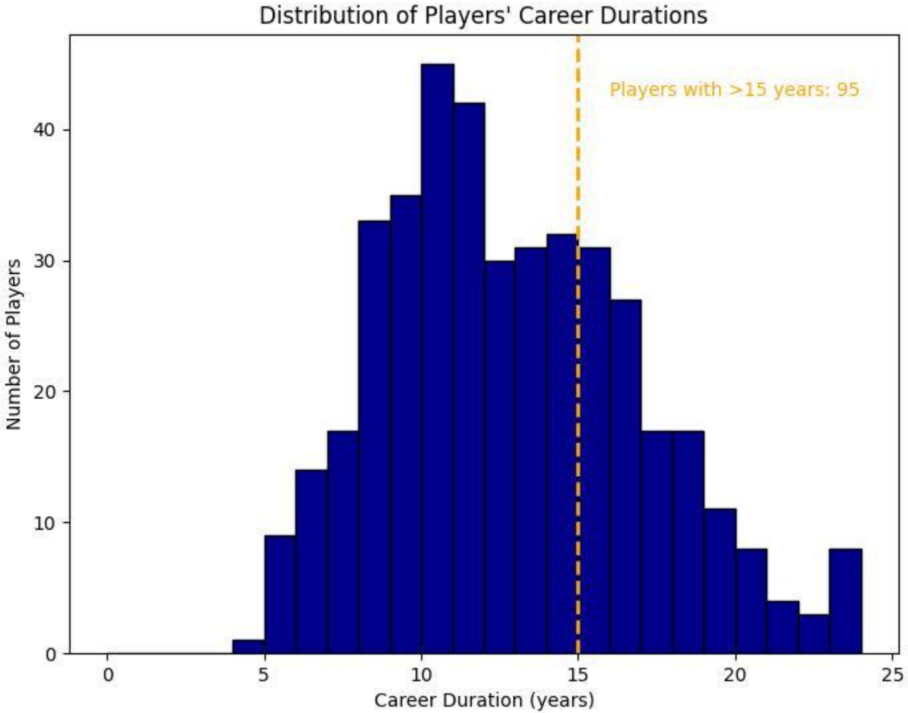
S : Statistics

S : Home/Away team score

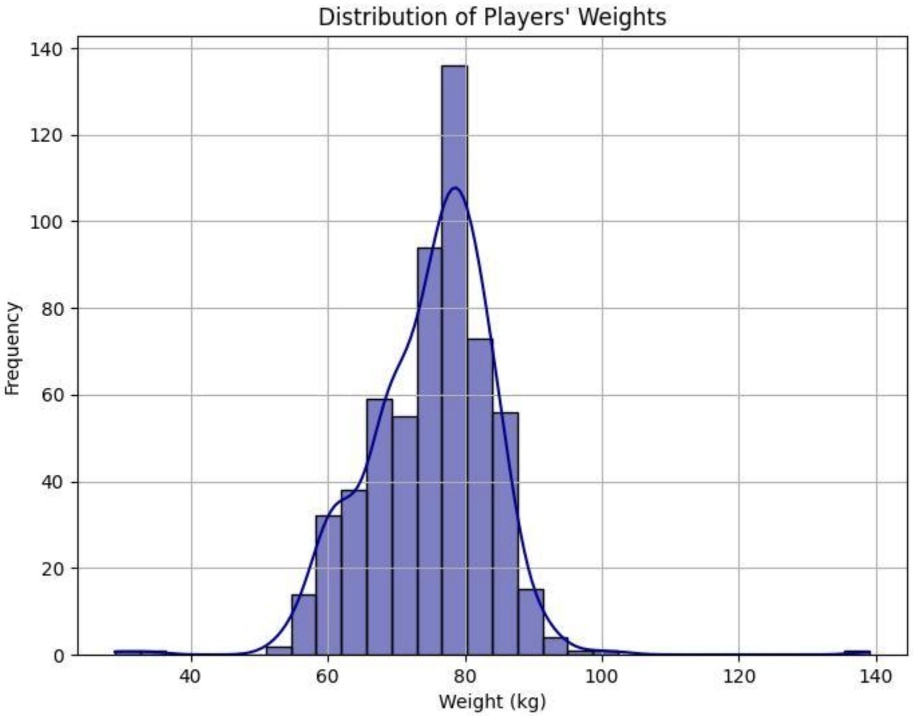
drop_duplicates(keep='last')

dropna()

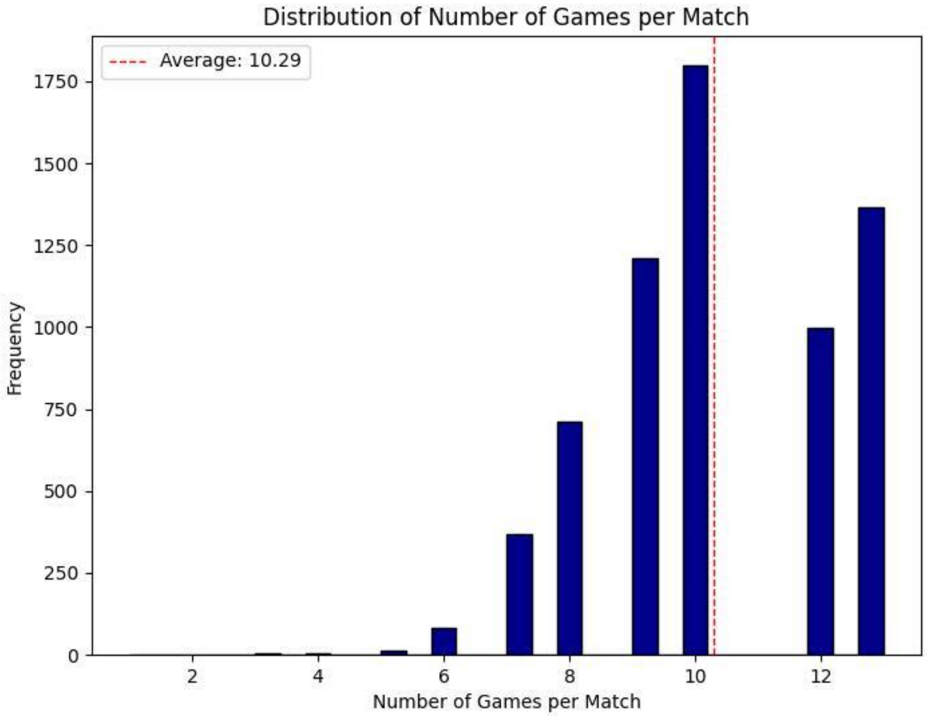
Distribution of players' career durations



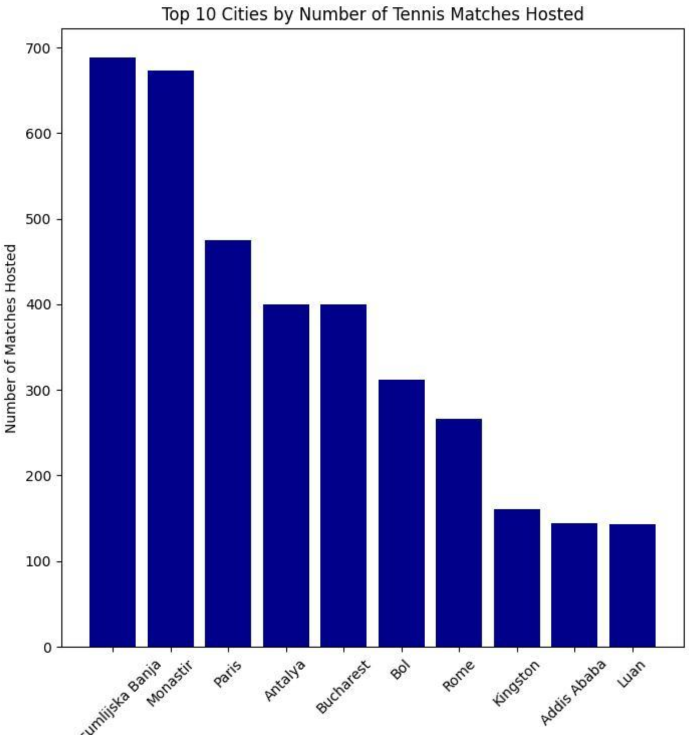
Distribution of players' weights in the dataset



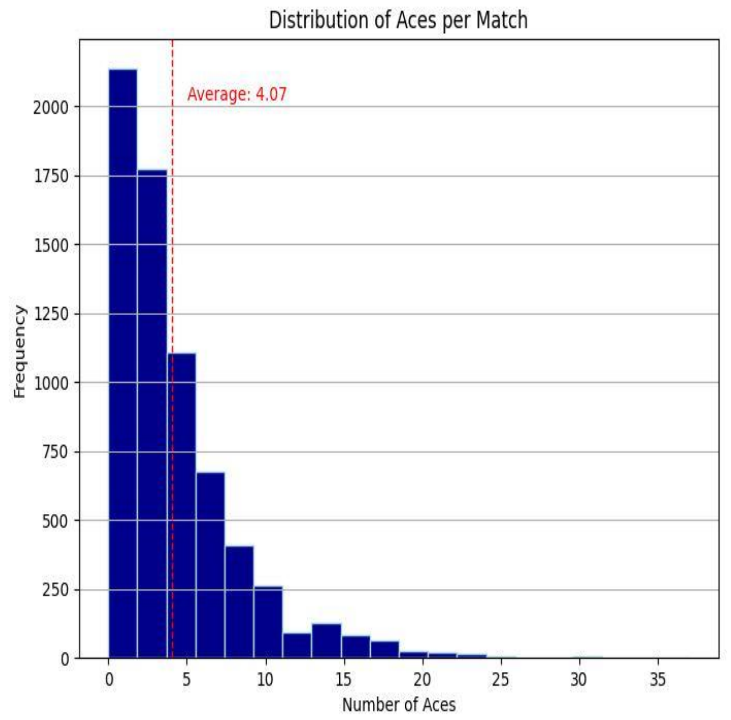
Average number of games per match



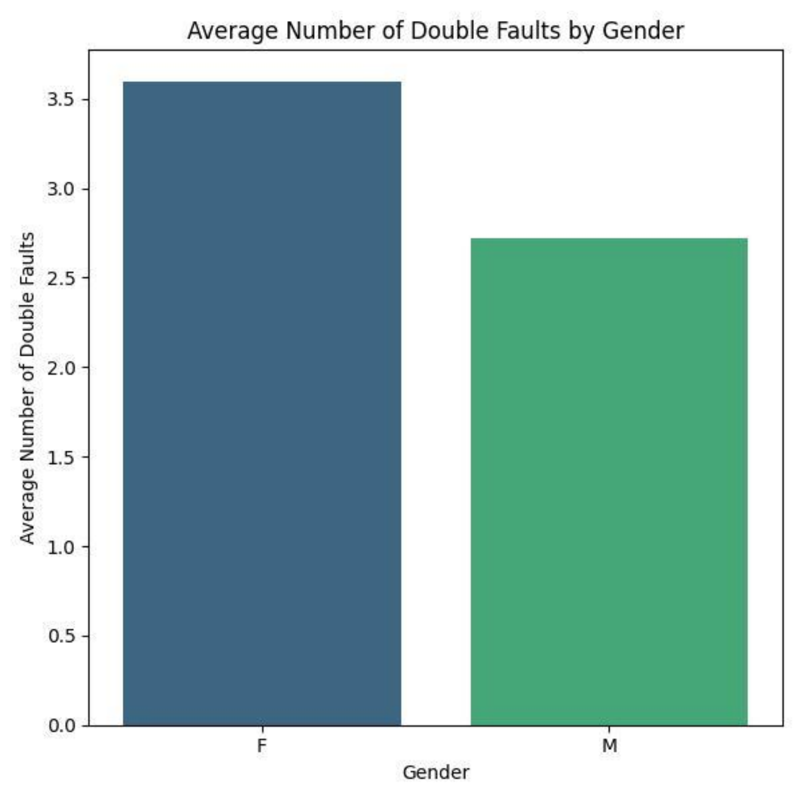
Cities which hosted the most tennis matches



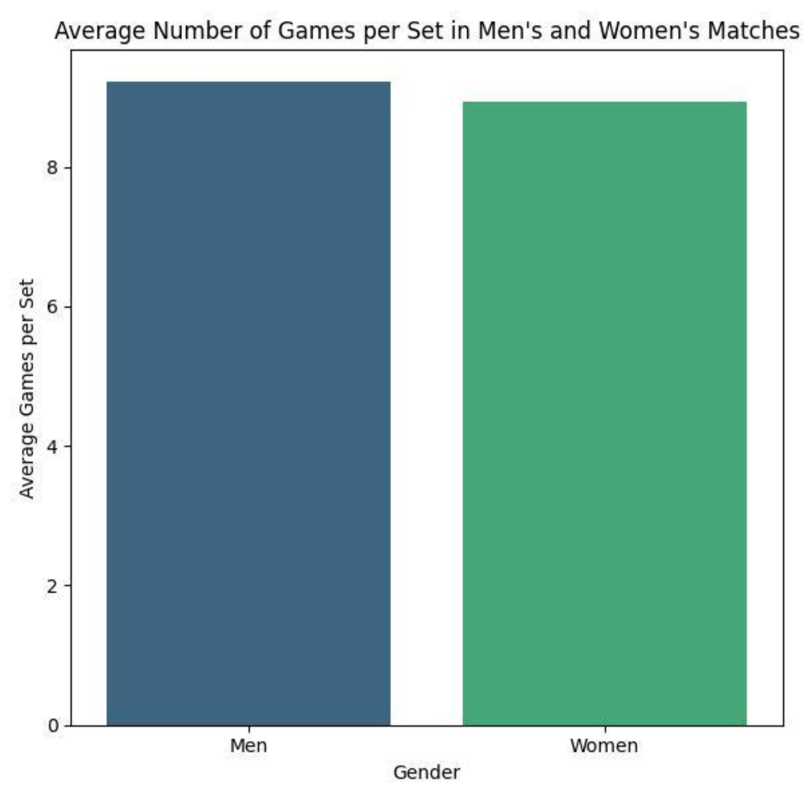
Average number of aces per match is 4.07



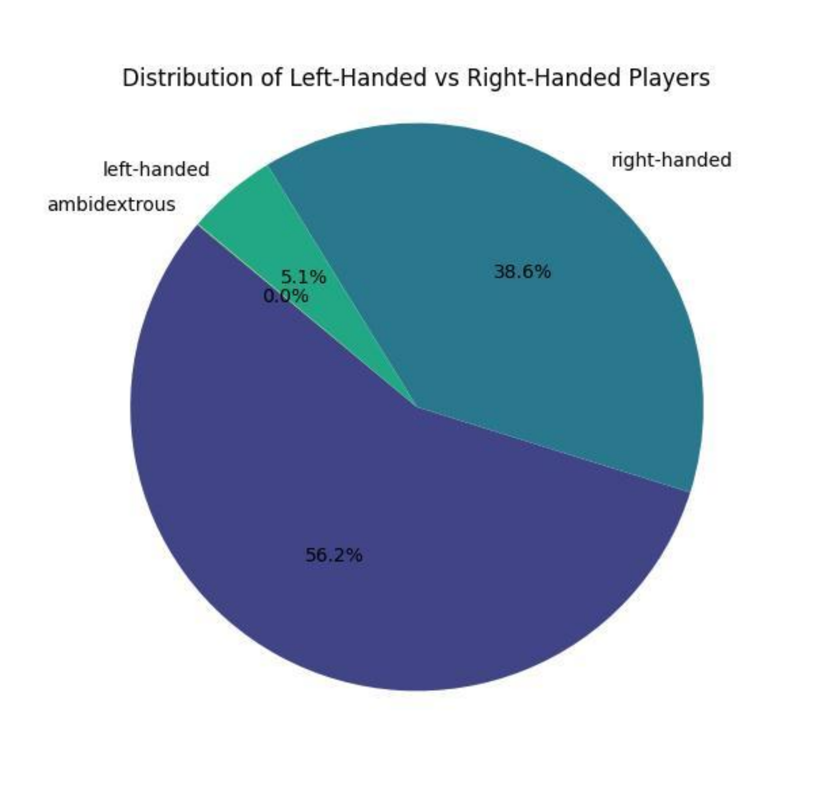
The average number of double faults for females is 3.58 and for males is 2.71.



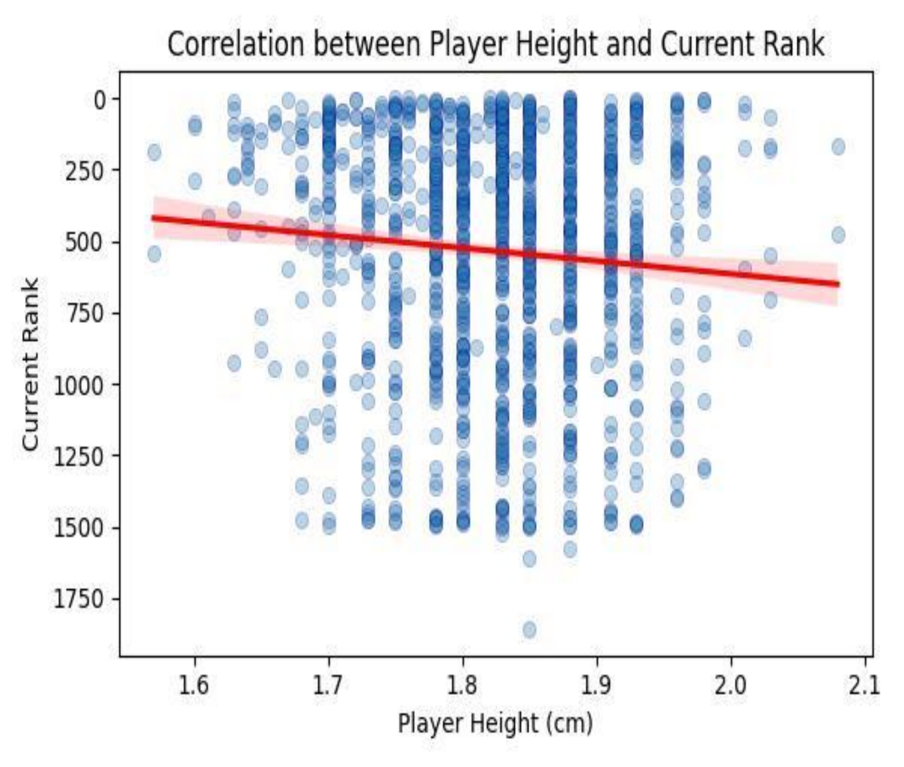
The average number of games per set in men's matches is 9.22.
The average number of games per set in women's matches is 8.93.



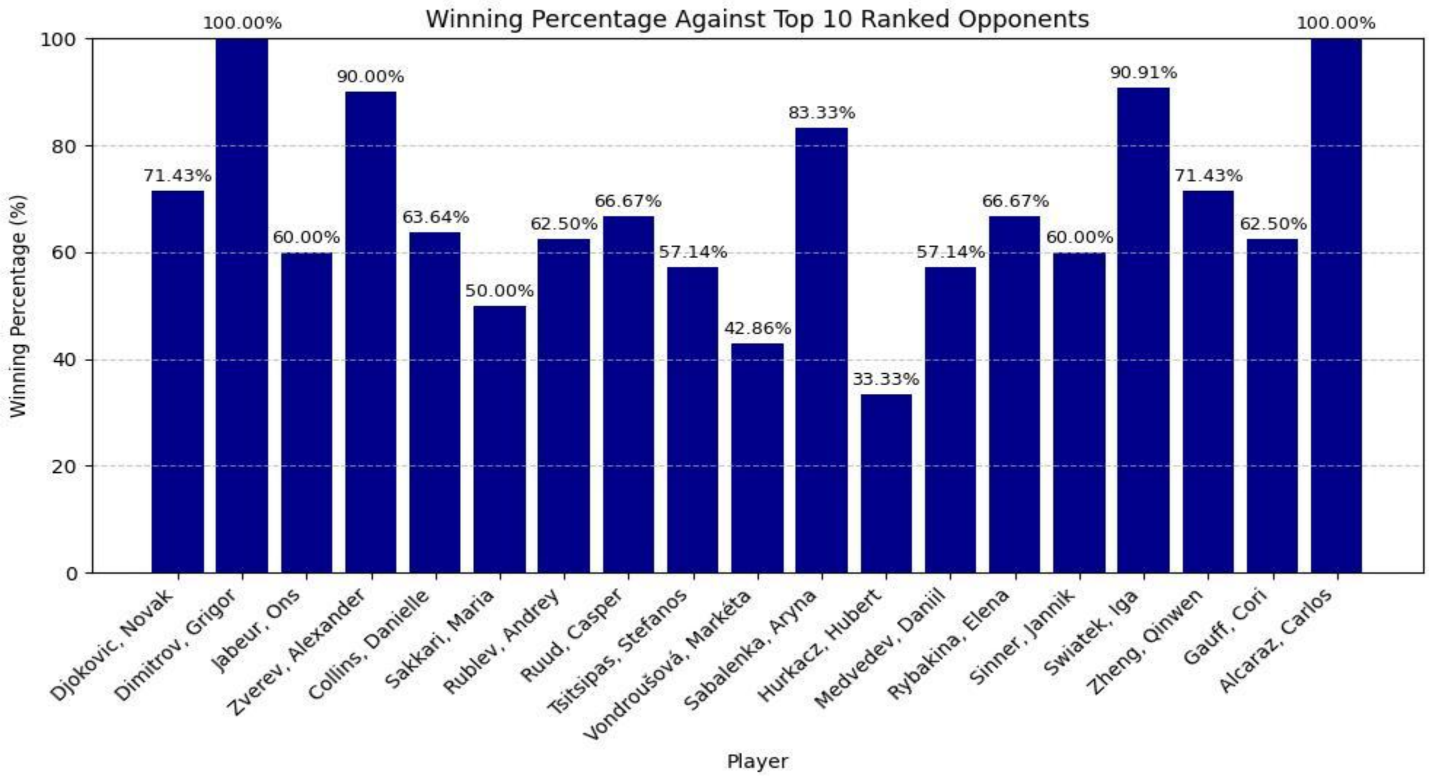
Distribution of left-handed versus right-handed players



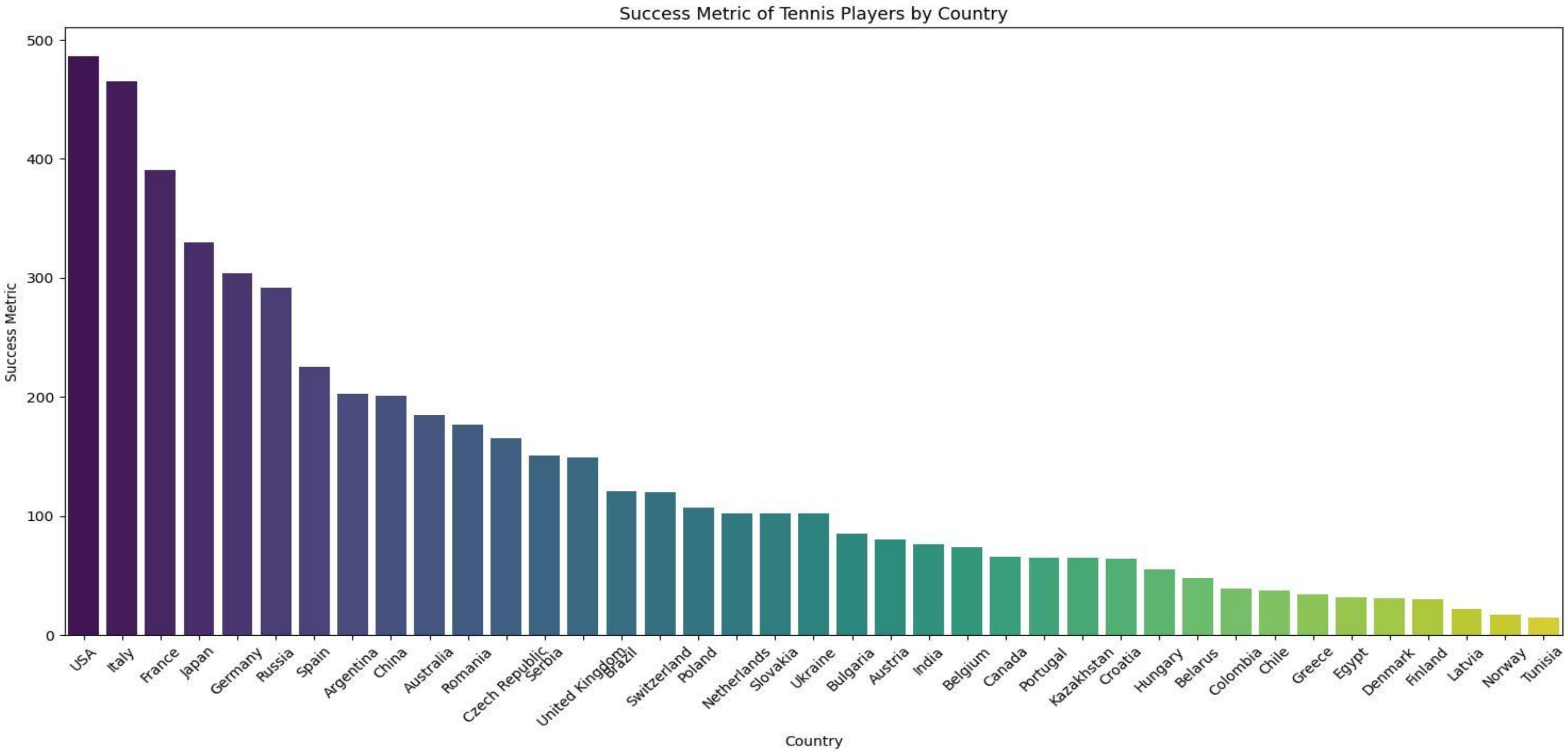
Correlation between a player's height and their ranking



Which player has the highest winning percentage against top 10 ranked opponents?



Data set includes 99 unique countries and the USA is The country with the most successful tennis players with a success metric of 516 (462 wins in May 2024 and 24 top 100 players).



Thanks for your attention