# Tennis Match Data Analysis

7th group:
Maedeh Bank
Zeinab roshandel

# About data

- one zip file which contains the data of tennis matches have been received
- It contains the data of 31 days in zip format. And each day data contains many parquet files
- Firstly we extract data by coding in python
- Then we read data and save it into 15 dataframes which named as dbdiagram table names
- Then the data frames have been saved into 15 csv files for using in solving the questions
- Following the results of solving the question are presented

# Question 1

**How many tennis players are included in the dataset?**

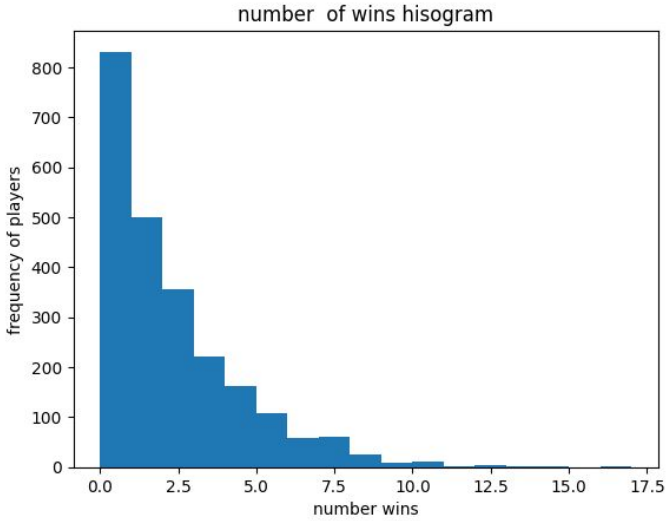
- For this data we need MatchHomeTeamInfo and MatchAwayTeamInfo tables
- By concating and dropping duplicated data the following results have been gained

```
Total number of tenis players equalls to 2352
51.0 % for men and 48.0% for women
```

# Question 3

**Which player has the highest number of wins?**

- Using MatchHomeTeamInfo and MatchAwayTeamInfo and MatchEventInfo tables
- Number of wins histogram has been plotted and it shows *Exponential* behavior




```
The highest number of wins is 17 whitch is gained by Uchijima M. with 253356 ID
```

# Question 5

**How many sets are typically played in a tennis match?**

- GameInfo has been used( PowerInfo can be used too)
- Mode and mean are equal
- The mean says that

```
Tenis matches has typiccally 2.3 sets
```

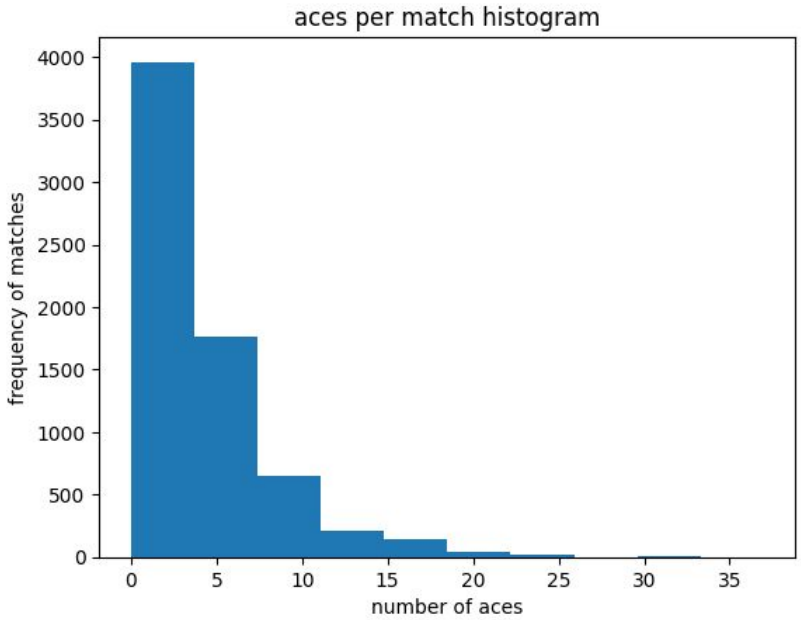
```
70.0 % of matches with 2 sets and 30.0% with 3 sets

set_id
2    4518
3    1920
4      29
5      20
```

# Question 7

- **What is the average number of aces per match?**
- PeriodInfo  table has been used. Counting home_stat and away_stat where stattics_name and period columns equal to ‘Aces’ and ‘ALL’ in order
- *Exponential* behavior

```
4.0 is the average of aces in each match
```



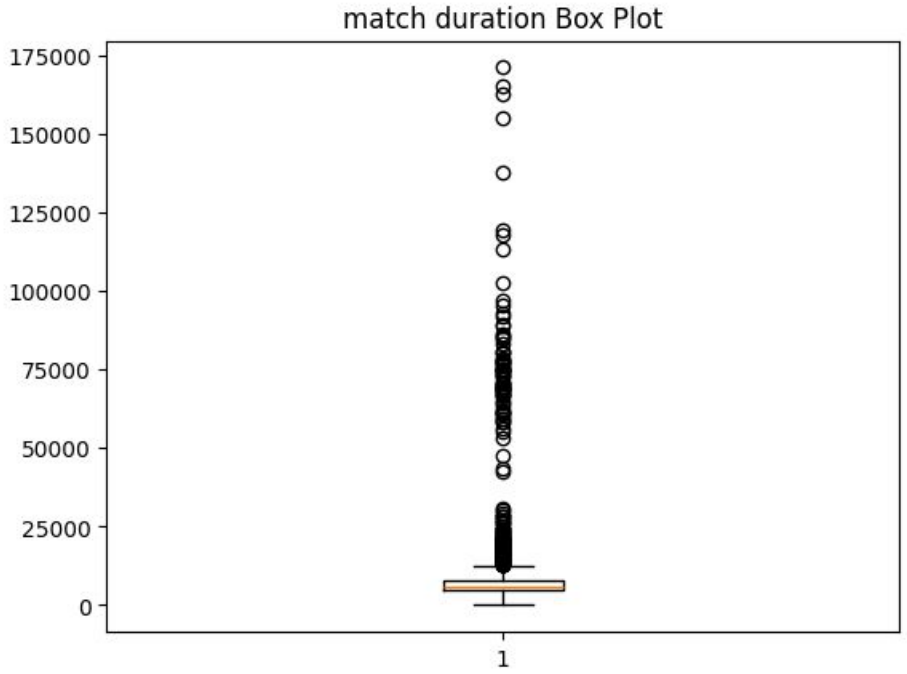
# Question 9

- **Which player has won the most tournaments in a single month?**
- Using MatchHomeTeamInfo and MatchAwayTeamInfo and MatchEventInfo and MatchTournamentInfo
- The winner of last match (according to start time) in each tournament is the winner of the tournament

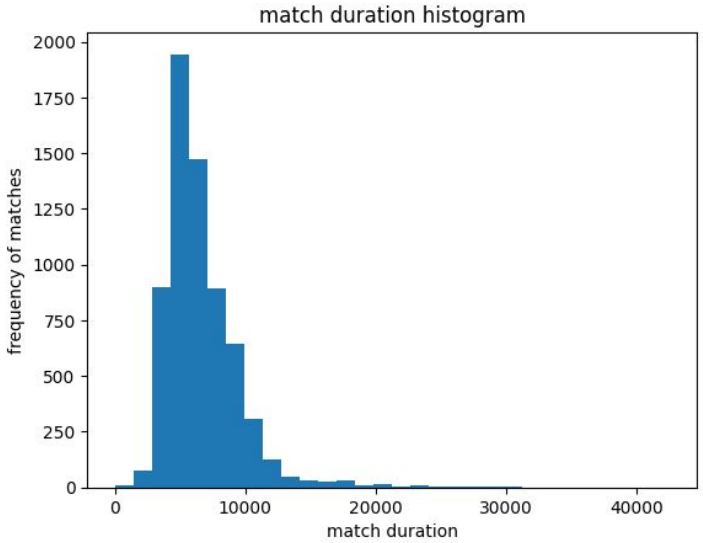
```
Player Świątek I. with 228272 ID has won 3 tournomernt which is most tournaments in a single month
```

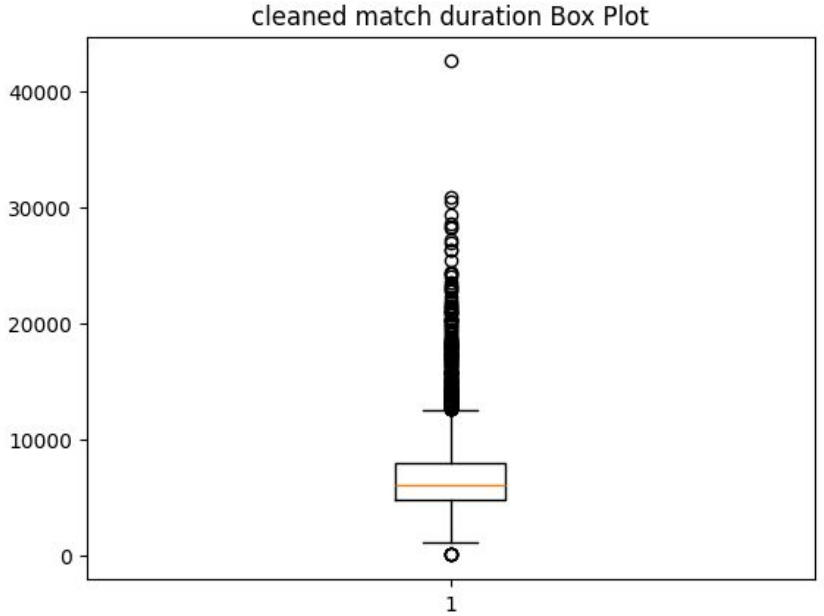
# Question 11

- **What is the average duration of matches?**
- MatchTimeInfo  table has been used. Dropping the match_ids which their period 1 and period 2 values are null and duplicated data
- Invalid Duration time dropped (more than 12 hours)



# Question 11

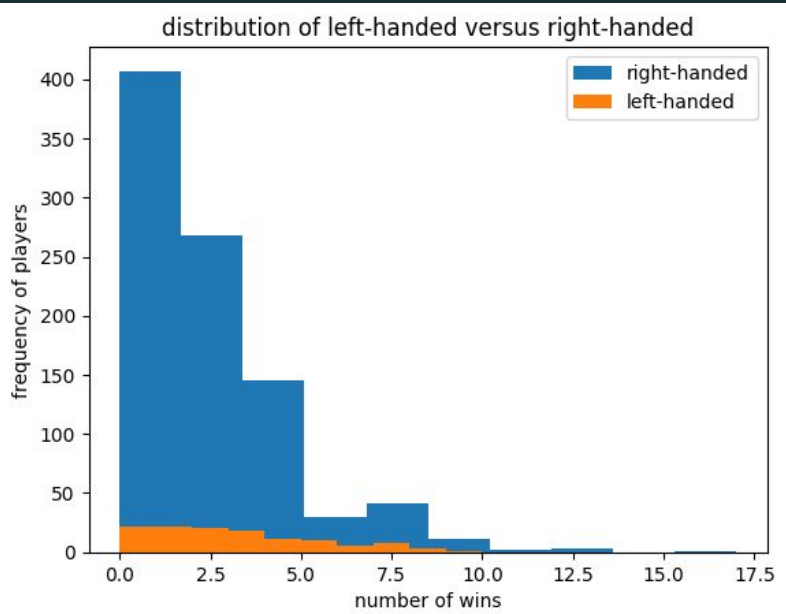



```
The average of match duration is 6625.516079865874
```

# Question 13

- **What is the distribution of left-handed versus right-handed players?**
- Using MatchHomeTeamInfo and MatchAwayTeamInfo and MatchEventInfo tables




```
12.0% of players are left_handed
88.0% of players are right_handed
```

Frequency table of right-handed players

| | bin | frequency |
|---|---|---|
| 0 | (0.0 , 1.7) | 407 |
| 1 | (1.7 , 3.4) | 268 |
| 2 | (3.4 , 5.1) | 145 |
| 3 | (5.1 , 6.8) | 30 |
| 4 | (6.8 , 8.5) | 41 |
| 5 | (8.5 , 10.2) | 11 |
| 6 | (10.2 , 11.9) | 2 |
| 7 | (11.9 , 13.6) | 3 |
| 8 | (13.6 , 15.299999999999999) | 0 |
| 9 | (15.299999999999999 , 17.0) | 1 |

Frequency table of left-handed players

| | bin | frequency |
|---|---|---|
| 0 | (0.0 , 1.0) | 22 |
| 1 | (1.0 , 2.0) | 22 |
| 2 | (2.0 , 3.0) | 20 |
| 3 | (3.0 , 4.0) | 18 |
| 4 | (4.0 , 5.0) | 11 |
| 5 | (5.0 , 6.0) | 10 |
| 6 | (6.0 , 7.0) | 5 |
| 7 | (7.0 , 8.0) | 8 |
| 8 | (8.0 , 9.0) | 3 |
| 9 | (9.0 , 10.0) | 1 |

# Question 15

- **How many distinct countries are represented in the dataset?**
- Using MatchHomeTeamInfo and MatchAwayTeamInfo and MatchVenueInfo tables. After dropping duplicated in merged MatchHomeTeamInfo and MatchAwayTeamInfo table

```
97 distinct countries are represented in as player country
34 distinct countries are represented in as vanue country
99 distinct countries are represented in the total dataset
```

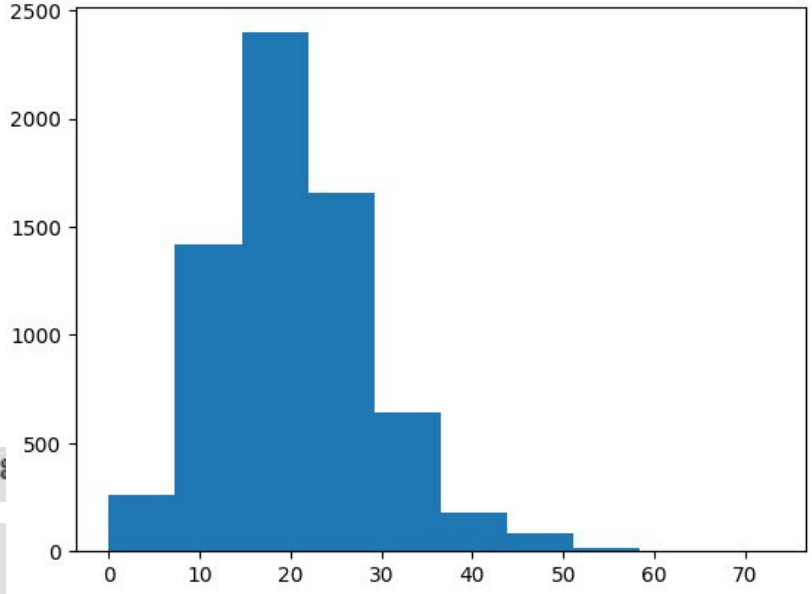
# Question 17

**What is the average number of breaks of serve per match?**

- Breaks of serve means win a game in which another player is serving
- Table PowerInfo has been used it represents  break occur in each game

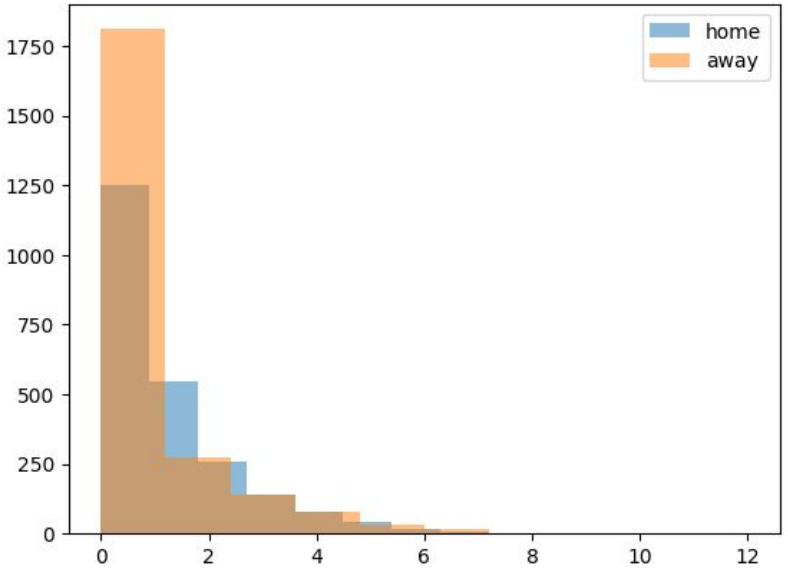
```
The average number of breaks of serve per match is 20.28
```

```
Shapiro-Wilk Test for Math Score: Statistic = 0.97, p-value = 0.0
T-statistic: 0.97, p-value: 0.0
Sample does not look Gaussian (reject H0)
```



# Additional Question no 1

- **Is there a difference in the percentage of wins based on hostage?**
- Using MatchHomeTeamInfo and MatchAwayTeamInfo and MatchEventInfo tables.
- The distribution of the win's number in home and away state have been




```
T-statistic: 1.2064401313236124, p-value: 0.22770852617328424
Samples dose not show significant difference  (fail to reject H0)
so there is not significant difference between the number of wins and hostage
```

# Additional Question no 2

- **Which tournament has the highest number of votes?**
  - Using MatchTournamentInfo and MatchVotesInfo
  - Adding home and away votes

```
Tournament Rome, Italy has max votes equals to 413648
Average of votes for each tournament is 10153.505084745762
Tournament Antalya, Singles Qualifying, W-ITF-TUR-17A has max votes equals to 10
```

# Additional Question no 3

- **The percentage of games have been played by women?**
- Using MatchHomeTeamInfo and MatchAwayTeamInfo and MatchEventInfo tables.

```
0.48 % of matches have been played by women
```

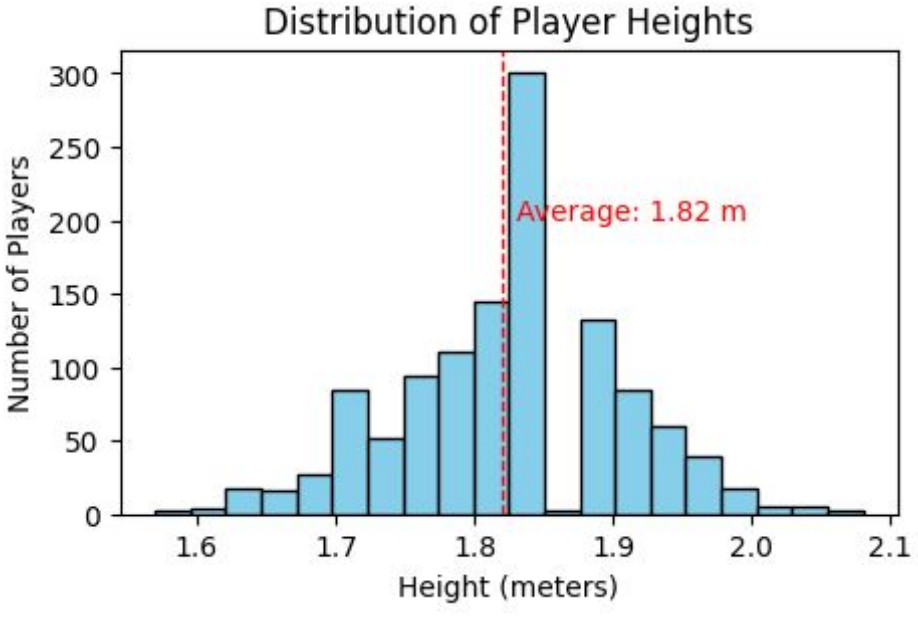
# Additional Question no 4

- **How many games are played in each stadium on average?**
- Using MatchHomeTeamInfo and MatchAwayTeamInfo and MatchVenueInfo tables.

```
the average of games played in each stadium is 59.909
```

# Query #2: What is the average height of the players?

- Table Used: MatchAwayTeamInfo and MatchHomeTeamInfo (concat: “players_info”)
- The duplicates of “players_info” dropped based on “player_id”
- Result:
  - **The average height of the players is: 1.82 m**
- The average height of 1.82 meters indicates that professional tennis players tend to be relatively tall.



# Query #4: What is the longest match recorded in terms of duration?

- Table Used: MatchTimeInfo
- The “total_duration” column generated by summing the periods columns.
- Initial result was unrealistic: approximately 23.9 hours!!
- Added a sanity check of 12 hours
- The longest recorded tennis match in history: 11 hours and 5 minutes!
- More reasonable result:
  - Match ID: 12366196.0
  - Duration: 42551.0
- Approximately 11.8 hours. So close to the history record!

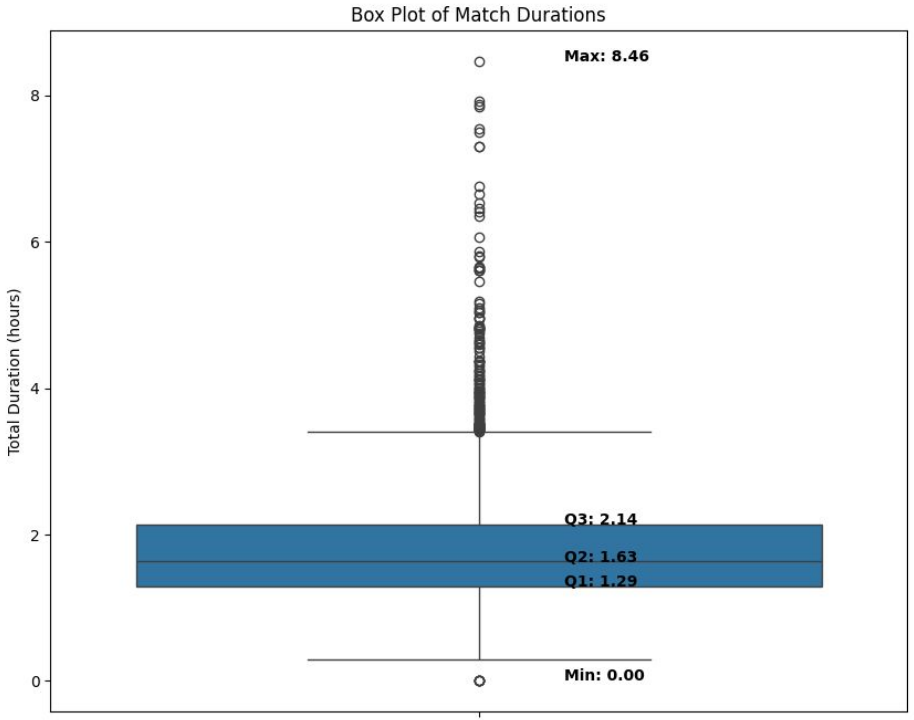
```
MatchTimeInfo
✓ 0.0s
```

| | match_id | period_1 | period_2 | period_3 | period_4 | period_5 | current_period_start_timestamp | total_duration |
|---|---|---|---|---|---|---|---|---|
| 0 | 12260075 | 3463.0 | 3855.0 | NaN | NaN | NaN | 1.714511e+09 | 7318.0 |
| 1 | 12260076 | 3032.0 | 2121.0 | 2674.0 | NaN | NaN | 1.714492e+09 | 7827.0 |
| 2 | 12260077 | 2747.0 | 3525.0 | 4074.0 | NaN | NaN | 1.714492e+09 | 10346.0 |
| 3 | 12260078 | 2519.0 | 2531.0 | 2121.0 | NaN | NaN | 1.714578e+09 | 7171.0 |
| 4 | 12260080 | 2616.0 | 2766.0 | NaN | NaN | NaN | 1.714483e+09 | 5382.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 19671 | 12384975 | 2218.0 | 4709.0 | 7140.0 | NaN | NaN | 1.717248e+09 | 14067.0 |
| 19672 | 12385017 | 3295.0 | 4903.0 | 4413.0 | NaN | NaN | 1.717258e+09 | 12611.0 |
| 19673 | 12385869 | NaN | NaN | NaN | NaN | NaN | NaN | 0.0 |
| 19674 | 12385873 | NaN | NaN | NaN | NaN | NaN | NaN | 0.0 |
| 19675 | 12386383 | NaN | NaN | NaN | NaN | NaN | NaN | 0.0 |

19676 rows × 8 columns

# Query #4: What is the longest match recorded in terms of duration?

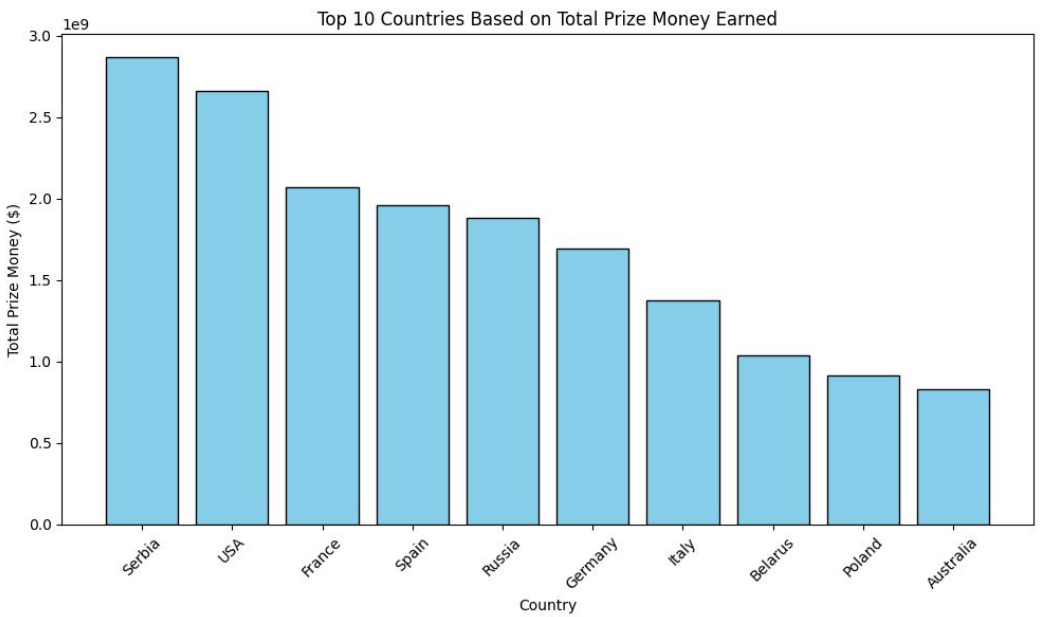
- Table Used: MatchTimeInfo
- The “total_duration” column generated by summing the periods columns.
- Initial result was unrealistic: approximately 23.9 hours!!
- Added a sanity check of 12 hours
- The longest recorded tennis match in history: 11 hours and 5 minutes!
- More reasonable result:
  - **Match ID: 12346456**
  - **Duration: 30465.0 seconds, equal to 8.4625 hours**



# Query #6: Which country has produced the most successful tennis players? (Prize Approach)

- Table Used: MatchHomeTeamInfo and MatchAwayTeamInfo
- “Total_prize_per_country” is generated using groupby function
- Result:

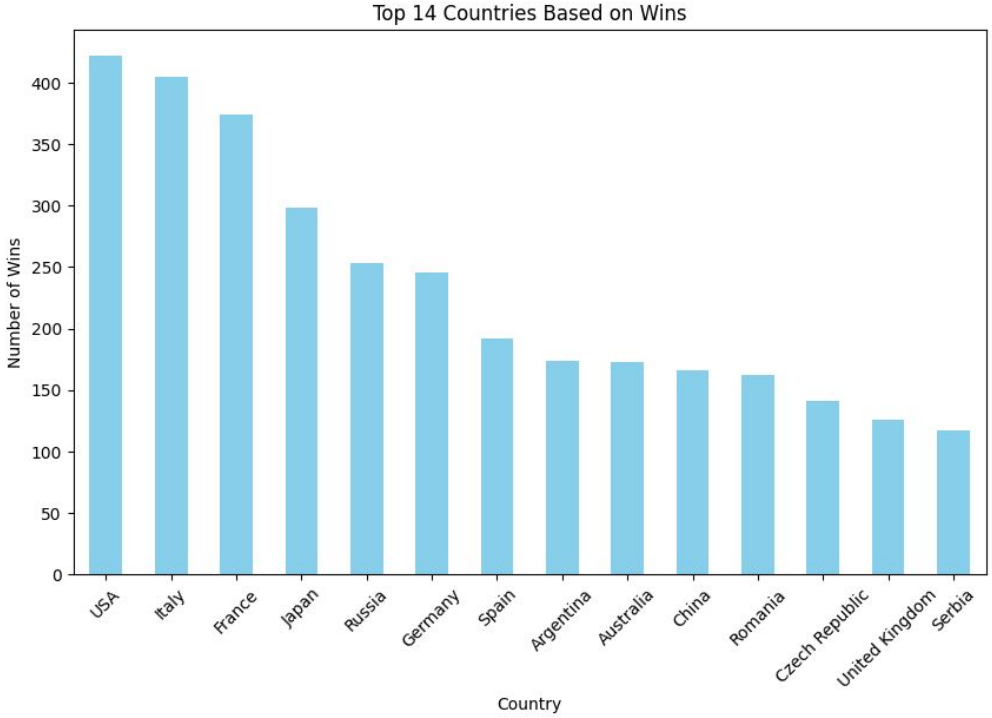
**“The country that has produced the most successful tennis players based on total prize money earned is Serbia with total prize money of $2864742758.”**



# Query #6: Which country has produced the most successful tennis players? (Win Counts Approach)

- Table Used: MatchHomeTeamInfo, MatchAwayTeamInfo and MatchEventInfo
- “Win_counts” series is generated using “value_counts()” function on merged dataframe
- Result:

**The country with the maximum wins is USA with 422 wins.**



# Query #8: Is there a difference in the number of double faults based on gender?

- Table Used: PeriodInfo, MatchHomeTeamInfo and MatchAwayTeamInfo
- The “ratios” series generated by this formula:

  **ratios = double_faults_by_gender / gender_counts**

- Result: The ratio of double faults to the total count for each gender

```
gender
F    3.541608
M    2.646076
dtype: float64
```

# Query #8: Is there a difference in the number of double faults based on gender? (the independent t-test)

- We performed a t-test by using scipy.stats library:

```
T-statistic: -18.598815378195127, p-value: 4.3727670450299826e-76
Samples shows significant difference (reject H0)
so there is significant difference the number of double faults based on gender
```

# Query #10: Is there a correlation between a player's height and their ranking?

The `corr` function in pandas calculates the Pearson correlation coefficient, denoted as `r`, between two variables. ranging from -1 to 1:

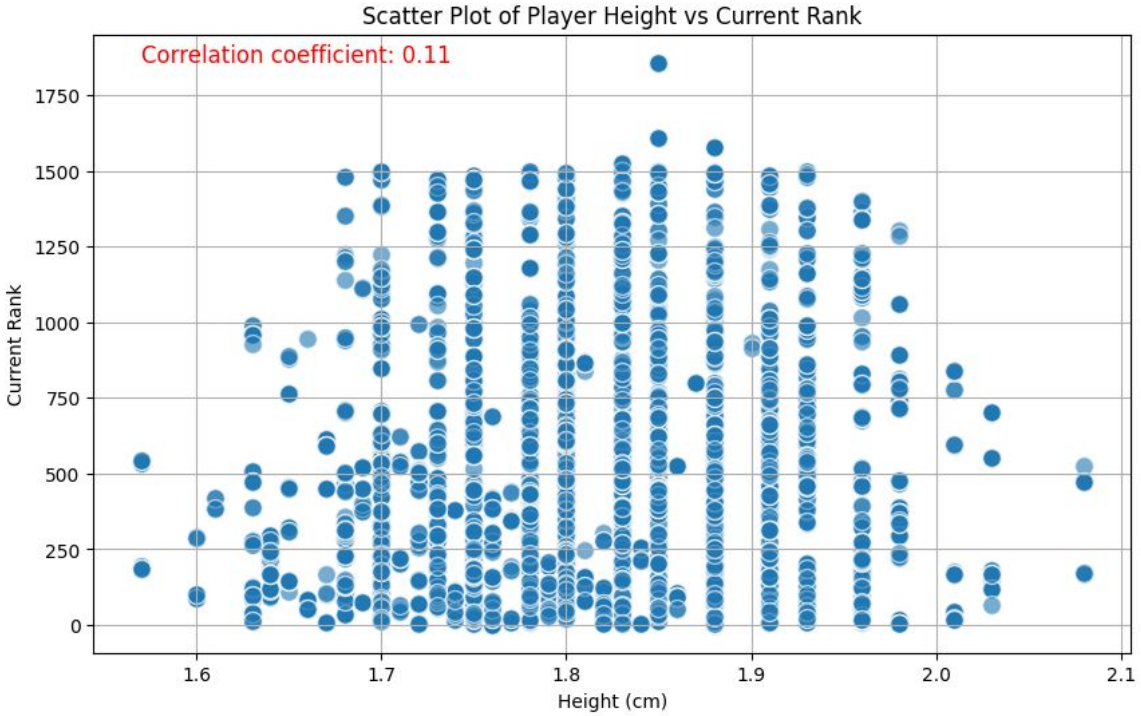
- **1**: A perfect positive linear relationship
- **-1**: A perfect negative linear relationship
- **0**: No linear relationship between the variables

We performed this function between ['height'] and ['current_rank'] of all players

**Result:**

```
Correlation coefficient between height and ranking: 0.11243439094485426
This value indicates a very weak positive correlation between these two variables
```

# Query #10: Is there a correlation between a player's height and their ranking? (scatter plot)



# Query #12: What is the average number of games per set in men's matches compared to women's matches?

- Tables used: MatchHomeTeamInfo, MatchAwayTeamInfo and GameInfo
- Created a table with columns of "match_id", "game_id" and "gender". The "game_id" is extracted by cleaning the duplicated data and keeping the last record of each "match_id".

GameInfo

✓ 0.0s

| | match_id | set_id | game_id | point_id | home_point | away_point | point_description | home_point_type | away_point_type |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 12260075 | 2 | 10 | 0 | 15 | 0 | 0 | 1 | 5 |
| 1 | 12260075 | 2 | 10 | 1 | 15 | 15 | 0 | 5 | 1 |
| 2 | 12260075 | 2 | 10 | 2 | 15 | 30 | 0 | 5 | 1 |
| 3 | 12260075 | 2 | 10 | 3 | 15 | 40 | 0 | 5 | 3 |
| 4 | 12260075 | 2 | 9 | 0 | 15 | 0 | 1 | 1 | 5 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 1467008 | 12385017 | 1 | 2 | 8 | 40 | A | 0 | 5 | 1 |
| 1467009 | 12385017 | 1 | 1 | 0 | 15 | 0 | 0 | 1 | 5 |
| 1467010 | 12385017 | 1 | 1 | 1 | 30 | 0 | 0 | 1 | 5 |
| 1467011 | 12385017 | 1 | 1 | 2 | 30 | 15 | 0 | 5 | 1 |
| 1467012 | 12385017 | 1 | 1 | 3 | 40 | 15 | 0 | 1 | 5 |

1467013 rows × 13 columns

# Query #12: What is the average number of games per set in men's matches compared to women's matches?

- Tables used: MatchHomeTeamInfo, MatchAwayTeamInfo and GameInfo
- Created a table with columns of “match_id”, “game_id” and “gender”. The “game_id” is extracted by cleaning the duplicated data and keeping the last record of each “match_id”.

**Result:**

```
Average number of games per set in men's matches: 9.2136869589116 39
Average number of games per set in women's matches: 8.931094383323682
```

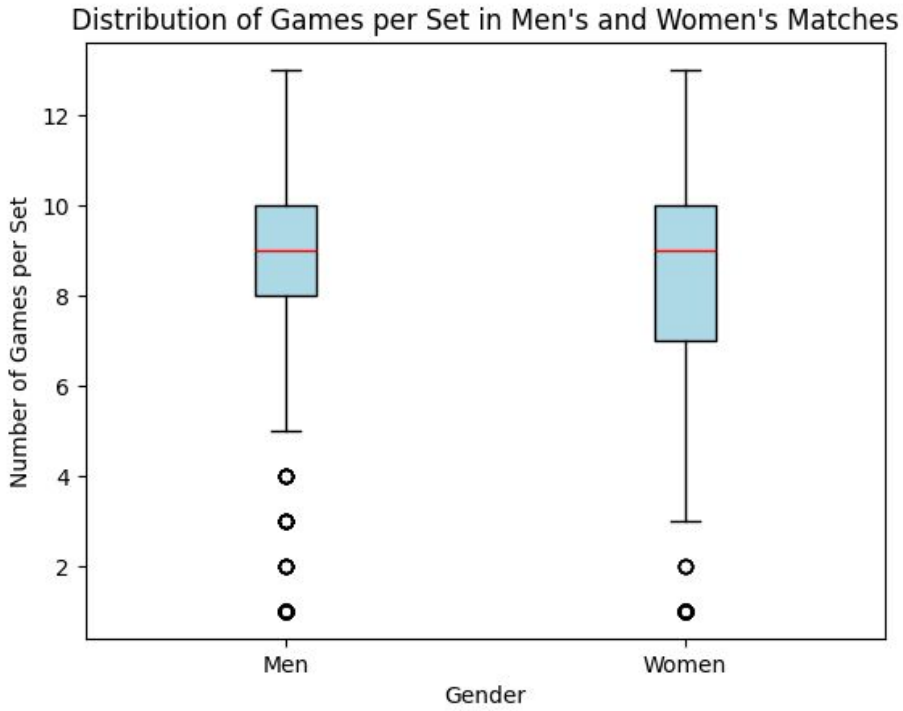
# Query #12: What is the average number of games per set in men's matches compared to women's matches? (t-test)

- We also performed a t-test independent between men and women data of number of games.

**Result:**

```
T-statistic: 7.739154137176257, p-value: 1.0681743235485147e-14
Samples shows significant difference (reject H0)
so there is significant difference between average number of games per set based on gender
```

# Query #12: What is the average number of games per set in men's matches compared to women's matches? (box plot)



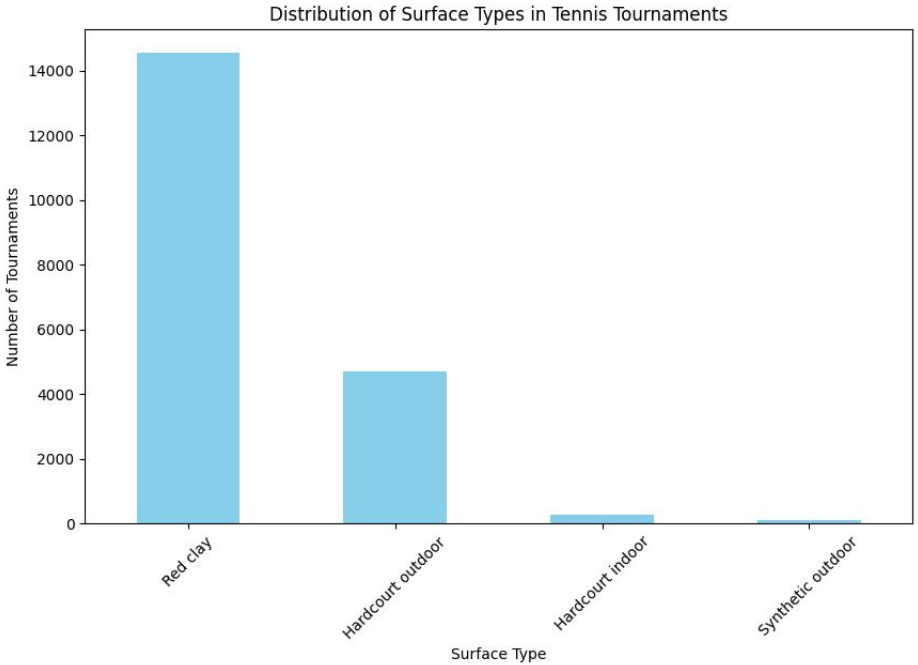
# Query #14: What is the most common type of surface used in tournaments?

- Table used: MatchTournamentInfo

**Result:**

The most common type of surface used in tournaments is: Red clay

Number of tournaments using this surface: 14545



# Query #16: Which player has the highest winning percentage against top 10 ranked opponents?

- Table used: MatchEventInfo, MatchHomeTeamInfo, MatchAwayTeamInfo
- Filtered the matches by top ten losers ranks
- Calculated the total numbers of wins of each winner, and number of wins against each top ranked loser
- Calculated the percentage

**Result:**

```
Winner ID with the maximum percentage of top wins: 14486.0
Maximum percentage of top wins: 50.00%
```

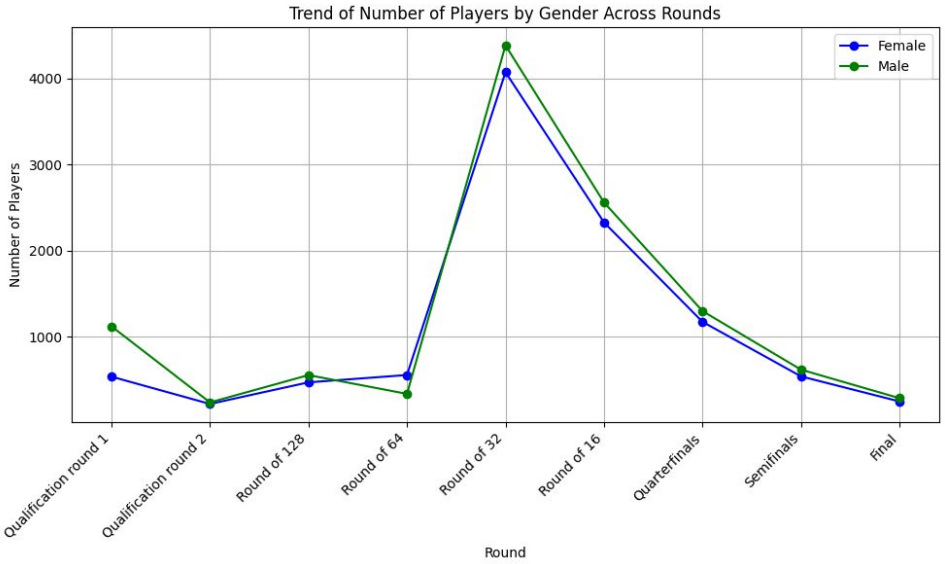
# Query #16: Which player has the highest winning percentage against top 10 ranked opponents?

This analysis helps in understanding which player performs exceptionally well against top-ranked opponents, showcasing their competitive edge and effectiveness in high-stakes matches. It provides valuable insights into player performance dynamics within the dataset.

```
Winner ID with the maximum percentage of top wins: 14486.0
Maximum percentage of top wins: 50.00%
```

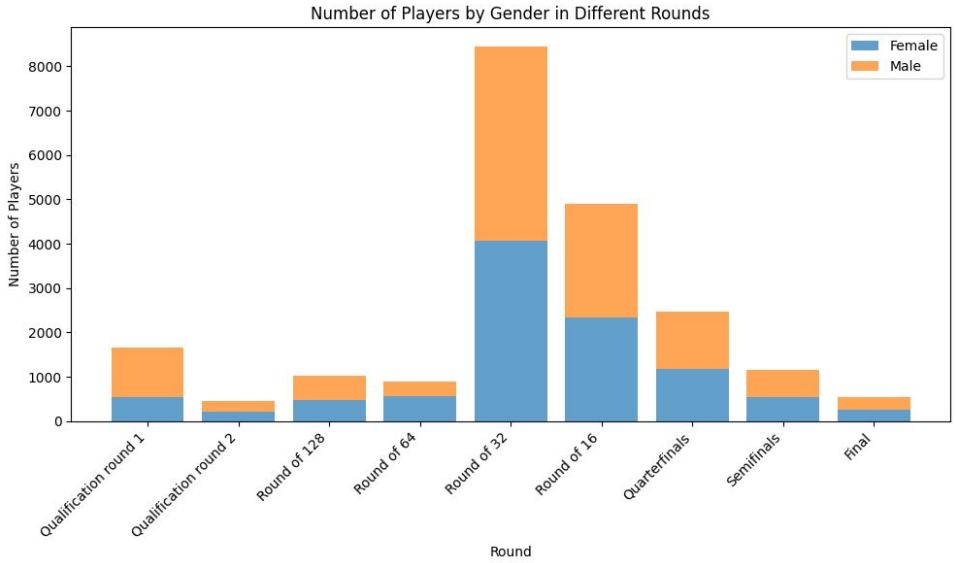
# Additional Query #1: Show the trend of the number of men and women in the different rounds.

- Used table: MatchHomeTeamInfo, MatchAwayTeamInfo and MatchRoundInfo
- Merging and dropping duplicates
- **Participation Trends**: The analysis shows that while more men tend to participate in the earlier qualification rounds, the gender gap narrows significantly as the tournament progresses to the intermediate and final rounds. This trend highlights the competitive nature and skill level of female players who advance through the rounds.



# Additional Query #1: Show the trend of the number of men and women in the different rounds.

- Used table: MatchHomeTeamInfo, MatchAwayTeamInfo and MatchRoundInfo
- Merging and dropping duplicates
- **Participation Trends**: The analysis shows that while more men tend to participate in the earlier qualification rounds, the gender gap narrows significantly as the tournament progresses to the intermediate and final rounds. This trend highlights the competitive nature and skill level of female players who advance through the rounds.



# Additional Query #2: Which player had the highest number of aces ?

- Used tables: PeriodInfo, MatchHomeTeamInfo
- The code extracts and interprets data to highlight significant achievements in serving within tennis matches, aiding in the analysis of player performance and match competitiveness.

```
Match with the highest number of aces by a single player: Match ID 12276774
The number of aces: 9
Player Name: Schoolkate T.
```

# Additional Query #3: Calculate "Win/Loss ratio of home players against opponents from the same country?

- Used tables: MatchEventInfo, MatchHomeTeamInfo and MatchAwayTeamInfo
- How effectively home players perform against opponents from their own country. A ratio greater than 1 indicates that home players tend to win more matches against compatriots than they lose, whereas a ratio less than 1 suggests the opposite.
-

| | country_home | win_loss_ratio |
|---|---|---|
| 0 | Israel | inf |
| 1 | Brazil | inf |
| 2 | Canada | inf |
| 3 | Chile | inf |
| 4 | Egypt | inf |
| 5 | Switzerland | inf |
| 6 | Netherlands | inf |
| 7 | Australia | 4.000000 |
| 8 | Russia | 3.500000 |
| 9 | Sweden | 2.000000 |
| 10 | South Korea | 2.000000 |
| 11 | Japan | 1.275000 |
| 12 | Spain | 1.272727 |
| 13 | Romania | 1.214286 |
| 14 | USA | 1.169811 |
| 15 | Mexico | 1.000000 |
| 16 | Serbia | 1.000000 |
| 17 | Austria | 1.000000 |
| 18 | Colombia | 1.000000 |
| 19 | Argentina | 0.933333 |

| | country_home | win_loss_ratio |
|---|---|---|
| 20 | Italy | 0.863636 |
| 21 | Czech Republic | 0.750000 |
| 22 | France | 0.741935 |
| 23 | China | 0.729730 |
| 24 | Germany | 0.538462 |
| 25 | Turkey | 0.500000 |
| 26 | Ukraine | 0.333333 |
| 27 | India | 0.111111 |
| 28 | United Kingdom | 0.083333 |
| 29 | Slovenia | 0.000000 |
| 30 | Slovakia | 0.000000 |
| 31 | Portugal | 0.000000 |
| 32 | Poland | 0.000000 |
| 33 | Peru | 0.000000 |
| 34 | Thailand | 0.000000 |
| 35 | Croatia | 0.000000 |
| 36 | Chinese Taipei | 0.000000 |
| 37 | Belgium | 0.000000 |
| 38 | Kazakhstan | 0.000000 |

# Additional Query #4: Which country has participated in the ATP tournaments the most?

- Used tables: MatchTournamentInfo, MatchHomeTeamInfo, MatchAwayTeamInfo
- Result:

**Country with the most ATP tournament participations: France (688 participations)**