Uncovering the Secrets of the 2024 ATP

Data analysis project

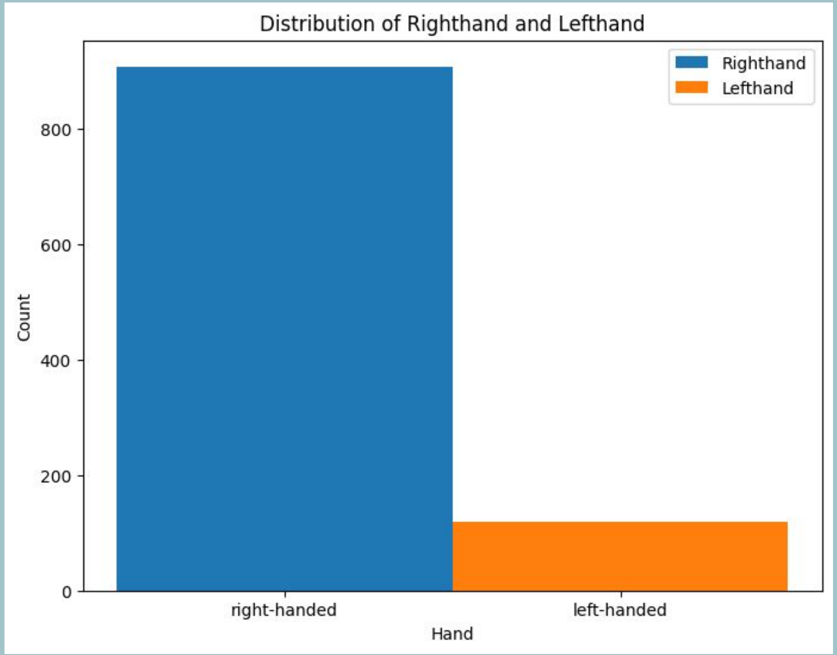
Introduction

- The 2024 ATP is one of the most prestigious tennis tournaments on the ATP Tour
- Deep into the data and uncover the hidden stories behind the matches played at this event.

Exploring Players Profile and participated countries Information

- Let's start by taking a closer look at the players who competed in the tournament
- The data reveals that the tournament featured a diverse mix of players, including stars like Świątek I.
- The total number of players played in the tournament was 2352.
- Interestingly, the average height of the players is 1.82 meters,
- A range of playing styles, including both right-handed and left-handed players.
  - right-handed 88.24%
  - left-handed 11.66%
  - ambidextrous 0.09%
- By analyzing the players' total prize money, we can see that Djokovic, Novak, Nadal Rafael and Murray, Andy have the 3 highest career earnings

This slide covers the answers of question numbers 1, 2, 13, Ex. Q. 3



full_name	total_prize
Djokovic, Novak	455008686
Nadal, Rafael	224653134
Murray, Andy	107560350

…Exploring Players Profile Information

- The player with the most winnings in matches was Uchijima M. with player id 253356 from Japan. The number of winnings for the player was 15 times.
- The country which produced the most successful tennis players was Italy with 1029 number of wins.
- The player who wins the most tournaments was Paquet C.
- There are 80 distinct countries participated in the ATP
- The player with the highest winning percentage against top 10 ranked opponents is: Świątek I. with 16.666666666666664 percent win

```
Italy          2237
USA            1946
France         1900
Japan          1235
Germany        1158
               ...
Azerbaijan        4
Ivory Coast       4
Kenya             4
Iran              4
Kyrgyzstan        4
Name: winner_country, Length:
93
```

player_name	No_winn_against_10top_rank_players
Świątek I.	3
Rublev A.	1
Cerundolo F.	1
Lehečka J.	1

This slide covers the answers of question numbers 3, 6, 9, 15, 16

...Exploring Players Profile Information

- The country with the most number of players is USA with 208 players (**`Table 1`**)
- The number of winning matches for each player has been shown in **`Table 2`**

`Table 1`

player_name	number_of_winns
Uchijima M.	15
Sherif M.	13
Sun F.	12
Urhobo A.	12
Wiskandt M.	12

`Table 2`

country	number_of_players
USA	208
Italy	181
France	153
Japan	122
Russia	118

This slide covers the answers of question numbers Ex. Q. 5, Ex. Q. 6

Analysing Match Duration

- According to the given information ,the longest match recorded in terms of duration is the match with match-id: 12346747 and players: Vulpitta G. & Pieri S.

match_id	period_1	period_2	period_3	period_4	period_5	Current_period_start_timestamp	Sumation_time_periods_match
12346747	167761	3392	0	0	0	1.72E+09	171153

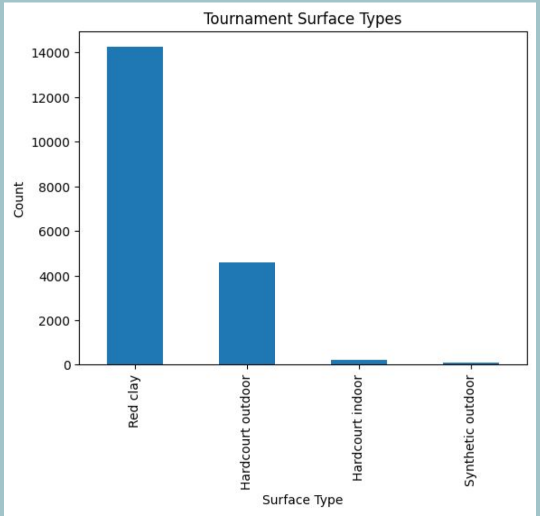
- The average duration of matches is 182.77 minutes.

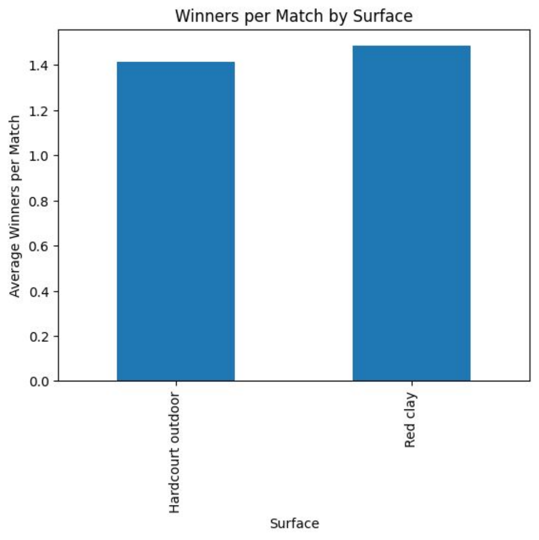
This slide covers the answers of question numbers 4, 11

Exploring Tournament Dynamic

- Now, let's take a closer look at the tournament dynamics and how ground type may have influenced the outcomes.
- The data shows that the tournament was played entirely on red clay courts, which is known to favor certain playing styles and strategies.
- The average number of winners per match differ between hard and clay court surfaces
- Results esults shows that 'Red clay' ground provides more winners in compare with 'HardCourt outdoor'.

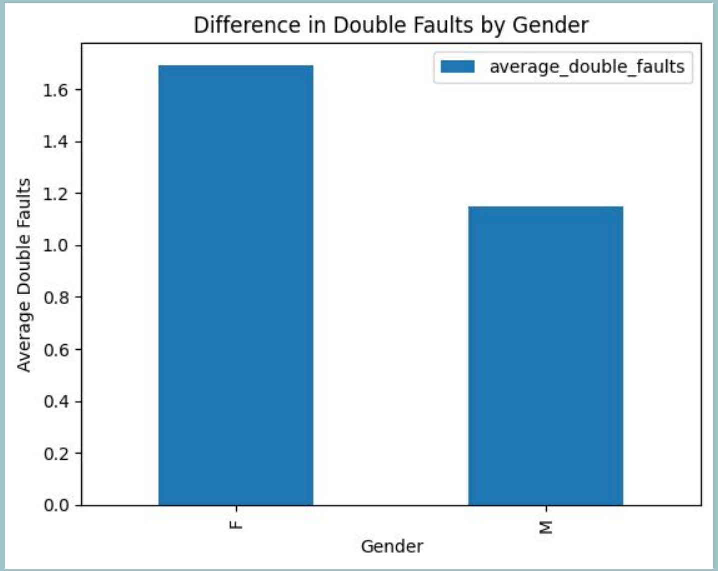
This slide covers the answers of question numbers 14, Ex Q. 4





Uncovering Insights from the Scoreboard

- Finally, let's dive into the scoreboard data and see what insights we can uncover about the team performances.
- Most matches typically had 2 sets played in a tennis match. Some of matches played 3 sets.
- The average number of aces per match was 6.71
- There is a difference in the number of double faults based on gender, as we can see in the following plot

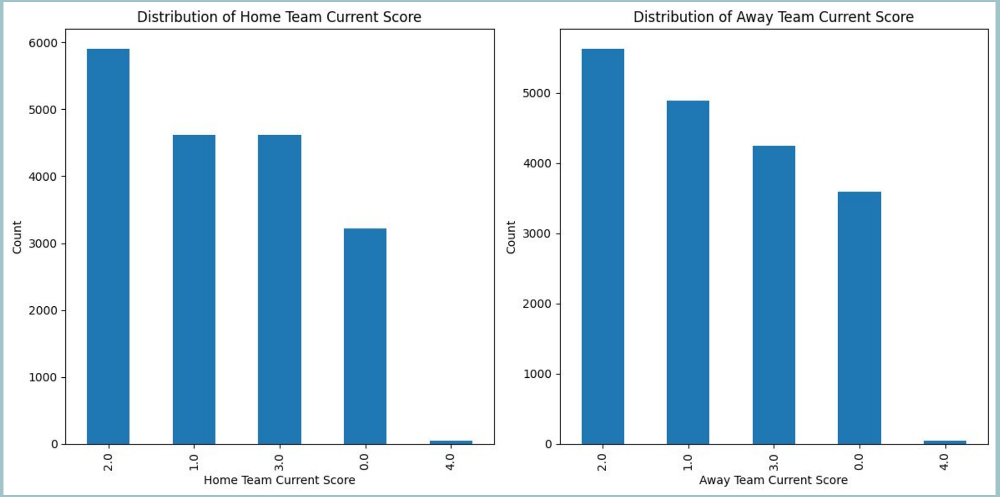


```
match_id
12260075    2
12260076    3
12260077    3
12260078    3
12260080    2
           ..
12384789    2
12384806    3
12384892    2
12384975    3
12385017    3
Name: set_num, Length:
6658
```

This slide covers the answers of question numbers 5, 7, 8

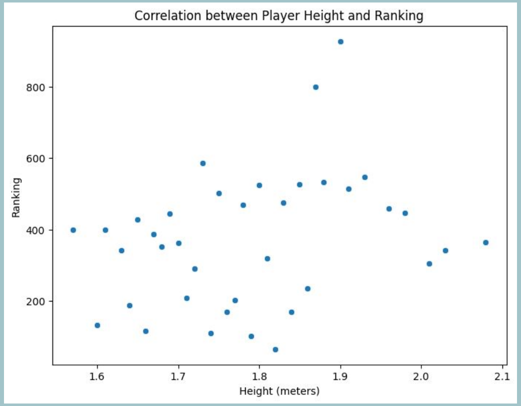
…Uncovering Insights from the Scoreboard

- The average number of games per set in men's matches was 9.18, while this value was 8.92 in women's matches.
- The average number of breaks per match was 16.37
- The distribution of the 'current_score' values in the df_home_score and df_away_score DataFrames.

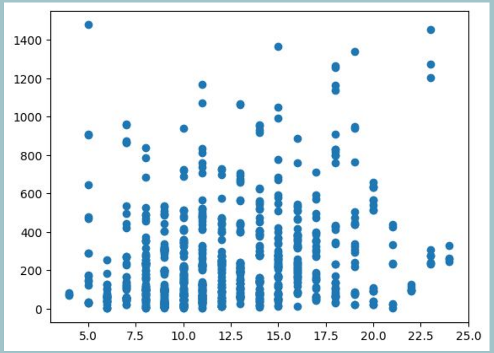


This slide covers the answers of question numbers 12, 17, Ex. Q. 8

Analyse some parameters correlations



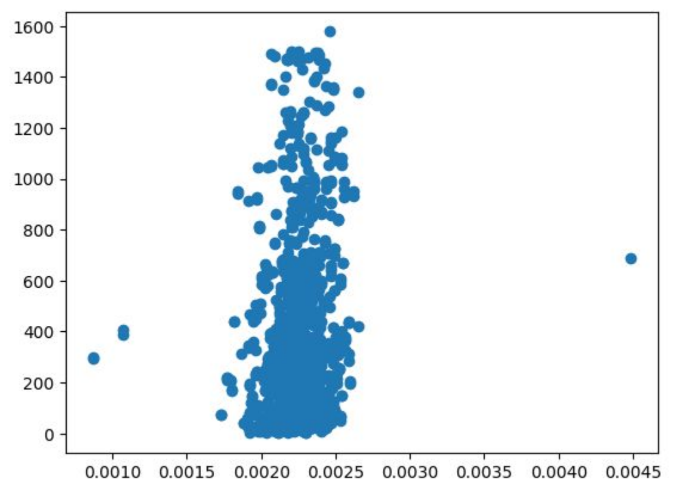
- The correlation between height and ranking was 0.10, which is very weak. The scatter plot verify this event
- There is a weak correlation between players experience and their rank
- There isn’t much difference between the ranks for right-handed and left-handed players. Right-handed players have an average ranking of 359.29, while left-handed players have an average ranking of 367.87.

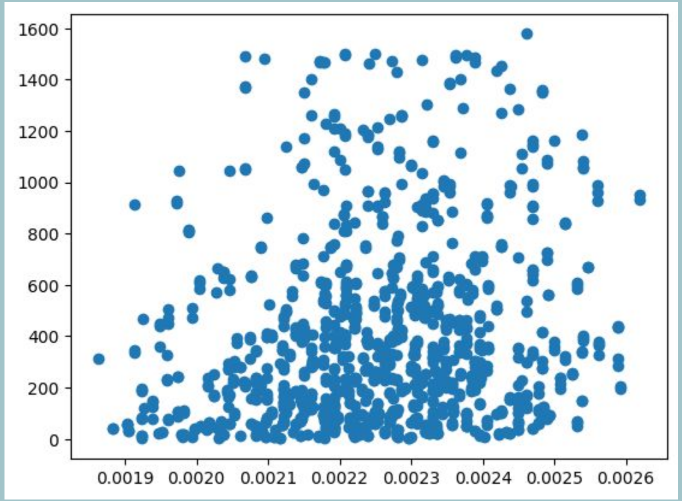


This slide covers the answers of question numbers 10, Ex. Q. 1, Ex. Q. 2

Analyse some parameters correlations

- There is no correlation between Height and weight ratio (BMI) and ranking value





This slide covers the answers of question numbers Ex. Q. 7,

Conclusion

- Through this data analysis project, we've uncovered a wealth of insights about the 2024 ATP
- From player profiles to match durations and tournament dynamics, the data has provided a fascinating glimpse into the world of professional tennis
- As we look ahead to future tournaments, this analysis can serve as a valuable resource for players, coaches, and fans alike, helping them better understand the factors that contribute to success on the ATP Tour.

Appendix: Detailed Answers for Questions

1- How many tennis players are included in the dataset?

- ○ We first concat two dfs based on their 'player_id', home and away teams informations. and count number of players by 'unique' function.

Total number of players: 2352

…Appendix: Detailed Answers for Questions

“player_id” and “ height” were considered from home and away match_team_info. By dropping nulls and duplicates, the data prepared for calculating **mean** value.

```
The result is:

Average height of players: 1.8200083263946711
```

According to the given information, the average height of the players in this series of matches is 1.82

Note:
There was one player with 2 different heights
His heighs were not considered in calculation

player_id	full_name	height
192862	Damas, Miguel	2.08
192862	Damas, Miguel	1.86

3. Which player has the highest number of wins?

We use 'game_df' that contains the points for each match separated by games and sets. We sum all points per each match for away and home teams. Then if any of sum of points for home or away team is greater we consider it as winner.

We then use that name of players take participate in each match. Every match has two players. We use this dataframe and the above df to have a tables containing match_id, home_points, away_points, winner_code and the player name that wins the match. The 'value_count' method is used to count the number of repeated player's names. Then using 'idxmax' get the name.

```
The player with the most winnings in matches is:
Uchijima M.

The number of winnings for the player is:  15
```

```
Uchijima M.          15
Sherif M.            13
Sun F.               12
Wiskandt M.          12
Urhobo A.            12
```

4. What is the longest match recorded in terms of duration?

For time info the “match_time_df” was used. Procedure of time calculation is as below:

1. Replacing the NAN values by 0
2. Remove duplicates
3. Calculating summation of each match periods time
4. Filter the maximum time duration

match_id	period_1	period_2	period_3	period_4	period_5	Current_period_start_timestamp	Sumation_time_periods_match
12346747	167761	3392	0	0	0	1.72E+09	171153

According to the given information ,the longest match recorded in terms of duration is the match with **match-id: 12346747   players: Vulpitta G. &   Pieri S.**

5. How many sets are typically played in a tennis match?

We use the power_df that has the number of sets for each match. We group by the power_df by 'match_id' and then count the unique sets of a match.

```
match_id
12260075    2
12260076    3
12260077    3
12260078    3
12260080    2
           ..
12384789    2
12384806    3
12384892    2
12384975    3
12385017    3
Name: set_num, Length: 6658
```

6. Which country has produced the most successful tennis players?

There are several views on this matter:

1. If most successful means **most winning** the **table 1** can describe the answer and Italy did perfect.
2. If most successful means being in **top 3** ranks the **table 2** shows the places.
3. If most successful means **average rank** the table 3 presents it.

```
Italy          2237
USA            1946
France         1900
Japan          1235
Germany        1158
               ...
Azerbaijan        4
Ivory Coast       4
Kenya             4
Iran              4
Kyrgyzstan        4
Name: winner_country,
Length: 93
```

Table 1

name	country	current_rank
Świątek I.	Poland	1.0
Djokovic N.	Serbia	1.0
Sabalenka A.	Belarus	2.0
Sinner J.	Italy	2.0
Gauff C.	USA	3.0
Alcaraz C.	Spain	3.0

Table 2

country	Average_rank
Philippines	161.000000
Andorra	294.666667
Malta	330.500000
Syria	344.000000
Burundi	371.000000

Table 3

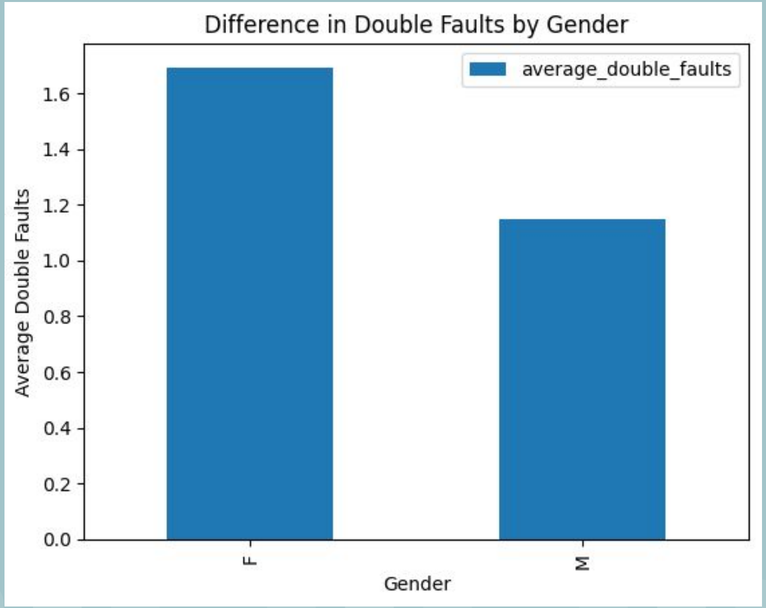
7. What is the average number of aces per match?

We use period_info data frames containing the 'statistic_name'. We select rows that their 'statistic_name' is 'aces'. Then we sum the 'home_value' and 'away_value'. The next step is to get the number of matches by 'nunique' method. The average aces per match is get by dividing the total_aces by number of matches.

```
The average number of aces per match is:  6.711020495761348
```

8. Is there a difference in the number of double faults based on gender?

In conclusion of the given information, there is a difference between male and female in the number of double faults. The average double faults value is more in females than in males.



gender	home_value	away_value	average_double_faults
F	1.650679	1.734638	1.692658
M	1.109402	1.1864	1.147901

9. Which player has won the most tournaments in a single month?

We use the player Winner dataframe previously produced in question 3. It contains 'match_id', 'name_player1', 'name_player2' and 'winner_name' fields. Now we join this dataframe with the Match Tournament Info data frame to get the player who won the most tournaments.

```
The player who wins the most tournaments is: Paquet C.
```

10. Is there a correlation between a player's height and their ranking?

For this case , corr() method

```
The correlation between height and ranking is:  0.10782530868262802
```