Performance Analysis in Grand Slams and ATP Rankings: A Decade of Data

Alberto Arienti[1] **and Sofia Turroni**[2]

[1]**Matriculation number: 860251, email: a.arienti10@campus.unimib.it**
[2]**Matriculation number: 871277, email: s.turroni@campus.unimib.it**

ABSTRACT

The purpose of this project is to develop a data management system focused on the world of professional tennis, with a focus on Grand Slam tournaments over the past decade. The main objective is to acquire, store and integrate data from different sources to analyze the performance of young tennis players.
The data were acquired through web scraping techniques from up-to-date authoritative sources.
The acquired data were structured and stored in a MongoDB database, chosen for its flexibility in handling semi-structured and unstructured data.
This data management project provides a robust and scalable platform for collecting, storing, and analyzing tennis data, with an emphasis on the performance of young tennis players. The integrated, structured approach facilitates in-depth analyses and enables meaningful insights that can influence strategic decisions in the world of professional tennis.

Keywords: Data Science - Web Scraping - MongoDB - Tennis - ATP Tour

CONTENTS

1 INTRODUCTION

Tennis is one of the most popular and followed sports in the world, characterized by a mix of physical, strategic and mental skills. Played professionally around the world, tennis attracts millions of spectators and players, with prestigious tournaments and a long historical tradition.
Player rankings, both men's (ATP) and women's (WTA), are determined by points accumulated at various tournaments throughout the year.
Points are awarded based on players' performance in tournaments, with **Grand Slam tournaments** offering the most points. The rankings are updated weekly and reflect players' recent performances.
In tennis, the most prestigious and coveted tournaments are the four Grand Slams:

- Australian Open: held in January in Melbourne, Australia. It is played on a hard surface;
- French Open (Roland Garros): held in May-June in Paris, France. It is played on clay;
- Wimbledon: it is held in June-July in London, England. It is the oldest tennis tournament

and is played on grass;

- US Open: it is held in August-September in New York, USA. This tournament is also played on a hard surface.

Winning one of these tournaments is considered one of the greatest achievements in a tennis player's career.

1.1 Research question

Over the past decade, many young tennis players have emerged on the world stage, showing significant promise. However, identifying which of these players will continue to develop into champions is a complex challenge that requires in-depth analysis of their performance data.
Which young tennis players have achieved the best results in Grand Slam tournaments?
In this project we will then go on to analyze the performance of young tennis players in Grand Slam tournaments, identifying the most promising ones and visualize the data through graphs and charts to highlight significant trends and patterns.

2 DATA ACQUISITION

To answer our research question, we extracted data relating to the ATP rankings before each Grand Slam tournament of the last 10 years, the overview of the players (height, weight, age...) and all the matches of these tournaments (also these for the last 10 years). To obtain this data, we used an automatic **web scraping** technology using Python along with the *Selenium* and *WebDriver Manager* libraries. This combination of tools allowed us to navigate and interact with dynamic websites to gather the necessary information efficiently.
The tools we used for our data acquisition were therefore:

- Python: A programming language used for automation and data analysis;
- Selenium: Python library that allows browser automation to simulate human interaction with websites;
- WebDriver Manager: Library that automatically manages browser drivers, making it easier to use Selenium.

2.1 Data sources

The links we used to collect this data are the following:

- flashscore.com: it is a site where it is possible to see the results of matches of all sports, including those of past years. From this site the data of all SLAM matches have been acquired (US open, Australian Open, French Open and Wimbledon), from the 1/64 of final to the final and from the years 2014 to 2024;
- atptour.com: is a global point of reference for everything related to the world of professional men's tennis. It offers a wide range of content and services such as news, results, updated ATP rankings and their history and player profiles.

Through the first website we identified all ten links (one for each year and ten for each SLAM) from the first link and with the use of a loop acquired the data given that the structure of the HTML page did not change.
While the second website was useful for the history of the ATP rankings before each tournament we went to consider for extracting the rankings of each player, after having identified all the players who appeared in the rankings before the SLAMs of the last ten years and always using a for loop they were acquired the data we were looking for.
The decision was to extract the data in JSON format, below is the structure of how the data was saved (Figure 1, 2 and 3). It is possible to see that the documents are key-value pairs where for each key (the year) an array is associated which contains for each observation a dictionary which can be a match or a position in the ranking (in the case of Figure 1 and 2), while the format changes for the data representing player overviews (Figure 3).

```
{
    "2023": [
        {
            "match type": "Final",
            "player1": "Alcaraz C.",
            "player2": "Djokovic N.",
            "points_p1": [
                "1",
                "7(8)",
                "6",
                "3",
                "6"
            ],
            "points_p2": [
                "6",
                "6(6)",
                "1",
                "6",
                "4"
            ],
            "winner": "Alcaraz C."
        },
```


Figure 1. Structure of the firs JSON file for the matches, the *year* acts as the key to which an array of key-value dictionaries is associated with *match type* (type of match played), *player_1* and *player_2* (names of the tennis players), *points_p1* and points_p2 (are array that contain the points made respectively by the two players in the game) and *winner* (winner of the match).


```
{
    "2024": [
        {
            "Name": "Novak Djokovic",
            "Rank": "1"
        },
        {
            "Name": "Carlos Alcaraz",
            "Rank": "2"
        },
        {
            "Name": "Daniil Medvedev",
            "Rank": "3"
        },
```


Figure 2. Structure of the second JSON file for the rank, the *year* which allows access to an array, which contain the *name* and *rank*(ranking position) before each slam.


```
{"Yoshihito Nishioka": {"year of birth": "1995", "weight":
"64", "height": "170", "turned pro": "2014", "hand": "Left-
Handed", "backhand": "Two-Handed Backhand"}, "Ivo Karlovic":
{"year of birth": "1979", "weight": "104", "height": "211",
"turned pro": "2000", "hand": "Right-Handed", "backhand":
"One-Handed Backhand"}, "Holger Rune": {"year of birth":
"2003", "weight": "77", "height": "188", "turned pro": "2020",
"hand": "Right-Handed", "backhand": "Two-Handed Backhand"},
```


Figure 3. This JSON file differs from the previous ones because the key is the player's name and not the year, which allows us to access a dictionary where there are characteristics for each player. The player information is as follows: *year of birth*, *weight*, *height*, *turned pro* (in which year did he become pro), *hand* and *backhand* (hand or hands with which he backhands).

3 CONFLICT RESOLUTION & DATA CLEANSING

Data quality and integrity are key elements in achieving reliable and meaningful results. Data cleansing and conflict resolution are two crucial processes for transforming raw, often messy and inconsistent data into useful and consistent information. Without these preliminary steps, data-driven analyses and decisions can be compromised by errors and biases.
For our tennis project, we particularly addressed the need to:

- **Clean the data**: remove errors, duplicates, and missing values to ensure that the data are accurate and complete.

- **Resolving conflicts**: Ensuring consistency of data from different sources, harmonizing information that may be discordant.

3.1 Conflict resolution:

One of the main problems encountered concerned the discordance in player names between the two data sources. As can be seen in Figure 1, the variables *player1* and *player2* representing a tennis player have the player's last name and the initials of the name dotted, while in Figure 2 the full last name and first name appear. Although representing the same person (as we can see from the Figures 1 and 2, for *Novak Djokovic* and *Carlos Alcaraz*) the name is represented in two different formats, this later when we want to enrich the dataset will give us problems, so it is necessary to find a way to represent the name in one format.
To solve this problem, we developed a **name transformation function** that can automatically convert full names into their corresponding abbrevi-

ated names. Using this function, we added a new variable called *Name Formatted* to the document containing rankings and player names. The new variable coincides with the *Name* variable in the other document.

3.2 Data cleaning and data quality evaluation:

Finally, an overall assessment of the quality of the data was made, considering various evaluation metrics; we checked their accuracy by comparing the data between the two sources to make sure there were no significant discrepancies.

4 DATA STORAGE

To effectively manage and organize the data in our tennis data management project, we chose MongoDB as our database system. MongoDB is a NoSQL database that offers flexibility, scalability, and ease of use, making it ideal for managing large volumes of heterogeneous data such as those related to tennis tournaments and player rankings, thus perfect for the situation in which we are.
Before doing so, however, we decided to change the data schema to be structured in a more readable way.

4.1 Schema transformation

Initially, the data were structured with a key representing the year, containing a list of dictionaries representing matches or rankings. This approach had some limitations in terms of flexibility and data accessibility. To improve data management and analysis, we decided to transform the schema by adding a "year" key within each dictionary, making data access and manipulation more direct and efficient, so the data would no longer be dictionary lists but simply dictionaries .

4.2 Database structure

We have created a database called *"tennis"* that contains several collections, each of which is designed to store a particular type of data. The collections are organized as follows:

- **Grand Slam Tournament Collections**: each Grand Slam tournament has its own collection to store match data, *Australian_Open*, *French_Open*, *Wimbledon* and *Us_open* are the names we have given to our collections. These collections contain a huge amount of documents, each document is a match of that tournament in the last ten years;

- **Ranking Collections**: for each tournament we created a collection to store the rankings before that tournament, so four collections (*RankingBefore_Australian_Open*, *RankingBefore_French_Open*, *RankingBefore_Wimbledon* and *RankingBefore_Us_open*):

- **Player Overview Collection**: this collection contains general information about players.

5 DATA ENRICHEMNT

In addition to structuring and transforming the data, we focused on enriching the collections containing batch documents. Data enrichment involves the integration of additional information that can provide a richer and more useful context for analysis.
Specifically, we enriched the match documents by including players' rankings positions before each tournament, using data from the rankings collections and player overview.

5.1 Data Enrichment Process

For each match, we integrated the players' positions in the world rankings before the start of the tournament. This enrichment was done by matching with the year of the tournament and using a new variable *"Name Formatted"* to equalize the *player1* or *player2*, in case a player is not in the top 100 of the ATP ranking, the value is replaced with *"not in top 100"*. After that in addition to the ranking positions, we added general information about the players coming from the players_overview collection, where instead we just matches with the two names coming from the two collections since the 'overview does not change every year we did not need to make time discriminations.

```
"player1": {
  "_id": {
    "$oid": "668a9c5de169e56598eae0cb"
  },
  "Name": "Jannik Sinner",
  "Rank": "4",
  "Name Formatted": "Sinner J.",
  "year": "2024",
  "backhand": "Two-Handed Backhand",
  "hand": "Right-Handed",
  "height": "188",
  "turned pro": "2018",
  "weight": "76",
  "year of birth": "2001"
},
```


Figure 4. The integrated document will then have this format, as you can see the *Name Formatted* variables have been added, the *Rank* and different characteristics regarding the 'overview of the tennis player.

6 VISUALIZATION & CONCLUSION

Data visualization is a crucial component in our data management project, as it allows us to transform complex data into easily interpretable graphical representations. We developed several graphs using queries on the integrated dataset we created, which can enable us to answer the research question we had set for ourselves.

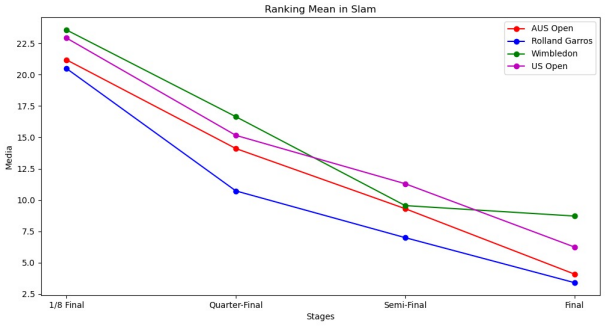

Figure 5. The graph depicts for matches ranging from the round of 16 to the final, based on Slam (broken down by color), the line drawn indicates the average of the *rank* variable for each type of match. From this graph we can see on average what age a player is when they play a type of match in the different tournaments.

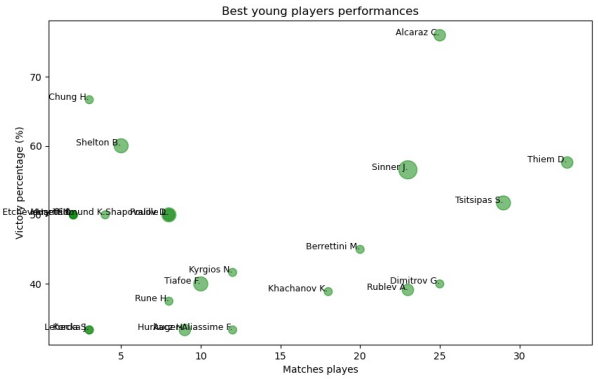

Figure 6. The graph depicts for matches where the type match variable takes values from 1/8 Final to Final,, we defined young tennis players as those players who were less than 25 at the time of the match. The size of the bullets indicates the number of times the player beat an opponent with the rank variable greater than his or her own..

6.1 Conclusion

We initially tried to observe whether there is a Slam in which it is easier for a young tennis player to win.

From Figure 5 it is possible to see that the average of the variable *Rank* of players is higher for matches in the Wimbledon Slam, a young player (who therefore generally has a higher rank) has more possible advances in the mentioned Slam than in the others.

Going then to look at the *Wibledon* winners we noticed that the average is so high because *Novak Djokovic* in 2018 won while being placed 18th in the ATP ranking, this value raised the average a lot causing it to be distorted.

So despite the fact that he seems more suitable for "outsiders," we cannot call him one.

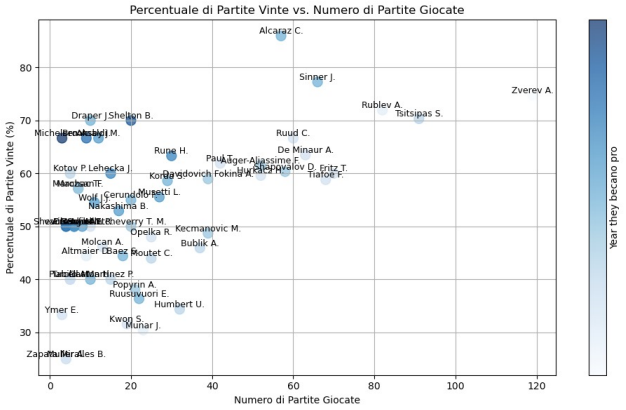

Figure 7. The graph depicts for matches where , we defined young tennis players as those born after 1997, so with a maximum age of 26 or 27 currently. On the x-axis are the matches they have played in tournaments while on the y-axis is the winning percentage, the graduation of the color scale represents the value of the variable *turned pro*, that is, how long they have been considered pro. In the case of a tennis player who has been considered pro for a long time the value will be lighter, the hue increases as the years decrease.

Wimbledon winners, the 'age they were when they won, and their ATP ranking are listed here.

'2023': ['Alcaraz C.', 20, 2],
'2022': ['Djokovic N.', 35, 1
'2021': ['Djokovic N.', 34, 1],
'2019': ['Djokovic N.', 32, 1],
'2018': ['Djokovic N.', 31, 21],
'2017': ['Federer R.', 36, 5],
'2016': ['Murray A.', 29, 2],
'2015': ['Djokovic N.', 28, 1]
, '2014': ['Djokovic N.', 27, 2]

While the Figure 6 and the Figure 7, we can see how *Sinner Jannik* and *Carlos Alcaraz* are the youngest tennis players who have recently entered the ATP rankings and have played a significant number of matches, both winning more than 75% of grand slam matches. One factor that distinguishes the two tennis players is that Sinner compared to Alcaraz has beaten people higher than him in the rankings many times, this means that he has potential to do even better.
Another player who has played fewer matches but still achieved good results is *Ben Shelton*, who compared to *Jannik Sinner* and *Carlos Alcaraz*, who are already established, could surprise in the upcoming. We can also notice how *Tsitsipas* always achieved good results in both

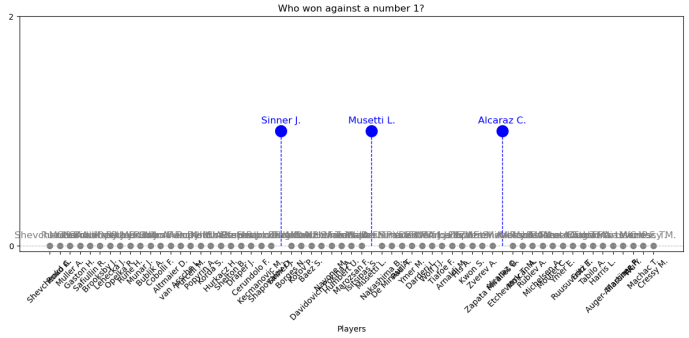

Figure 8. The graph depicts under definition of young tennis player of the 7, tennis players who won against the first position in ranked ATP.

The Figure 8 shows only 3 young tennis players who met the condition, this means it is not for everyone and among the names Jannik Sinner and Carlos Alcaraz are present, another signal that they are destined to compete for first place in the coming years.