Tennis Player Performance Analysis

Professor Adam G. Anderson
Digital Humanities 100

Xinyue (Sherry) Liu
05/28/2021

1. Project Description

Tennis, being an Olympic sport played at all levels of society and all ages, although not as popular as football and basketball, has always been favored by many people.

Since the 1890s the rules and forms of tennis tournaments have nearly remained unchanged, yet generation after generation of tennis players has been put onto the stage and competed fiercely with each other, replacing the old generations with the new ones.

This project aims to find out the Greatest Players of All Times through a detailed and comprehensive data analysis. Is the most renowned player necessarily the best player in history? Are there any players with great performance but were not noticed by anyone? Aside from pure statistical analysis, the popularity of the player and comments from others will also be taken into account.

2. Dataset Description

This dataset collects data of the winner and the runner-up of each year's grand slams over the span of 141 years (from 1877 to 2018), as well as the names of the tournaments.

I added the birth date of some outstanding players so that through data manipulations, I could get the information of their age when winning the grand slam.

3. Research Question

Who are the Greatest Players of All Times? (Based on their overall performance from the following aspects)

- What is the general trend of tennis matches? Does the competition become fiercer over the years or the opposite?
- What are the statistics of each player? Are there any outstanding players? Or most of them are close to average?
- What is the distribution of each player's performance within his career lifetime?

4. Tools and Methods

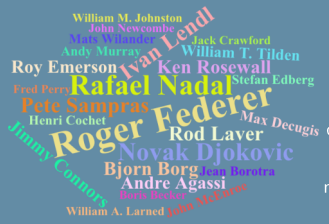
R: used for data cleaning, data manipulation, and data visualizing packages used in R: readr, dplyr, ggplot2, wordcloud2

Python: used for data manipulation and data visualizing packages used in Python: pandas, matplotlib

I used R to clean the data and conduct exploratory data analysis, the outcomes of which are shown below. I first imported the data, added information about the birth dates of some players to the dataset, and got the ages of winning the grand slam of each player. Then, I started doing exploratory data analysis.

I explored the data from five aspects: the number of grand slams won by each player, the number of grand slams won in each year, the distribution of player's grand slam data, the number of times each player won the grand slam or was the runner-up and the grand slam distributions of the four most famous tennis players. I reached some conclusions for each of the analyses, but I got rid of the less important and less attractive ones. The ones that I deemed to be worth mentioning have been displayed below.

5. Exploratory Data Analysis



I manipulated the data to get the number of times each player won the grand slams or was the runner-up, and created a word cloud based on the frequency of this number. Name with a larger size means more winning.

Intuitively, some of the names that pop up include Roger Federer, Rafael Nadal, Novak Djokovic, and Ivan Lendl. Some of the others also have a moderate name size. For demonstration purposes, I only include players that have won above nine grand slams, therefore the distribution may seem scattered. But if all the names were included, since many of them have only won one grand slam, the four players mentioned above would pop up even more, since the number of winning and being the runner-up of the above-mentioned four players are 30, 24, 21, and 19.

Number of Grand Slams Won by Each Player

Number of Times Each Player Won the Grand Slam or Was the Runner-Up

These two circular bar plots represent the number of grand slams won by each player. The left one only includes the cases where the player was the champion, whereas the right one also includes the case where the player was the runner-up.

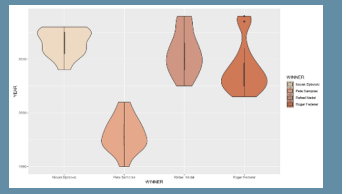
From the polar bar plots, it is not hard to see that there exist some outstanding bars among the others. If the names of these outstanding bars in two plots correspond with each other, we would admit that these are the players that outperform the others.

This chart analyzes the grand slam distributions of the four famous tennis players, aiming to visualize their performance at different stages of their career lifetime. Unexpectedly, the distribution looks so different from each other, since I would assume that all the players would have their best performance approximately from age 25-30.

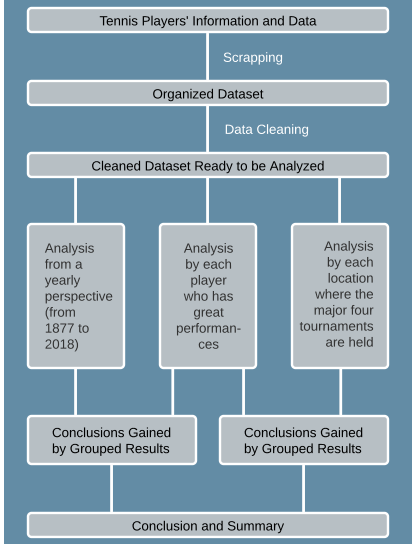
But from the graph, we can see that only Pete Sampras and Rafael Nadal roughly follow this pattern.

Grand Slam Distributions of the Four Famous Tennis Players

Novak Djokovic's performance grew better with his age, and Roger Federer's performance follows a much more bizarre pattern.



6. Project Workflow



7. Interpretation of Results

The exploratory data analysis provides me with three different aspects of this dataset and the performance of each tennis player. Just by looking at the number of grand slams won, Roger Federer, Rafael Nadal, Pete Sampras, and Novak Djokovic outperforms the others. Among these four players, three of them are players of our generation, which means that from the 1870s to 2020, the trend is that the players' abilities have become less converged and more polarized.

Although further analyses and research is needed, I would probably conclude that the best players of all times are more or less the four players mentioned above based on their grand slam counts, career performances, and the general trend of tennis tournaments.

Works Cited

https://en.wikipedia.org/wiki/Tennis
https://thenounproject.com/search/?q=tennis
https://www.kaggle.com/manish2104/tennis-grand-slams-data
https://nycdatascience.com/blog/student-works/tennis-player-performance-analysis/

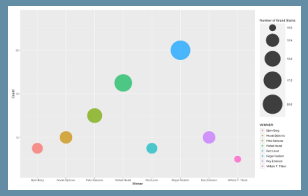
Tennis Player Performance Analysis

Professor Adam G. Anderson
Digital Humanities 100

Xinyue (Sherry) Liu
05/28/2021

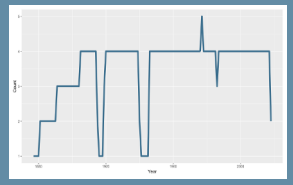
8. Data Analysis

1) Number of Grand Slams Won by Players



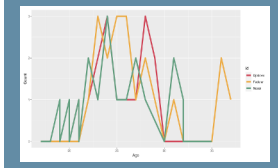
Although I took out all the players that have won less than 10 times to make the graph cleaner, the counts of the top three players are still very outstanding. From the graph we can see that the player with the highest count corresponds to Roger Federer, followed by Rafael Nadal, and Pete Sampras. The number of times winning is 20, 17, and 14 respectively. Followed by them are Novak Djokovic and Rod Laver, both of whom have won 12 times. These five players together have won 75 grand slams in total, which is fifteen percent of the total grand slams recorded in this dataset. Admittedly, their performances are very outstanding compared to other players.

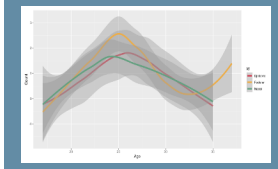
2) Number of Grand Slams Won Yearly



Theoretically, there should be four winners each year. Yet from this graph we can see that during some years there are only one, two, or three champions, and for one year there are five champions. This means that there are some limitations to this dataset. For some years there might be missing champions. The data is not complete.

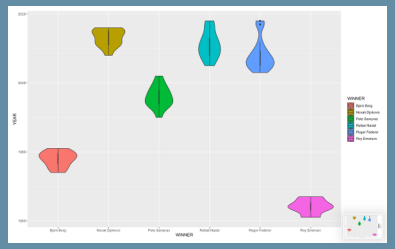
3) Distribution of Player's Grand Slam Data





It might not be very intuitive to see from the first plot but from the second plot, we can see that among the three players, Roger Federer has a slightly better performance. Rafael Nadal and Novak Djokovic have similar performance curves, but Nadal's curve seems to be more shifted to the right, and Djokovic's curve is more shifted to the right. The curves of Nadal and Djokovic both follow an approximately normal (Gaussian) distribution, with a maximum point at around age 25, whereas Federer's curve follows a stranger pattern, indicating his golden time might not just be around age 25, but also during his late 30s or even early 40s, which is certainly unusual for tennis players.

4) Grand Slam Distributions of the Famous Players



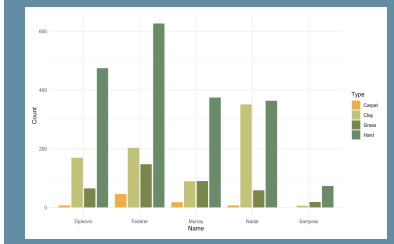
This plot visualizes the distribution of winning grand slams of each individual player. I have selected the six players with the most outstanding performances. As we can tell from the graph, the distribution of most players follows a roughly normal (Gaussian) distribution, since between age 23 to 28 are the golden time for players to get the best scores.

Before 23, players do not have time to gain enough tactical experience. And after 28, players' physical strengths start to decline.

From the graph, we can see that Pete Sampras, Rafael Nadal, Bjorn Borg, and Roy Emerson all follow this pattern. Yet surprisingly, the distribution of Novak Djokovic and Roger Federer's curves do not quite follow this pattern. It seems like that Djokovic's winning rate grows with his age, and just started to decline in the year 2017, but the trend is not obvious since the data only records matches until the year 2018.

Federer's curve is more bizarre. It looks like two normal distributions connecting each other, with a pit in the middle, and his curve seems to be growing in a rather steady trend, which makes his future game anticipating, and the curve is definitely worth investigating.

5) Number of Matches Won by Each Player on Different Types of Surface



From the plot, we can see that each player has different types of the court they are good at. There are four types of courts that are mainly used in tournaments worldwide. Clay courts are traditionally used in the French Open. These courts characteristically have a slower game and give balls a higher bounce. They tend to favor baseline players.

The grass is the fastest type of court because of its low bounce capacity. This means that players with stronger serve-and-volley skills will generally perform better. The grass court is the signature of Wimbledon.

Hard courts can range from faster to slower speeds depending on the quantity and size of sand mixed into the paint coating. An acrylic hard court is used in the US Open and a synthetic for the Australian Open. Carpet courts are removable tennis court surfaces. In general, carpeted courts make for a fast game.

Comparing and contrasting the win records of the five players, we can see that Federer has the best performance on the carpet court. As for clay courts, Nadal far exceeds other players, which shows that he is better at strength, and he is a good baseline player.

Works Cited

https://en.wikipedia.org/wiki/Tennis
https://sportsbyapt.com/types-tennis-courts/
https://www.kaggle.com/manish2104/tennis-grand-slams-data
https://www.masterclass.com/articles/getting-to-the-majors#what-is-a-grand-slam-in-tennis

Tennis Player Performance Analysis

Professor Adam G. Anderson
Digital Humanities 100

Xinyue (Sherry) Liu
05/28/2021

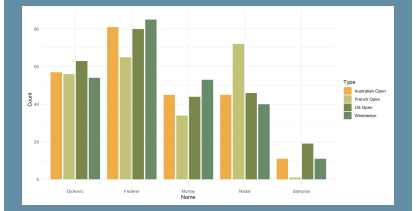
Federer and Djokovic have a similar win rate with each other, followed by Murray and Sampras. As for the grass court, Federer has outperformed the other players. Since grass is the fastest type of court, it proves that Federer is better at serve-and-volley skills and has better flexibility and agility.

From this data, it seems like hard courts are the most used court in the tournaments. Federer is also better at this type of court compared to other players. It makes sense since hard courts also have fast speeds compared to clay and grass courts.

This favors players that have better mobility and agility, but baseline players can also take advantage of this type of court, since it also has the high bouncing capacity that is shared by clay courts.

Admittedly, there might be a limitation to this dataset. Djokovic, Federer, Murray, and Nadal are players of our generation, yet Pete Sampras is a player from decades ago. Therefore, this dataset may not record all the matches that he has been played, resulting in a low count of winning. Therefore, this data is not as representative compared to the other players that I have included in this bar plot.

6) Number of Matches Won by Each Player in Four Major Tournaments



The Australian Open is the first Grand Slam event of the year. The tournament is held in Melbourne over two weeks in mid-January. The French Open, also known as Roland Garros, is a two-week event that takes place towards the end of May in Paris, France. This tournament is the only major championship to utilize the advantage set to determine the winner of the match, and the only one played exclusively on outdoor clay court surfaces.

Wimbledon is the oldest tennis championship in the world, commonly referred to as “The Championships.” Wimbledon takes place in late June/early July and is the only major tournament since 1988 to take place on a grass court. Last but not least, the US Open takes place on the last Monday of August, stretching over two weeks. This tournament is played on hard courts. The US Open is the only major tournament to use the 12-point tiebreak scoring system.

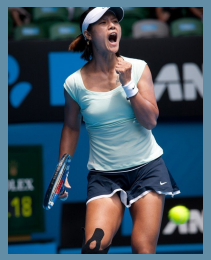
We can see the data from this graph correspond to the conclusion that we have reached from the last graph: Nadal won more times than others in the French Open, which uses clay courts.

Federer has a good performance in every major tournament except the French Open since he is good at the carpet, grass, and hard courts. Compared to the others, Djokovic seems to have the most stable performances in each type of tournament, unlike Nadal, who has a great advantage in French Open. Both Murray and Sampras do not perform very well on French Open, as well as Federer.

We are able to find a pattern and group players into three different types by taking a closer look at this graph. We can see that the winning rate distribution of Federer, Murray, and Sampras looks similar to each other. They are both good at all the other courts aside from the clay court, so the three of them fall into the same category.

On the opposite side, Nadal is good at clay courts and performs better on clay courts than any other types of courts. Therefore, players like Nadal fall into another category.

Lastly, Djokovic has stable performance in every type of tournament. So, players like Djokovic that perform equally well on any type of courts and any type of tournaments fall into the third category.



7) Number of Times Each Player Won the Grand Slam or Was the Runner-Up

In this wordcloud, the name Roger Federer and Rafael Nadal definitely pop up, with several other noteworthy names such as Novak Djokovic, Pete Sampras, Ivan Lendl, and Jimmy Connors.

By comparing these two plots, we can notice that the outstanding bars are almost the same. The first five highest bars correspond to Roger Federer, Rafael Nadal, Novak Djokovic, Ivan Lendl, and Pete Sampras. And the counts are 30, 24, 21, 19, and 18 respectively. They outperform the other players in terms of the number of grand slams and the number of being the runner-up, since more than eighty percent of the players have a count below 15.

9. Limitations

- There might be missing values in the dataset. For some years, not all the champions were recorded.
- There are limited features in this dataset. The whole dataset only contains four columns. Therefore the information we can extract from this dataset is very limited.
- The size of this data is relatively small. The data only contains the four major tournaments from 1877 to 2018. Although it seems to be a very long period, yet the number of tournaments and matches recorded is limited.

10. Summary

Through the data analysis from five different perspectives, we have gained an understanding of the general trend of the tennis tournaments, the distribution of players' performances, and the differences between each player's performance.

By looking at each result, it is not hard to reach the conclusion that Roger Federer and Rafael Nadal are definitely the two most outstanding tennis players among all the other tennis players. Novak Djokovic and Pete Sampras are very noteworthy as well, especially from the violin plot we can see that there is still more growing space for Djokovic in the future. What is also noteworthy is that among the first four outstanding tennis players, Pete Sampras is the only player that is not of our generation. This leads to the conclusion that the competition indeed has become more fierce during recent decades.

Admittedly, there are some limitations to this dataset, such as the size of the data is not very large, and there might be some missing data that could impair our decision, but in general, the dataset is quite trustworthy. By just looking at the dataset and by conducting exploratory data analysis, I have examined that there are little flaws and mistakes in this dataset, so the conclusion should be rather accurate and trustworthy.

Works Cited

https://en.wikipedia.org/wiki/Tennis
https://thenounproject.com/search/?q=tennis
https://www.kaggle.com/manish2104/tennis-grand-slams-data
https://nycdatascience.com/blog/student-works/tennis-player-performance-analysis/