Andy McDill

10/19/2023

Greatest Tennis Player of All Time



**Table of Contents**

[**1 Introduction 2**](#_Toc148612922)

[**1.1 Purpose of Analysis 2**](#_Toc148612923)

[**1.2 Original Records 2**](#_Toc148612924)

[**2 Data Preprocessing 3**](#_Toc148612925)

[**2.1 Derivations/Transformations 3**](#_Toc148612926)

[**2.2 Data Errors 4**](#_Toc148612927)

[**3 Analysis 4**](#_Toc148612928)

[**3.1 Overall Success 4**](#_Toc148612929)

[**3.2 Circumstantial Success 5**](#_Toc148612930)

[**3.3 Domination factor 7**](#_Toc148612931)

[**3.4 Career Impedances 7**](#_Toc148612932)

[**4 Conclusion 8**](#_Toc148612933)

[**5 Bibliography 8**](#_Toc148612934)

# Introduction

This report is designed to explain the rationale and techniques behind the analysis that was completed to address an ongoing controversy within the tennis community. Data preprocessing, completed in SAS Enterprise Guide 7.1, and visualizations, generated in Tableau Desktop 2023.2.2, were utilized in order to draw accurate conclusions.

## Purpose of Analysis

Beginning around 2005 and continuing to present day, we have lived in a significant time within the world of tennis, a golden era per se, in which three players that have a chance to hold the title of greatest of all time (GOAT) overlap their careers nearly entirely. These players I am referring to are Roger Federer (2000 – 2023), Rafael Nadal (2005 – Present), and Novak Djokovic (2006 – Present).

Any admirer and/or athlete of tennis are generally a fanatic of one of the three GOAT prospects and will not hesitate to defend their case on why their favorite player should be awarded this title. These arguments undoubtedly are filled with bias based on the players style, technique, appearance, highlights, and overall demeanor; therefore, a detailed analysis using unbiased data is necessary to ensure the conclusion of this controversy.

## Original Records

The original data source contained 60,000+ records of every tennis match from 2000 – 2023 and is updated daily. The source code and dataset were found on the popular data science forum, Kaggle.

The original variables that were captured are described in the table below:

|  |  |  |
| --- | --- | --- |
| Variable | Data Type | Description |
| Tournament | Categorical | Name of tournament |
| Date | Date (YYYY-MM-DD) | Date match is played |
| Series | Categorical | Series of tournament (ATP 250, 500, Masters, Grand Slam) |
| Court | Categorical | Indoor or Outdoor |
| Surface | Categorical | Surface of court (Hard, clay, grass, carpet) |
| Round | Categorical | Round of the tournament |
| Best of | Numeric | Number of sets required to win (3 or 5) |
| Player\_1 | Categorical | First player in the match |
| Player\_2 | Categorical | Second player in the match |
| Winner | Categorical | Winner of the match |
| Rank\_1 | Numeric | ATP ranking of Player\_1 |
| Rank \_2 | Numeric | ATP ranking of Player\_2 |
| Pts\_1 | Numeric | ATP points of Player\_1 |
| Pts\_2 | Numeric | ATP points of Player\_2 |
| Odds\_1 | Numeric | Odds that Player\_1 wins the match |
| Odds\_2 | Numeric | Odds that Player\_2 wins the match |
| Score | Categorical | Final score of the match |

# Data Preprocessing

## Derivations/Transformations

The original data that was obtained via Kaggle was implemented into my own Python 3.9.12 interface and updated to output a filtered dataset that included only matches that Federer, Nadal, and Djokovic competed in. The reason for doing this was because a few important matches were being removed due to another variable that was unseen in the Kaggle dataset.

This new filtered dataset was imported into SAS to commence the data transformations and derivations to properly conduct the analysis.

Initially, the variable *series* was provided with more than the four categories that the tennis community is familiar with. *Tourney\_weight,* a numeric variable, was created to hold the weight of each tournament in exactly four categories. 250 was assigned when *series* was ‘International’ or ‘ATP250’, 500 was assigned when *series* was ‘International Gold’ or ‘ATP500’, 1000 was assigned when *series* was ‘Masters’ or ‘Masters 1000’, and 1500 was assigned when *series* was ‘Grand Slam’ or ‘Masters Cup’. These tournament weights represent the ATP points awarded when a player wins the tournament and are indicative of the difficulty and importance.

*Game\_ratio\_1* and *game\_ratio\_2* were created from *score* as a calculation of total games won over total games lost for each match played by one of the GOAT prospects. The \_1 and \_2 relate to if our players in question fall under *Player\_1* or *Player\_2*.

Once these variables were created, a transpose was done around *Player\_1* and *Player\_2* to create an additional record each time one of the GOAT’s played against each other and to create new variables that pertain to only the three prospects as opposed to any player within *Player\_1* and *Player\_2*. These variables were *player* (containing only ‘Federer R.’, ‘Nadal R.’, or ‘Djokovic N.’), *rank*, *points*, *odds,* and *game\_ratio*. This dataset was output as **tennis\_final1** and was used to create a majority of the analysis in Tableau.

In order to evaluate a player’s entire career time line, an additional dataset needed to be created. The Association of Tennis Professional’s (ATP’s) process is that a player’s points are updated on Monday of each week and their rankings are adjusted as such compared to other players. The structure of the data was that each match was recorded on the day of the match along with their rankings and points at that time as well; however, some weeks of the year did not have a match and therefore were missing. Therefore, a new dataset that contained all the weeks of each year from 2000 to 2023 was merged onto the original dataset to fill in these missing weeks. The *rank* variable then needed to be filled in for these missing weeks and were done so by using the LAG function in SAS to capture the previous value and assign it to the current record. The LAG function was iterated 50 times to ensure all the missing values of *rank* would be assigned a value but this would result in overcounting. To counteract this, the number 1 ranked player was taken from the source dataset (prior to filtering in Python) and used to prevent any one of the GOAT prospects from obtaining too many records at world number 1.

All the weeks from each year were included in this dataset so another variable was created, *absent,* as a flag to indicate the start and stop time for a period where one of the three players were absent for more than 12 weeks. Finally, this dataset was output as **tennis\_final2** and was used to complete the rest of the analysis in Tableau.

## Data Errors

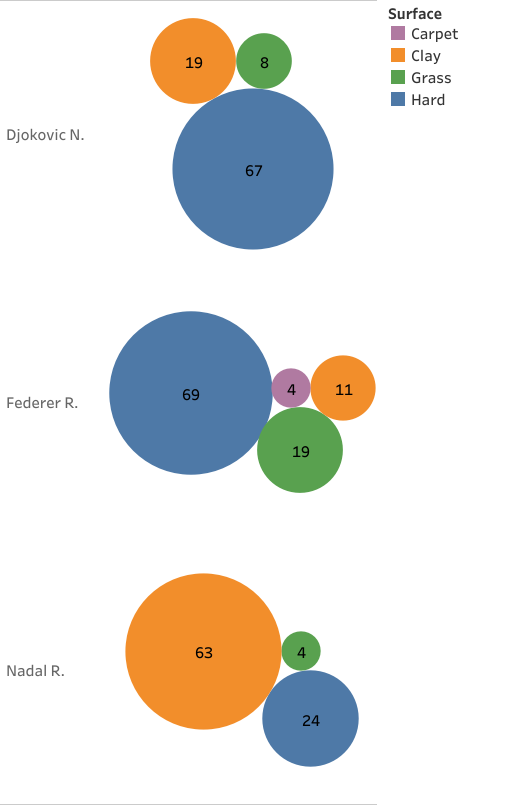
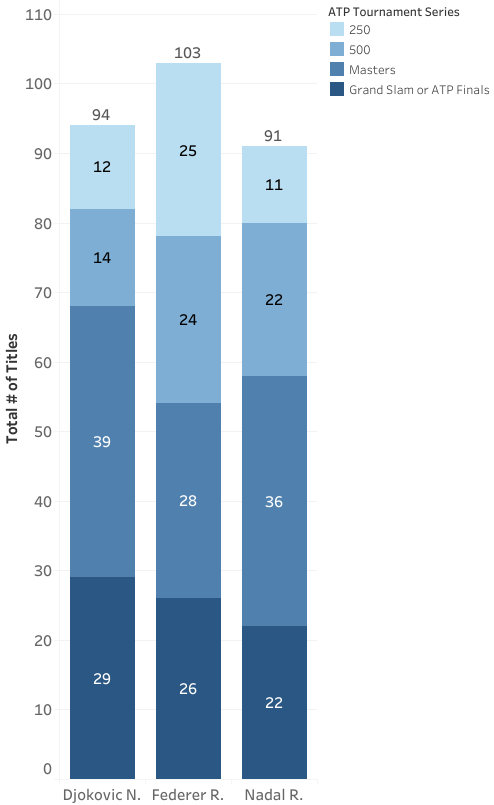
An error was found in the dataset where the China Open in 2013 was recorded as 2012, which caused the China Open to occur twice in 2012. This data error was manually fixed by researching the true records during this tournament.

# Analysis

## Overall Success

A large portion of the current controversy has always revolved around the number of titles a player has won, especially the number of grand slam titles. The four grand slam tournaments throughout the year are the Australian Open, French Open (Roland Garros), Wimbledon, and U.S. Open and these tournaments are the most difficult to win and award the highest amount of ATP points. However, there are more factors to address when determining the GOAT and evaluating a player’s overall success throughout their career.

The figures below show the total number of titles won, broken up by the tournament series and the court surface.



At first glance, Roger Federer has won the most titles, however, he has fully completed his career and has been on the tour the longest. Novak Djokovic has won the second most titles whilst his career is still ongoing and has also won more Masters 1000 and Grand Slam events than Federer and Nadal. Rafael Nadal has won a large amount of tournaments as well while his career remains unfinished, however, a vast majority of his victories are on clay courts, hence the reason for his title “King of Clay”. Regardless of his 91 titles, his career has lacked the versatility required to be labelled as the GOAT.

A screenshot of a social media post

Description automatically generatedA screenshot of a computer screen

Description automatically generated

To further understand each players success throughout their career, we should evaluate how long they were ranked at number 1 and their total career record. Novak Djokovic is the victor of both these categories by maintaining the number 1 ranking for the most number of weeks by a large margin and by having the largest win/loss ratio.

## Circumstantial Success

A few more factors can be understood that aren’t always very direct and these are the rankings of the GOAT prospects opponents, how many times were they eliminated early in a tournament and what was their record against other GOATs.

A screenshot of a table

Description automatically generated

A screenshot of a graph

Description automatically generatedA screenshot of a white and black text

Description automatically generated

Djokovic has won the most against players with a higher ranking by a small margin compared to Nadal but has fallen short to Federer when losing prior to the quarterfinals in a tournament. Regardless, Djokovic has the highest win rate against Federer and Nadal which places him at the top amidst this debate.

## Domination factor

One last evaluation is required for a players career and that is their domination factor. The score of each match was used to calculate a ratio of total games won over total games lost and this is visualized below.

A graph with blue and white lines

Description automatically generated

Each player looks to be equivalent for the domination factor with a few exceptions which indicates how close each of them are in skill level and how talented all professional tennis players are.

## Career Impedances

Steering away from the success within their career, some things may have happened throughout their career that impeded them from achieving more success. The figure below is their rankings throughout their career and it is shown if there were any dips in performance which would indicate that something is not quite right. And the dots represent a gap in their career for more than 3 months in case of an injury.

A screenshot of a graph

Description automatically generated

Federer dipped in the rankings at the end of 2016 to 2017, Djokovic lost steam from 2017 to 2018 and Nadal did not have any significant fluctuation throughout his career besides the natural ups and downs within a competitive sport. Not to mention, they all were unable to play for a portion in 2020.

# Conclusion

The ongoing controversy can be laid to rest now that we have detailed the important aspects of Federer, Djokovic, and Nadal’s career. The true greatest tennis player of all time can be awarded to Novak Djokovic who has won the most important titles, maintained the number one ranking for the longest period of time, and is the most versatile player on the tour. Now we can only speculate if another player will ever surpass his records.

# Bibliography

*Kaggle*. (n.d.). Retrieved from Kaggle: https://www.kaggle.com/datasets/dissfya/atp-tennis-2000-2023daily-pull