INTRODUCTORY PROGRAMMING FOR DATA SCIENCE

FINAL PROJECT:

EXPLORATORY ANALYSIS OF ATP TOUR DATA

Ollscoil
Teicneolaíochta
an Atlantaigh

Atlantic
Technological
University

Shane Brady(S00278397),
MSc Computing (Data Science),

DEPARTMENT OF COMPUTING,
& ELECTRONIC ENGINEERING,
ATLANTIC TECHNOLOGICAL UNIVERSITY

2025

Contents

Project Scope

In this project, a statistical analysis of accessible data relating to the ATP (Association of Tennis Professionals) Tour in the 21^{st} century[1] was performed. The data used has been gathered from publicly available datasets from 'Kaggle' and has been supplemented with various information, gathered through web scraping. The goal of this project is to investigate trends, correlations and patterns in professional men's tennis performance, player attributes, and ranking dynamics.

The analysis will cover areas such as physical attributes of players (specifically height and weight of players), geographic trends, generational performance, ranking dynamics, surface specific performance and a basic predictive model to model the probability of future match wins.

Together these analyses will provide a data driven perspective on the evolving nature of men's professional tennis over the past two decades.

Description of the Data

Firstly, before delving into the specifics of the statistical analysis, one must understand what information there is, how this information is being gathered, and how to store the data in a meaningful way. The primary data source for this project will be the Kaggle dataset 'atp_tennis.csv' from 'ATP Tennis 2000 - 2025 Daily update'[1]. This dataset comprises of historical match data spanning 25 years. Within the dataset, specific information such as tournament name, date, surface type, players, winner, rankings and score can be found. This dataset is well structured and can be easily loaded using the *pandas* library.

This data is then supplemented with various other available information such as players height, players weight, players nationalities and players age. Searching the web for this information, two reliable websites, 'tennisexplorer.com' [2] and 'tenniscompanion.org' [3] [4] were found. Using libraries such as *BeautifulSoup* for web scraping and *pandas* for data manipulation, this information can be extracted from these websites and temporarily stored as data frames.

Once the additional data is collected, it is incorporated into the main dataset as a table in the database, which will be discussed in the following section.

When extracting the data from the various websites mentioned, it is important to check the code for errors as one attempts to gather the information. Checks such as raising errors based on HTTP status code's during the scraping process. Additionally some pieces of information, such as player nationalities were split over multiple pages on the website and so a check to make sure all tables have been gathered can be executed. With the information spread over multiple pages, the column headers can be retrieved for the first table only or be hard coded so that no further issues may arise.

Database

Once the data has been loaded in, it is important to store the information from the various sources in a meaningful way. A way that can be easily accessed and formatted correctly, so that the analyses can be done efficiently on the data gathered. The first step is to add the data from the different sources to a database. In this project the database will be named 'tennis_data.db'. Each

[1]The data used pertains to the period from January 2000 up until April 2025

source of information will represent a table in the database. In total, the 'tennis_data' database will contain the following tables; 'birthdate', 'height_weight', 'matches' and 'nationalities'.

To ensure consistency across the data, some formatting adjustments were necessary. For example in the 'matches' table it can be seen that the players names and the name of the winner is stored as 'Last Name' and initial of first name i.e. 'Alcaraz C.' , whereas in other tables it is stored in the traditional manner of 'First Name' and 'Last Name'. In order to perform the analysis, specifically referring to the JOIN function in SQL the data must be stored in the same manner. To resolve this potential issue, an extra column was added to the relevant tables with uniformly formatted names.

Analysis/Visualization

A wide range of analyses and visualisations were conducted in this project so that insights may be extracted from the data. These analyses look at player performance, demographic trends and match characteristics, which help us visualise and differentiate between the current top ranking tennis players. A breakdown of the analyses done an the visualisations performed will be shown in this section.

Physical Attributes and Ranking: Players were grouped into ranking bands (1-10, 11-20 and so on for the top 100 players) to investigate whether the players average height and average weight played a part in their success. This investigation was performed as some people may see height to be a significant advantage when it comes to service games and overall coverage of the court, whereas for weight, a look at whether stronger but heavier players may perform better than lighter players (potentially quicker and more agile) was done. To visualise the results of the analysis, a combined bar chart was plotted so that any significant changes in features may be identified easily. Looking at the results of the analysis, it can be seen that for the lower ranked players (lower the better) their average height is taller than those ranked further down the list. The first four groups (rankings 1-40) all have average heights 187 cm or above, whereas from groups 5 onwards the average height is below 187 cm. Taking a look at the average weight of the groups, it is not so obvious whether this plays a part in the players success or not.

Country Vs. Success: This analysis aimed to identify which countries are most successful in producing top tennis players. Using information on the top 100 ranked tennis players, players were grouped by nationality and the total number of match wins for each country was calculated. This analysis was conducted to see if certain countries consistently produce high-performing players, which could be due to stronger tennis programs or greater national investment in the sport. Visualising the results, a pie chart was used to display the percentage wins per country for the top 100 ranked players. It can be seen that certain countries like Serbia, USA and France hold large portions of the total wins. Serbia's significant share may be a results one specific tennis player, arguably the best of all time, 'Novak Djokovic' and his dominance over the past 25 years, whereas the USA seem to produce a greater number of successful tennis players overall.

Ranking Trends Over Time: For this analysis, the ATP rankings of the current top 10 players have changed over the past year were tracked. This topic was chosen because rankings are a direct measure of a player's performance over time, and visualising these changes can reveal which players are consistent, improving, or declining. A heat map was used to show monthly changes in rankings, where lighter colours represent higher rankings (i.e., better positions). The heat map clearly shows that some players have maintained stable top-10 positions, while others

have moved up or down significantly throughout the year. Players such as 'Jack Draper' and 'Lorenzo Musetti' both young players have had significant movement over the year, due to recent success whereas players such as 'Carlos Alcaraz' and 'Jannik Sinner' have been consistently at the top of their respective games for the period of the year. In the heat map, it can be seen that there are some months with no data. This is the result of a player being injured and not competing for that month (most likely) or a player recently breaking into the top 10 (i.e. 'Alex De Minaur') or a player suspension (which is the case of 'Jannik Sinner's' most recent missing months).

Younger Vs. Older Players: This section explores the hypothesis that younger players (< 25) may be performing better than older players in recent years. The players were divided into age groups and a comparison of their win counts was done. The idea was to investigate whether a generational shift is occurring in professional tennis, where younger athletes are increasingly successful wining more matches and in turn climbing the rankings. A basic hypothesis test (T-test) was performed using *scipy.stats* and the results showed that there is NOT strong enough evidence to suggest that younger players are performing better than older players (p-value greater than 0.05). It can be said that experience still plays a significant role, and older players often remain among the highest-ranked. However, there are signs of increased impact from younger competitors in the top rankings in the last few seasons.

Performance by Surface: In this analysis, the performance of the current top 10 players on different surfaces—clay, grass, and hard courts was examined. The reason for this analysis was to understand whether certain players are surface specialists or whether they are consistent across all conditions. Win percentages were calculated separately for each surface type and visualised using a combined bar chart for easy comparison. The results show clear surface preferences for some players—for example, stronger performance on clay or hard courts ('Jack Draper' on hard and 'Casper Ruud' on clay)—while others demonstrate a high level of versatility across all surfaces ('Novak Djokovic' and 'Taylor Fritz').

Match Length by Surface: This analysis looked at whether matches tend to last longer on certain surfaces. The number of games per set and sets per match were used in order to calculate the average match duration by surface type. The motivation behind this was that surface type can affect playing style and rally length. From the results it is observed that grass courts on average had the longest matches in terms of number of games played, while clay and carpet were the shortest. This finding is somewhat counterintuitive, as grass courts are known for their faster ball speeds, which typically lead to shorter rallies. However, the speed of the surface may also make service games more dominant and harder to break, potentially resulting in more extended matches with additional games required to determine a winner. In contrast, the slower pace of clay courts allows for longer rallies and reduces the server's advantage, making it easier for players to break serve and conclude sets more quickly.

Match Outcome Prediction: Exploring predictive modelling for the tennis data gathered, a basic logistic regression model was created to predict match outcomes based on two main features: the difference in ATP rankings and the difference in surface-specific win percentages between the two players. The idea was to see whether these two features alone could predict with reasonable accuracy and precision. While the model is relatively simple, it did show some ability to predict outcomes correctly, suggesting that surface expertise and ranking gaps are meaningful indicators of match results.

Conclusion

This project looked at investigating a wide range of analytical perspectives on professional mens tennis, using relevant data like player statistics, match results, and features like surface type and nationality. From conducting these analyses meaningful insights were drawn about factors affecting player performance, success and match outcomes.

Physical characteristics such as height showed some correlation with higher rankings, suggesting a possible advantage in terms of success while weight was not such a decisive factor. Looking at the nationalities of the players it could be shown that certain nationalities held significant portions of wins over the past 25 years, displaying the dominance of these countries in producing successful tennis players.

Time and surface-based analyses gave an insight into how player performance can change over time and on specific surfaces. Tracking the rankings of the top 10 ranked players, allowed us to determine the longevity and consistency of the players, while surface specific performance analysis identified clear preferences and strengths that could be related to the playing style of that particular player. Contrary to conventional expectations, match duration analysis showed that on average, matches played on grass tend to be longer than those on hard and clay despite being considered the fastest surface.

Using a simple T-test to test the hypothesis that younger players (< 25) are performing better than older players, concluded that there is NOT strong evidence to back this claim, which leads me to believe that experience is still crucial in determining the results of matches.

Overall, this project demonstrates how important data driven approaches are in understanding the nature of tennis. In the future with more time, a better predictive model may be developed (more meanignful features) so that higher accuracy and higher precision may be obtained in predicting the winner of a particular match. Finally, more data may also be obtained in order to conduct a more complete and comprehensive analysis of all aspects of the game of tennis.

References

[1] *ATP Tennis 2000 - 2025 Daily update — kaggle.com*. https://www.kaggle.com/datasets/dissfya/atp-tennis-2000-2023daily-pull. [Accessed 10-05-2025].

[2] LiveSport s.r.o. *Tennis Explorer: Tennis Rankings, WTA & ATP Rankings — tennisexplorer.com*. https://www.tennisexplorer.com/ranking/atp-men/. [Accessed 10-05-2025].

[3] *ATP Male Tennis Player Height & Weight Data + Stats — tenniscompanion.org*. https://tenniscompanion.org/players/male/height-and-weight/. [Accessed 10-05-2025].

[4] *A Detailed List Of ATP Player Ages & Birthdays + Stats — tenniscompanion.org*. https://tenniscompanion.org/players/male/birthdays/. [Accessed 10-05-2025].