Tennis: What Makes You Win?

Sin Sze Li
School of Mathematics, Computer Science and Engineering
City, University of London

***Abstract* — In this study, various factors in a tennis match are investigated to determine the most important factor towards match outcome. Match statistics and player's physical characteristics are used for analysis. A match is determined mostly by the percentage of points win. k-means clustering is also used to categorize players according to their match statistics, with three clusters of players corresponding to different levels identified. Finally, match statistics of the GOATs (Greatest Of All Time) is analyzed to observe the change in performance for the best 3 players in the world throughout the years.**

Keywords— tennis, match statistics, k-means clustering

I. Introduction

Tennis is one of the most popular sports in the world. It is complex and exciting that requires both stamina and strategic planning.

It is distinct from other ball games because of its scoring system. A tennis match consists of three scoring level: sets, games and points. There are usually 3 or 5 sets of games in a match, depending on the tournament level. In each set, the player who first won 6 games wins the set. Each game consists of several points, and the player who first won 4 points wins the game. Since tennis is scored set by set, and game by game, winning the majority of points or games does not necessarily mean that player wins the match.

Men's professional tennis is governed by the Association of Tennis Professionals (ATP). There are about 68 tournaments each year. ATP players compete on different tournaments to gain ATP ranking points. The higher the tournament level, the more ranking points the player gains if he won. The ATP rankings are updated every Monday.

At the moment, Roger Federer, Rafael Nadal and Novak Djokovic are considered the GOATs (Greatest Of All Time) in tennis. Their exceptional performance on court makes them on the top of the ATP rankings constantly.

In this study, data of ATP matches are used to find out several possible factors that affect match results. The first group of factors is the physical characteristics of players, like height and handedness. The other group of factors is the match statistics, for examples, the percentage of first serve and the percentage of points won. These are more related the strategies players adopt in a tennis match. The findings may be insightful for players who wish to improve their games and increase their chances of winning.

II. Analytical Questions

For the first group of factors, the following analytical questions are raised:

1. Do left-handed players have higher chance in winning than right-handed players?
2. Do taller tennis players serve better? Do they win more ace points and commit fewer double faults?

These questions are interesting because the factors are physical characteristics of players, so they are determined mainly by genetics and cannot hardly be changed by training. The answers to these questions may give an idea of how much hard work and practice contribute to winning a tennis match.

For the second group of factors, the following analytical questions may provide some interesting findings:

3. What are the most important factors that affect match outcomes?
4. Is it possible to categorize players based on their match statistics?
5. What makes the GOATs the greatest tennis players in the world?

The findings of these questions may give a more general understanding on what controls the result of a tennis match. Based on that, some areas of training can be derived for players to work on, so as to improve their games more effectively.

The data used in this study are available on DataHub [1] and GitHub [2]. Data from DataHub contains overviews of all ATP players, while additional players information, ATP tournament match statistics and weekly ATP rankings from 1968 to 2021 can be found from GitHub.

III. Data (Materials)

For simplicity, data from 2010 to 2020 are used for analytical questions 1 to 4, while data from 2000 to 2020 will be used for analytical question 5. Temporal data can provide another perspective to the analysis in terms of the variable's trends over the years. The data are scattered in numerous csv files which are then combined into three files corresponding to players overview dataset, rankings dataset and match dataset.

Players overview dataset contains information of 673 players, including their names, id, handedness, backhand styles and heights. Using each player's unique id, these data are merged to the match data and rankings data for further analysis.

In rankings dataset, there are 996561 rows. Each row contains the ranking date, player id, rank and ranking points. The amount of row is huge because ATP rankings are updated weekly, and there are over 10 years of data. From here the change in ranks of players with different handedness and backhand styles are compared using line plots.

In match dataset, there are 23726 rows and 48 columns. Each row represents an ATP tournament match. Information about the match including tourney name, surface of court, draw size, date and duration can be found. Moreover, winner's and loser's overview and match statistics such as their seed numbers, ranks, number of serve points, total points won, break points saved and aces are included. Correlation between each variable with the match result is used to find out the determining factors of winning or losing a match. k-means clustering is also used as an attempt to put players into different groups based on their match statistics. Finally, the changes in the GOATs' match statistics are visualized to find out how they manage to keep being on top of their games.

IV. ANALYSIS

To find out the answers to the analytical questions, a series of analysis is done as followed:

A. Data preparation

The raw data downloaded from DataHub and GitHub were not clean. They contained loads of missing values and outliers. These must be dealt with in order to achieve ready-to-use datasets as described in the previous part.

Missing values are handled depending on the importance of their corresponding variables. For example, in the match dataset, most of the missing values come from player's entry method. Since entry method is related to player's seed and player's rank, and these two variables are already present in the dataset with fewer missing values, the entire entry method columns are deleted. On the other hand, missing values can also be found in variables such as player's handedness and heights, which are the foci of this study. Therefore, it is unreasonable to delete the columns or replace the missing values with 0. Fortunately, those missing values only contribute to 5% to the total number of rows, these rows were deleted.

To search for outliers, boxplots for the variables are used. Figure 1 shows the boxplot for players height in cm. It is very unlikely that an ATP players have height of nearly 0, hence, those two datapoints are identified as outliers and the corresponding rows are deleted.

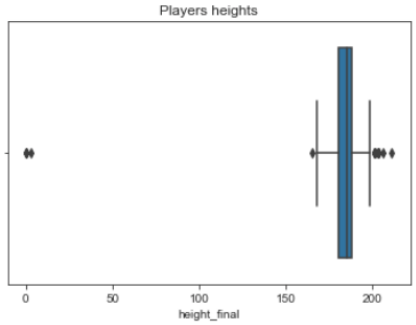

Fig 1. Distribution of players heights before data cleaning

After cleaning the data, distribution of players handedness and backhand styles can be shown using bar charts as in figures 2 and 3.

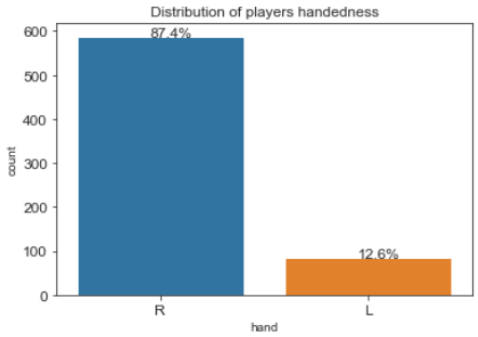

Fig 2. Distribution of players handedness

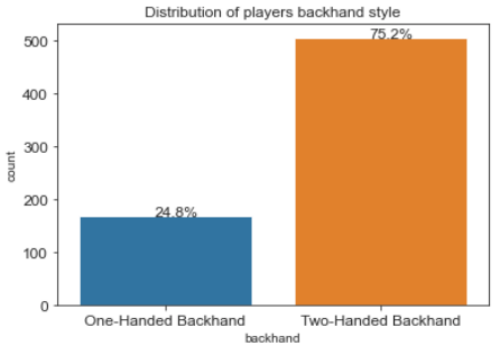

Fig 3. Distribution of players backhand style

B. Data derivation

In the match dataset, the statistics are presented as raw numbers, meaning there are absolute numbers for serve points, total points won, break point saved and aces etc. These numbers depend on the duration of the match. The longer the match, the higher these numbers. An extreme example is the Isner-Mahut match at the 2010 Wimbledon first round match. The match lasted for 11 hours and 5 minutes in three days. Isner and Mahut scored 113 and 103 aces respectively, while in a normal-length tennis match, the number of aces is usually about 5 to 10. Therefore, percentages corresponding to these variables are derived. For example, instead of using 'number of first serve in', 'percentage of first serve in' is derived from number of first serve in divided by the total number of serve points.

To compare different players, average match statistics in the last 10 years should be derived for individual players. Moreover, it may be interesting to also include the proportion of time players spend on hard, grass and clay courts. Hence, percentages of matches in hard, grass and clay courts and their corresponding winning percentage are calculated.

C. Construction of models

In order to find out the determining factors affecting a tennis match, the numerical variables are used to construct a correlation heatmap using Spearman's correlation. Part of the correlation heatmap corresponding to match result is selected, and the values of correlation are sorted to find the most correlated factors. The correlations between the top-correlated variables and match result are then visualized using regression plots.

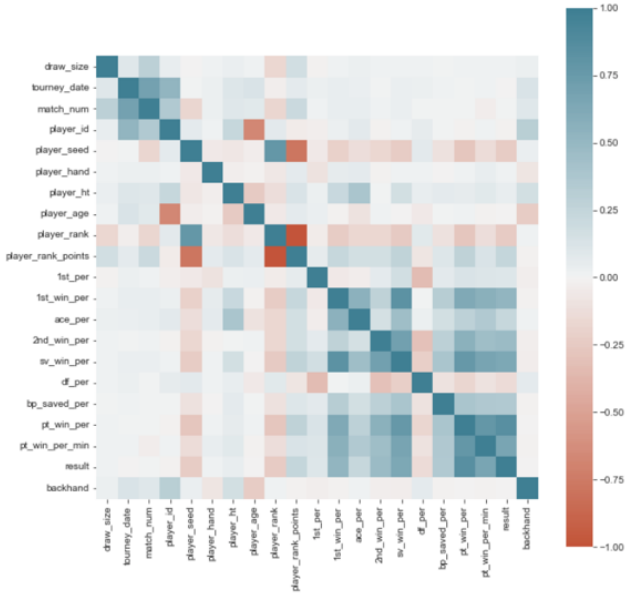

Fig 4. Correlation heatmap using Spearman's correlation

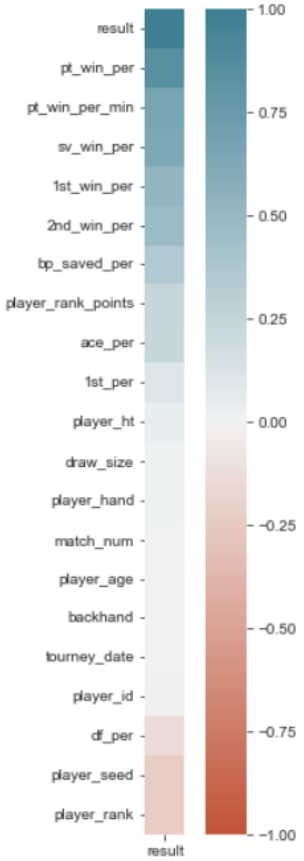

Fig 5. Correlation heatmap for match result

k-means clustering is used to identify different groups of players. Only some of the variables are chosen to train the model. They include player's first serve percentage, first serve win percentage, second serve win percentage, ace percentage, serve points win percentage, double faults percentage, break points saved percentage. The purpose of using k-means clustering is to find out if it is possible to categorize players according to their strategies so only their match statistics but not seed numbers, ranks, player's physical characteristics is used.

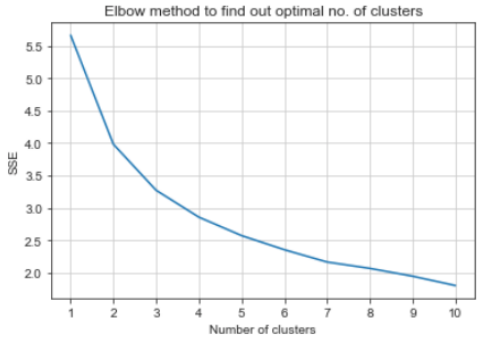

Fig 6. Elbow method for k-means clustering

The elbow method is applied to determine the optimal number of clusters, which is found to be 3. Since the data are multidimensional, PCA is used to visualize the result. The returned classes are assigned back to each player. According to different clusters, the average of players information and match statistics is computed, hence the difference between each cluster can be seen.

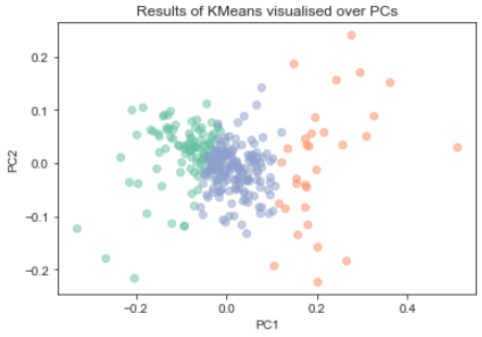

Fig 7. Scatter plot using PCA

D. Validation of results

For the k-means clustering, originally all the players information and match statistics are used to train the model. However, looking at the loadings of the returned labels, players seed numbers and rankings contribute the most in determining the clusters, which is not what the original purpose of clustering in this study.

V. FINDINGS, REFLECTIONS AND FURTHER WORK

A. Findings

Complete finds can be found in the Jupyter notebook. In this report, the findings are expressed according to the analytical questions:

1. Do left-handed players have higher chance in winning than right-handed players?

The difference in winning percentage for different handedness is insignificant. However, it is shown that between year 2015 to 2019, left-handed players have higher ranks than right-handed players. In fact, the average rank over the past 10 years for left-handed players is 471, while that for right-handed players is 501. So, there are more left-handed players in higher ranks than in lower ranks.

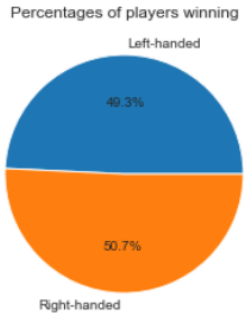

Fig 8. Players winning percentage in the last 10 years matches

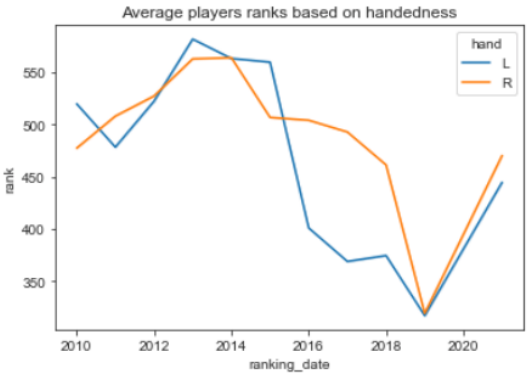

Fig 9. Average player ranks based on handedness

2. Do taller tennis players serve better? Do they win more ace points and commit fewer double faults?

Taller players do tend to win more ace points. A possible explanation is that when taller players serve, the ball has a larger vertical momentum, so it rebounds higher and faster, making it more difficult for the opponent to return the shot. However, double fault percentage does not change much when players height increases.

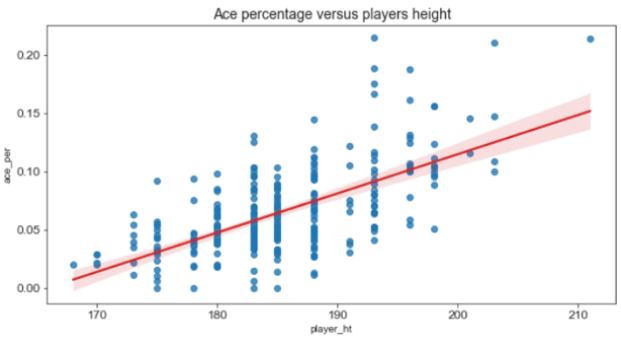

Fig 10. Relationship between ace percentage and player's height

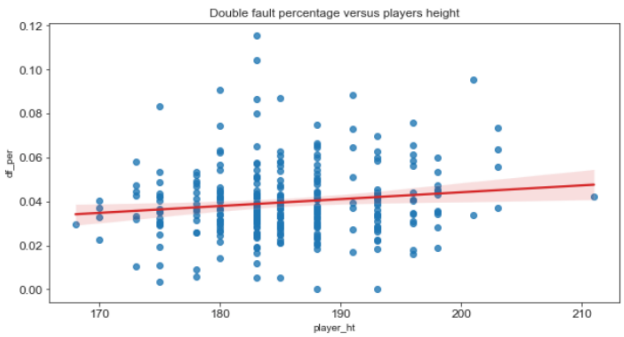

Fig 11. Relationship between double faults and player's height

3. What are the most important factors that affect match outcomes?

According to the correlation heatmap, the percentage of points won correlates the most to the match result, followed by average points win per minutes and percentage of serve points win.

4. Is it possible to categorize players based on their match statistics?

The average ranks for the three clusters are 141, 158 and 184, meaning the clustering model can somehow differentiate different levels of players based on their match statistics. One interesting finding is the percentages of points win for top-ranked, middle-ranked and low-ranked players are 0.49, 0.47 and 0.42, which is surprisingly close. If a player can win just slightly more points, his rank would rise notably.

5. What makes the GOATs the greatest tennis in the world?

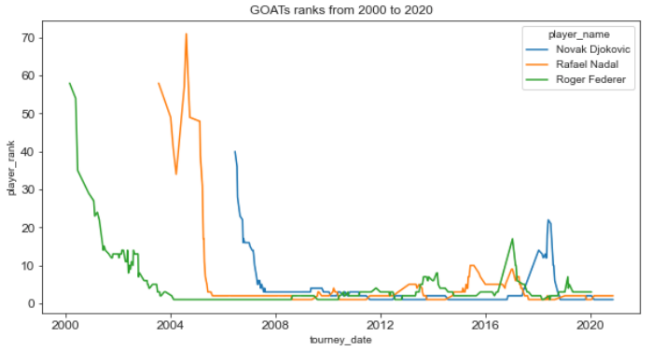

Fig 12. GOATs rank from 2000 to 2020

Starting from 2007, the GOATs are constantly within top 20 players. However, it is shown from the plots that there is no significant change in match statistics before and after 2007 that leads to their rising ranks, maybe because as the previous analytical question suggests, the difference between the match statistics of different ranked players is extremely small. The GOATs have very steady performance, that is why their ranks are constantly high.

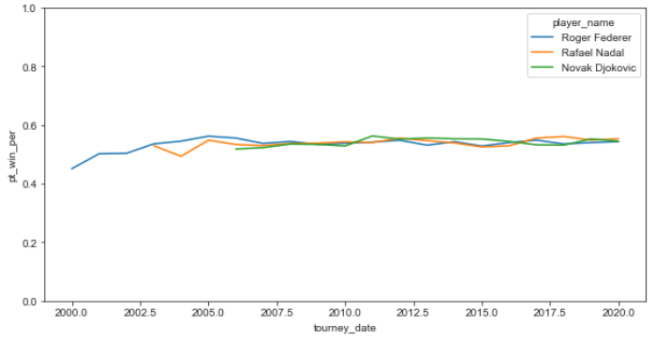

Fig 13. Percentage of points won by GOATs

B. Reflections

First, the datasets used in this study turn out to be less informative than expected because the match statistics is

somehow highly correlated to the match results and tells little about players playing style. Having more statistics about serve speed, distance that a player has run in a match, unforced errors percentage etc. can give a more comprehensive idea of what the player's playing style is [4].

Second, there are more personal qualities [5] in a tennis player which contributes to being a great player, and they cannot be shown in the dataset.

C. Further work

Other than match statistics, player's racket [6] and weather can also be interesting factors to look into. Moreover, it would be interesting to compare the match statistics between male and female players, as 'male players comply more with Prospect Theory exhibiting more loss aversion and reflection effect' [7]. Finally, a predictive model such as decision tree or Naïve Bayes classifier can be trained to predict the outcomes.

REFERENCES

[1] Datopian, "ATP World Tour tennis data," *DataHub*. https://datahub.io/sports-data/atp-world-tour-tennis-data#resource-player_overviews_unindexed (accessed Dec. 22, 2021).

[2] J. Sackmann, *ATP Tennis Rankings, Results and Stats*. 2021. Accessed: Dec. 22, 2021. [Online]. Available: https://github.com/JeffSackmann/tennis_atp

[3] "Federer's Focus: Time Is Of The Essence | ATP Tour | Tennis," *ATP Tour*. https://www.atptour.com/en/news/www.atptour.com/en/news/infosys-match-length-federer-2016-february (accessed Dec. 23, 2021).

[4] Y. Cui, × Miguel-Ángel Gómez, B. Gonçalves, and J. Sampaio, "Performance profiles of professional female tennis players in grand slams," p. e0200591, Jul. 2018, doi: http://dx.doi.org/10.1371/journal.pone.0200591.

[5] J. Higgins, "Why Roger Federer is a GOAT: an account of sporting genius," *J. Philos. Sport*, vol. 45, no. 3, pp. 296–317, Sep. 2018, doi: 10.1080/00948705.2018.1520126.

[6] T. Allen, S. Choppin, and D. Knudson, "A review of tennis racket performance parameters," *Sports Eng.*, vol. 19, no. 1, pp. 1–11, Mar. 2016, doi: 10.1007/s12283-014-0167-x.

[7] N. Anbarci, K. P. Arin, C. Okten, and C. Zenker, "Is Roger Federer more loss averse than Serena Williams?," *Appl. Econ.*, vol. 49, no. 35, pp. 3546–3559, Jul. 2017, doi: 10.1080/00036846.2016.1262527.

WORD COUNTS

Part	Word count
I. Abstract	93
II. Introduction	295
III. Analytical questions and date	227
IV. Data (Materials)	293
V. Analysis	756
VI. Findings, reflections and further work	500