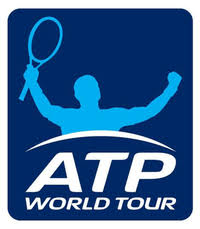
ATP Data Analysis

INSY 6500 Semester Project



By: Nick Colosi

**I. Introduction**

For my semester project, I decided to analyze men’s professional tennis data from the ATP tour. I decided to analyze this dataset primarily because I wanted to look at a sports dataset. Additionally, I picked this specific dataset because it was structured (ready for analysis) and because I have some domain expertise as I played tennis growing up. The goal for this project was to perform an exploratory analysis on the data provided. My objective was to explore different relationships and practice using the skills taught in the course.

**II. Data Overview**

I obtained my data from a [GitHub repository](https://github.com/JeffSackmann/tennis_atp) that I found through another [repository](https://github.com/awesomedata/awesome-public-datasets) that has links to other public datasets. The data was supplied in a series of csv files each containing professional match data from one year. Each record comprises of information regarding one specific match of professional play. The raw data has 49 fields per record, though some of the earlier years had missing points due to certain data collection methods being implemented in the more recent years. A list of the fields can be seen in appendix A. After compiling all the records, the dataset comprised of 156,516 match records.

As mentioned, each record contained information about one match. Each record has match information that describes the year it was played, the tournament it was part of, the players that played, the final score, and the length of the match. Each record also contains player biographical information such as their countries of origin, their heights, their ages, and their dominant hands. Lastly, each match record has statistics for each player containing metrics such as the number of aces, the number of double faults, the number of first serves in, the number of break points saved, and more.

With a plethora of fields, the data presented many options regarding the potential relationships to explore. Because the older match records lacked some of the match statistics, the majority of the analyses did not use the entire dataset.

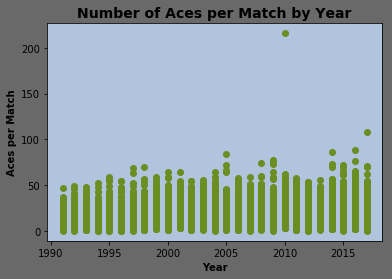
**III. Methodology**

For my project, I exclusively used Python with the Pandas, Matplotlib, and Seaborn packages. The first step I took involved importing the data by looping through a series of csv files and appending the records to an existing dataframe. After loading the data in, I was able to add columns using Pandas to fit the need of my desired analyses. I focused on using plots to explore potential relationships and identify trends in the data.

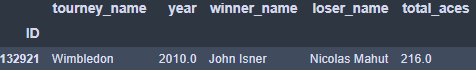
**IV. Analysis**

Aces:

I started my analysis by looking into aces. I picked this part of the game because of its importance to the sport and the difficulty of serving. I began my analysis with a simple scatter plot that showed the total aces per match over time. This plot can be seen below.

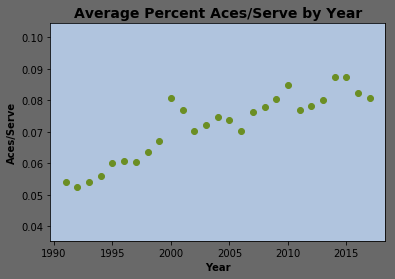


By looking at this plot I was first interested in the data point that appeared to be an outlier. 200 aces in a single match logically seems impossible. I pulled the record by filtering the dataframe to get more information about the match. This record can be seen below.



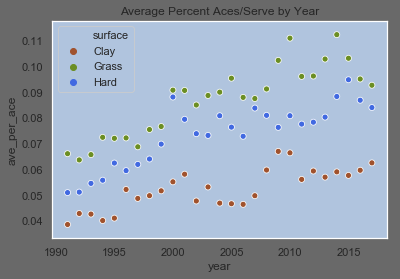
After looking up this match, I found that it was not a typo that occurred during the data entering process. It actually corresponds to the longest match in tennis history that lasted over 11 hours across two days. The final set lasted over eight hours. It was interesting to see this datapoint reflected in reality.

Next, I noticed a possible upward trend of the data showing a potential increase in the ace rate. To investigate further, I plotted the average number of aces per serve grouped by year against the year to see if there was a clearer trend. The resulting scatter plot can be seen below.

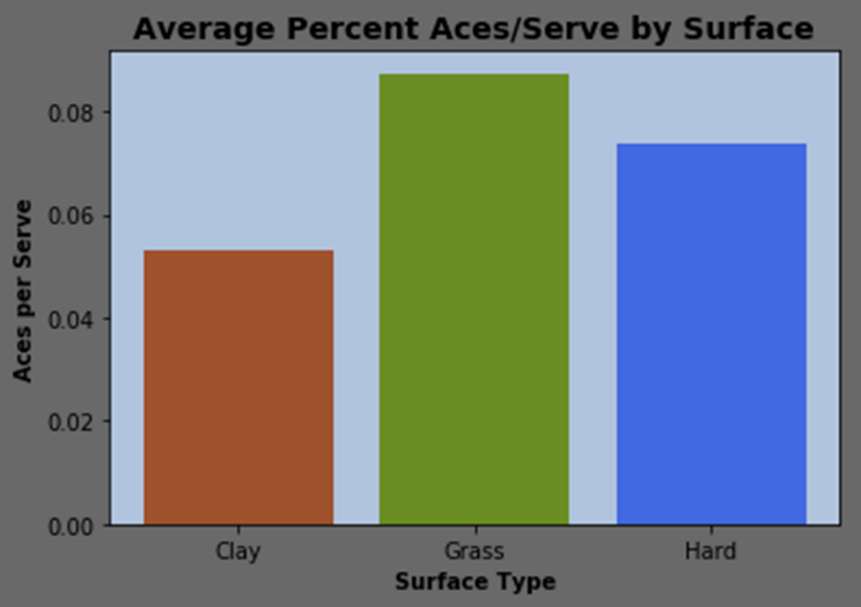


A clear upward trend can be seen showing that the number of aces per serve is increasing. This can be possibly explained by improving racquet technology and better sports science in general. It shows that the game is evolving.

I then wanted to see whether the court type had an effect on aces, so I plotted a similar plot as the one above but also grouped the data points by surface type using Seaborn. The resulting plot can be seen below.



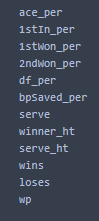
This plot shows that the upward trend of the number of aces being hit per serve persists across all surface types, but also shows a distinct difference in the number of aces per serve between the surface types. Its clear that the most aces are hit on grass and the least on clay. I plotted a bar chart to display the average number of aces per serve by court type, which can be seen below.



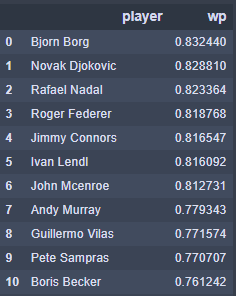
After doing some research, I found clear reason that explained the differences in the aces per serve between the court types. On grass, the ball bounces faster and lower making it more difficult for the returner to get his racquet on the ball. On clay, the ball bounces higher and slower making it easier for the returner to get a racquet on the ball and avoid an ace. Because of the physics behind the court types, players are said to have adjusted their serves to take advantage of the court type. For example, on grass, players hit the ball harder knowing that the opposing player will have a more difficult time returning the ball. On clay, players know that their serve is less effective, so they aim for placement and attempt to set themselves up for an easier show after the returner hits the ball back. Hard courts appear to be more neutral to serves.

Player Statistics:

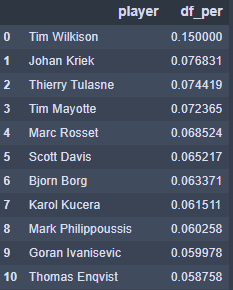
My next analysis focused on analyzing some of the top tennis players of all time. I created a dataframe that contains career statistics for a variety of the metrics provided in the dataset. The list of career statistics I computed can be seen below.



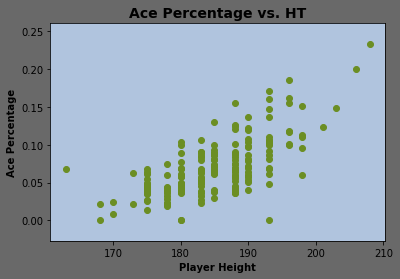
By filtering the data by individual fields, I was able to look at the players with the best of each statistic. The top 10 players in regard to win percentage are shown below.



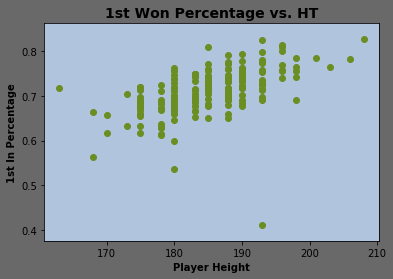
This list of players largely reflects the players who are considered the best of all time. After seeing this table, I looked at other rankings in search of interesting results. I found that Bjorn Borg, the player with the highest win percentage, had the sixth highest double fault percentage, which is displayed below.

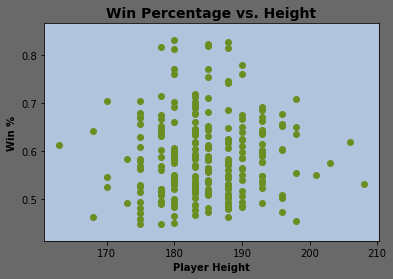


After looking at some of the other statistics I decided to see if any had relationships with each other. I started by looking at the relationship between player height and serving. I chose this relationship because I wanted to see if height truly has an effect on a player’s ability had to serve effectively. Below, I have displayed a plot showing the relationship between height and ace percentage.



The plot clearly shows that taller players have, on average, greater ace percentages. This follows common logic as a taller player has a more favorable angle for serving. Similarly, the plot below shows that height gives players a greater chance of winning his first serve, likely due to the number of aces they hit and the power of the serves they hit.



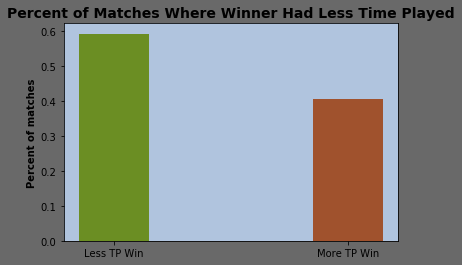
After seeing that height had a clear effect on the ability to serve aces, I wanted to see if taller players were more successful at the sport. To investigate this relationship, I plotted career win percentage against player height. The results can be seen below.

It appears that the players with the best win percentages are not the tallest nor the shortest. They are in the middle of the height spectrum. My logic tells me that this height range lends itself to being agile, having reach and the ability to serve the ball well. Taller players likely struggle with agility and shorter players struggle with reach at the net and the ability to serve at the highest level.

Time Played Analysis:

My last major analysis focused on determining whether the total time played the tournament of each player affected the outcome of a match. To do this analysis, I created a script that looped through the data and computed the total time played in the tournament prior to a given match. While the code I wrote proved to be someone inefficient for the task, I validated the results by collecting sample matches from a variety of matches and manually computing the values.

After computing the proper statistic, I created a new, binary field that was True if the player with less time won and False if the player with more time won. Using this field, I was able to calculate the percent of time each instance occurred. The results are graphed below.



The results show that 60% of the time the player with less time played won the match. This difference tells me that fatigue definitely plays a role in the outcome of a match. A statistic such as this could be used in developing a model to predict the outcome of a match.

**V. Future Work**

Moving forward, I would like to focus on building a model that can be used to predict the outcome of tennis matches based on the data points in this analysis. Through further exploration, I can identify more factors that affect the outcome and select the most prominent in building a model. For example, time played, the round of the tournament, and the differences in players could be used as factors in a model.

**VI. Appendix A**

