**EXECUTIVE SUMMARY**

Serve and return are, perhaps, the most important parts of the game of tennis. Additionally, most tennis matches are decided by a few points, with a player rarely winning more than 60% of the total points played. Therefore, looking at serve, return, and under pressure numbers are fundamental to the sport.

The goal of this project is to analyze how different factors affect how well a tennis players serves, returns, and plays under-pressure, as well as to see how important those numbers really are in terms of how good a tennis player is (based on their ATP ranking). Another goal is to perform a cluster analysis and see if another patterns show up.

**COLLECTING DATA**

All of the data collected came from ATP’s official website. The data was scrapped from ATP.com using python. Selenium and a Google Chrome Driver were necessary to collect the data, due to the pages being dynamically rendered using JavaScript. The Python scripts used for scrapping can be found at the local folder of the project. Two Python scripts were needed - one to scrape serve, return and under pressure stats; another to scrape info from each individual player, due to the diversity of countries, the decision was made to convert each country into the country’s region (EMEA = Europe, Middle East, and Africa; LATAM = Latin-America; APAC = Asia-Pacific; NAm = North America). This conversion had to be done in 3 parts: 1 – collect each player’s country 3 letter code; 2 – convert the 3 letter code into the country’s full name; 3 – convert the country into its respective region.

**DATA CLEANING / PREPROCESSING**

When collecting the data, it had to be separated into different tables. These 5 tables are – player\_info; region; serve\_stats; return\_stats; pressure\_stats. The columns for player’s names and region’s names were deleted where redundant, and a foreign key was added to each table so the tables are connected. The tables are connected in the following way:

serve\_stats (Serve\_Rank PK), (Player\_ID FK)

player\_info

(ATP\_Rank PK), (Region\_ID FK)

region (Region\_ID PK)

return\_stats (Return\_Rank PK), (Player\_ID FK)

pressure\_stats (Pressure\_Rank PK), (Player\_ID FK)

Some players from outside the top 100 appeared on the data, and the info for those had to be manually added to the player’s info table.

**EXPLORATORY DATA ANALYSIS**

**VISUALIZATION #1**

Chart, bar chart

Description automatically generated

This histogram represents the count of players by the region of the country they represent. From looking at it, we can conclude that tennis, at the top level, is dominated by the EMEA (Europe, Middle East, and Africa) region, while the others regions have roughly the same amount of players.

**VISUALIZATION #2**

**Chart, box and whisker chart

Description automatically generated**

By looking at a boxplot of the players’ ages divided by region, we can make some assumptions. Firstly, most players from the APAC region are in the same age group (25-29). The EMEA region has the highest age range distribution (difference in age from the youngest to the oldest player), perhaps due to the higher sample size (it has by far more players than any other region). Finally, the NAm region is likely the most promising, having a lot of players on around the 25 years old range.

**VISUALIZATION #3**

**Chart, box and whisker chart

Description automatically generated**

What stands out from this boxplot is that the NAm region is the region with the heaviest players. Every measure reaches that conclusion: higher mean, higher values for one standard deviation from the mean, and heaviest players. This can be partly due to the two outliers the region possesses, with weights of over 220 pounds (very unusual for a tennis player).

**VISUALIZATION #4**

**Chart, box and whisker chart

Description automatically generated**

What makes this boxplot interesting is the distribution of height on the LATAM region. Most players have very similar height (between 180 and 185 centimeters), with very few exceptions. When it comes to the other regions, the average North-American player seems to be tall, while Asian players are short. In terms of the height range distribution, all other regions (besides LATAM), seem to have different types of players (some tall, some short).

**VISUALIZATION #5**

Chart, scatter chart

Description automatically generated

This plot of weight against height shows that for top tennis players, the expected relationship between the variables is true: the taller a player is, the heavier he is.

**VISUALIZATION #6**

**Chart, scatter chart

Description automatically generated**

Most would agree that the older a player is, the longer he has competed for, and the more experienced he is. This logic leads us to assume that the older a player is, the better results he would have when faced with pressure situations. This plot of age against the player’s pressure rating (calculated by the ATP), shows that the opposite relationship is true: as a player gets older, its pressure rating declines (although not by a lot).

**VISUALIZATION #7**

Chart, scatter chart

Description automatically generated

As expected, the taller a player is, the better are his service stats.

**VISUALIZATION #8**

Chart, scatter chart

Description automatically generated

This relationship shows us that the taller a player is, the worse it tends to do in returning games. This can be explained by the fact that shorter players have lower center of gravity, which helps with their returns.

**VISUALIZATION #9**

Chart, scatter chart

Description automatically generated

One would expect that a higher first serve percentage would result in a higher service games won percentage. The plot shows us that this is not true, with a player’s first serve percentage not seeming to be a determinant factor of the percentage of service games won.

**VISUALIZATION #10**

**Chart, scatter chart

Description automatically generated**

An assumption thrown around in tennis is that taller tennis players, due to their potent serves, tend to have a better chance of winning tiebreaks. This plot tries to explore this relationship. It shows that a player’s tiebreak winning percentage does not change with as a player grows taller.

**VISUALIZATION #11**

Chart, scatter chart

Description automatically generated

This plot explores the relationship between a player’s quality of serve and a player’s quality of return. It shows that a player’s service rating is inversely related with the player’s return rating. That means that the better the players serves, the worse he tends to return.

**VISUALIZATION #12**

Chart, bar chart

Description automatically generated

The bar chart of players divided by their age range shows us that the vast majority of top tennis players are in between the ages of 21 and 35 years old, with the most populated group between 26 and 30 years of age.

**VISUALIZATION #13**

Chart, timeline, bar chart

Description automatically generated with medium confidence

This facet grid of players by age range divided into their respective regions can be used to strengthen the conclusions made with visualization #2 (boxplot of age by region).

**PRINCIPAL COMPONENT ANALYSIS**

For every area analyzed, I attempted to reduce the dimensionality of the data through principal component analysis. In most cases this ended up being redundant, because performing PCA did not successfully reduce the dimensionality of the data with my desired threshold of 95% for the minimum desired variability explained by the principal components. Additionally, not all related columns were used, as ranking and ratings are resulting of the other area-related variables, so they would cause a bias in my analysis. All used variables have been scaled prior to the PCA analysis.

**SERVE**

In this area, performing PCA was helpful, as my 95% threshold was with 4 components, which reduces the variables to be analyzed from 6 to 4. Therefore, the PCA successfully reduced the dimensionality of the data. By creating a scree plot, we can reassure this dimensionality reduction, as the optimal number of principal components is 4.

Through a biplot, we can also observe how the variables being studied relate to the principal components created. I also used a correlation matrix to achieve the same goal.

**RETURN**

When it comes to return data, performing PCA allowed me to reduce the dimensionality of the data from 4 variables to 3. However, the scree plot tells us that, perhaps, the optimal number of principal components to use should be 2. For the remainder of the analysis, I decided to go with 3 principal components for this data due to it surpassing the 95% minimum desired variability explained threshold.

Through a biplot, we can also observe how the variables being studied relate to the principal components created. I also used a correlation matrix to achieve the same goal.

**PRESSURE POINTS**

Finally, data taken regarding pressure points had a different result. To achieve my 95% threshold, the minimum number of principal components to use is 4, equal to the number of variables being studied. Therefore, PCA, in this case, was not needed. I still decided to use the PCA data for the remainder of the analysis (as opposed to the original variables), for consistency’s sake.

Through a biplot, we can also observe how the variables being studied relate to the principal components created. I also used a correlation matrix to achieve the same goal.

**CLUSTERING**

**SERVE**

**RETURN**

**PRESSURE POINTS**

**CLUSTERING VISUALIZATIONS AND CONCLUSIONS**

**VISUALIZATION #14**

**VISUALIZATION #15**

**VISUALIZATION #16**

**VISUALIZATION #17**

**VISUALIZATION #18**

**VISUALIZATION #19**

**VISUALIZATION #20**

**VISUALIZATION #21**

**LINEAR REGRESSION**

**SOURCES**

<https://www.atptour.com/en/rankings/singles>

<https://www.atptour.com/en/stats/leaderboard?boardType=serve&timeFrame=52Week&surface=all&versusRank=all&formerNo1=false>

<https://www.iban.com/country-codes>

<https://help.adjust.com/en/article/countries-by-region>