**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**

To help with the clustering process, I added a function wssplot() to the script, which creates the within sum of squares plot. As I’ll perform a k-means clustering analysis, the WSS plot will help decide on the optimal number of clusters to use.

OBS: Clustering results obtained are different every time you perform a new analysis. The results shown here are the results I obtained in my first analysis.

**SERVE**

For the serve components, the WSS plot shows us that the optimal number of clusters is 3. Using the kmeans() function, it is simple to create a new variable with the clustering results. By looking at this variable, it’s evident that the k-means algorithm does not do a great job, as the within clusters sum of squares by cluster obtained is 42.5%. This means that the clusters explains only 42.5% of the total variance in the data set comprised of the serve variables.

Finally, by looking at the summary results for each cluster, we can conclude that the cluster number 1 has average-level servers; the cluster number 2 has the worst servers; the cluster number 3 has the best servers. The division of the data points by cluster is pretty accurate. Although it does not match the ATP rankings/ratings, the division from cluster to cluster is evident. Therefore, this clustering divides the data fairly well.

**RETURN**

In the case of returns, the WSS plot tells us that the optimal number of clusters is 3. By looking at the variable created from the kmeans() function, we note that the within clusters sum of squares by cluster obtained is 62.2%, slightly better than our k-means performance for the serve variables. This means that the clusters explains only 62.2% of the total variance in the data set comprised of the return variables.

Finally, by looking at the summary results for each cluster, we can conclude that the cluster number 1 has average-level returners; the cluster number 2 has the worst returners; the cluster number 3 has the best returners. An interesting thing to note about the return clustering is that our results exactly match the rankings/ratings calculated by the ATP.

**PRESSURE POINTS**

Finally, for pressure points, the WSS plot tells us that the optimal number of clusters is 4. The within sum of squares by cluster obtained is 45.6%, so the clusters only serve to explain 45.6% of the total variance in the data set comprised of the pressure variables.

Contrary to the other areas of study, the pressure points clustering does not do a good job dividing the data. By looking at the medians and means of the ATP ratings for each cluster, I classified clusters 1 and 3 as being comprised of players that are generally good under pressure, while clusters 2 and 4 have players that are not so good under pressure. One interesting observation to note is that the clusters 1 and 2 are very similar in every variable besides the deciding set won measure. Cluster 1 players are easily the best in deciding sets won, while cluster 2 players are clearly the worst.

**CLUSTERING VISUALIZATIONS AND CONCLUSIONS**

To help interpret the results obtained from our clustering analysis, I added multiple binary variables to our original data set. These variables are:

* Returners, comprised of players on the best return cluster group, and on the worst serve cluster group. This group contains 14 players.
* Servers, comprised of players on the best serve cluster group, and on the worst return cluster group. This group contains 8 players.
* Worst\_allaround, comprised of players on the worst cluster for all of the studied measures (serve, returns, and pressure points). No players fall under this category.
* Worst\_allaround\_nop, comprised of players on the worst cluster for the serve and return variables. No players fall under this category.
* Best\_allaround, comprised of players on the best cluster for all of the studied measures. 8 players match this criteria.
* Best\_allaround\_nop, comprised of players on the best cluster for serves and returns. The 8 players in this group are the same as the ones in the Best\_allaround group.

**VISUALIZATIONS #14 & #15**

**Chart, bar chart

Description automatically generated**

By looking at this bar plot, we can observe the distribution of players by height (in black), and the distribution of returners by height (in yellow). It is evident that the percentage of players who fall in the group of “Returners” against the total number of players drops as height increases. Therefore, the taller a player is, the lesser the chances it is in the “Returners” group.Chart, bar chart

Description automatically generated

The conclusion established by looking at the previous visualizations become even more clear when creating a percent stacked bar plot. We can see that, as we have an increase in height, the proportion of returners drops.

**VISUALIZATIONS #16 & #17**

Chart, bar chart, histogram

Description automatically generated

As there are very few players who belong to the “Servers” group, it is hard to make any conclusions on this data. However, it is interesting to note that the 3 taller players in our analysis belong to this group.

Chart, bar chart, histogram

Description automatically generated

Although the numbers of players in this group is small, this visualization tells us that top players with a height over 2 meters are all similar in terms of serves and returns. They have elite serves, while having bad returns when compared with other top players.

**VISUALIZATIONS #18 & #19**

Chart, bar chart, histogram

Description automatically generated

Above, we see that, although the best all-around are not the tallest players in the circuit, they tend to be tall, with the ideal height for a tennis players being anywhere between 185 cm and 200 cm.

Chart, bar chart

Description automatically generated

Similarly, the proportion of the best all-around divided by height gives us some more information, telling us that, perhaps, players with a height between 190 cm and 200 cm have an edge. This makes sense, as short players tend to be worse when it comes to serving, while the tallest players have a disadvantage in returning position and movement (which can be exposed in returning games).

**VISUALIZATIONS #20 & #21**

Chart, bar chart, histogram

Description automatically generated

Chart, bar chart

Description automatically generated

Since no players fall under the criteria of being in the worst cluster group for every category, there is nothing we can conclude from the plots created from the Worst\_allaround group.

**LINEAR REGRESSION**

**Height VS Serve Rating – How does height affect a player’s serve?**

*Do the variables seem related?*

To answer this question, I created a scatter plot of the independent variable (height) against the dependent variable (serve rating):

Chart, scatter chart

Description automatically generated

Additionally, a correlation of 0.61 was computed between the variables. This shows us that the variables are positively related in a moderate strength.

*Residuals: do they comply with our regression assumptions?*

The visualization below gives us an idea of how the variability of the data changes in our linear regression model:

Chart, scatter chart

Description automatically generated

This plot shows us that there are some outliers in the middle range for the height variable of the data. I could subset the data to remove those results but decided to proceed with the original data for the analysis since the disparity is not enough to compromise the model.

*Regression results*

By running the built-in summary() function on our model, we can obtain its results. First, let’s look at the regression equation 🡪 Service rating = 43.30 + 1.19 \* Height (in cms). This means that an increase of 1 cm in a player’s height causes an expected increase of 1.19 in service rating.

An R2 of 0.3724 implies that our model explains 37.24% of the variability of the service rating variable. This means that our independent variable (height) explains 37.24% of the variability in service rating. Additionally, a low p-value (less than 0.0001) tells us that the independent variable is statistically significant in our model, that is, height is meaningful when predicting service rating. A low p-value (good) and weak R2 value means that, although we should keep the variable height in our model, it would benefit from the addition of more significant independent variables.

*Actual VS Predicted Values*

Chart, scatter chart

Description automatically generated

Although some values seem to deviate from the model’s prediction, the regression line seems to capture a trend in the data: the taller the player, the better his serve.

**Height VS Return Rating – How does height affect a player’s return?**

*Do the variables seem related?*

To answer this question, I created a scatter plot of the independent variable (height) against the dependent variable (return rating):

Chart, scatter chart

Description automatically generated

Additionally, a correlation of -0.38 was computed between the variables. This shows us that the variables are negatively related in a moderate/weak strength.

*Residuals: do they comply with our regression assumptions?*

The visualization below gives us an idea of how the variability of the data changes in our linear regression model:

Chart, scatter chart

Description automatically generated

Similarly to our previous model, this plot shows us that there are some outliers in the middle range for the height variable of the data. I could subset the data to remove those results but decided to proceed with the original data for the analysis since the disparity is not enough to compromise the model.

*Regression results*

By running the built-in summary() function on our model, we can obtain its results. First, let’s look at the regression equation 🡪 Return rating = 257.88 – 0.63 \* Height (in cms). This means that an increase of 1 cm in a player’s height causes an expected decrease of 0.63 in service rating.

An R2 of 0.1436 implies that our model explains 14.36% of the variability of the service rating variable. This means that our independent variable (height) explains 14.36% of the variability in service rating. Additionally, a low p-value (less than 0.0001) tells us that the independent variable is statistically significant in our model, that is, height is meaningful when predicting return rating. A low p-value (good) and very weak R2 value means that, although we should keep the variable height in our model, it needs the addition of more independent variables to improve its explanatory power, as height is not enough.

*Actual VS Predicted Values*

Chart, scatter chart

Description automatically generated

Although some values seem to deviate from the model’s prediction, the regression line seems to capture a trend in the data: the taller the player, the worse his return.

**First Serve % VS Service Games Won % - Does first serve percentage matter in terms of service games won?**

*Do the variables seem related?*

To answer this question, I created a scatter plot of the independent variable (first serve %) against the dependent variable (service games won %):

Chart, scatter chart

Description automatically generated

Additionally, a correlation of 0.03 was computed between the variables. These results are an early indication that the variables are very weakly related, therefore, we should build this model with caution, as it might not be meaningful.

*Regression results*

The R2 of our model is extremely low (0.0009), and the p-value of the independent variable used is too high (0.746), therefore, there is no point in using this model because these values tell us that first serve % does not matter when predicting the % of service games won. So our model is not meaningful and we should drop it.

**Age VS Pressure Rating – Does age help predict pressure rating?**

*Do the variables seem related?*

To answer this question, I created a scatter plot of the independent variable (age) against the dependent variable (pressure rating):

Chart, scatter chart

Description automatically generated

Additionally, a correlation of -0.11 was computed between the variables. These results are an early indication that the variables are very weakly related, therefore, we should build this model with caution, as it might not be meaningful.

*Regression results*

The R2 of our model is extremely low (0.01291), and the p-value of the independent variable used is too high (0.227), therefore, there is no point in using this model because these values tell us that age does not matter when predicting a player’s pressure rating. So our model is not meaningful and we should drop it.

**Height, ATP Ranking, Return Rating VS Serve Rating – How will adding more variables to our model affect how we predict the serve rating?**

*Regression results*

By running the built-in summary() function on our model, we can obtain its results. First, let’s look at the regression equation: 🡪 Service rating = 196.1628 + 0.7386 \* Height (in cms) - 0.1604 \* ATP ranking - 0.4172 \* Return rating.

This means that an increase of 1 cm in a player’s height causes an expected increase of 0.7386 in service rating; a fall of 1 position in the rankings causes an expected decrease of 0.1604 in service rating; and an increase of 1 in return rating causes a decrease of 0.4172 in service rating.

First of all, the p-value for all variables in the model is below 0.05, therefore, they are all meaningful. A new R2 of 0.5248 shows that the multi-regression model is more powerful than our previous one (the original R2) was of 0.3724. This means that, by adding more statistically significant independent variables resulted in an increase of around 15% in the model’s explanatory power.

An R2 of 0.1436 implies that our model explains 14.36% of the variability of the service rating variable. This means that our independent variable (height) explains 14.36% of the variability in service rating. Additionally, a low p-value (less than 0.0001) tells us that the independent variable is statistically significant in our model, that is, height is meaningful when predicting return rating. A low p-value (good) and very weak R2 value means that, although we should keep the variable height in our model, it needs the addition of more independent variables to improve its explanatory power, as height is not enough.

*Conclusions*

* Taller players seem to have a better service rating.
* On the other hand, the higher (worst) a player's ranking is, the worse its serve stats seem to be.
* And the better the player's return rating, the worse its service rating.

**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>