**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**

**VISUALIZATION #2**

**VISUALIZATION #3**

**VISUALIZATION #4**

**VISUALIZATION #5**

**VISUALIZATION #6**

**VISUALIZATION #7**

**VISUALIZATION #8**

**VISUALIZATION #9**

**VISUALIZATION #10**

**VISUALIZATION #11**

**VISUALIZATION #12**

**VISUALIZATION #13**

**PRINCIPAL COMPONENT ANALYSIS**

**SERVE**

**RETURN**

**PRESSURE POINTS**

**CLUSTERING**

**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>