Tristan Savella
Data Immersion: Task 6.1 (from CareerFoundry)
Project Overview: ATP Data (Men's Tennis)

Description: This document provides an overview of my workflow for this project, as well as a description of the original dataset and wrangling/cleaning steps. In addition to this document, two other documents will be provided for this project: a storyboard hosted on Tableau and an outline of the storyboard.

Table of Contents

Part 1: Data Source

Data Source and Content

This dataset contains information about the Association of Tennis Professionals, or the ATP, and contains three CSV Files (listed below). The following description, referring to the "Matches" CSV file, was written under "About Dataset" in the Kaggle link below:

> "The data set contains the details about all the ATP matches played since 1968. The match statistics are available for matches since 1991."

This dataset is updated annually and also cites the following two credits:

- <http://www.tennisabstract.com>
- Jeff Sackman

Link to Data
Downloaded From: <https://www.kaggle.com/datasets/sijovm/atpdata/data>

CSV Files:

1) Matches (Until 2022)
2) Rankings (Until 2022)
3) Players (Until 2022)

Data Limitations

The match statistics provided in this dataset focus only on serve statistics; there are no other match statistics, such as number of winners, unforced errors, points played at net, etc.

> There are many matches with missing statistics, including: all matches prior to 1991, those played at smaller tournaments, and more.

Part 2: Premise of the Project & Potential Questions to Explore

Motivations: Why this project? What problem needs to be solved?
For this project, there are two motivational factors that I would like to explore.

1) Uncover any correlation between serve statistics and surface?
2) Follow the careers of the "Big 3" (Federer, Djokovic and Nadal);
    a) Do any serve or break point statistics jump out at their matches when compared to each other, or other tennis players?
    b) Did their serve statistics change over time?
    c) Did their serve statistics vary by surface?

Potential Objectives: What questions will I answer to generate solutions for the project?

1) Which serve statistic correlates highest to number of wins?
    a) Does this vary by surface? Best of 3 vs. 5?
2) Is there anything in the serve statistics that helps explain the high success of the Big 3?
    a) Did certain statistics improve/get worse during their stronger or weaker years?
    b) How do their stats compare to other players?
    c) Did their statistics vary by surface or tournament level?
3) Is there a correlation between a player's height and their serve statistics?
    a) Does this vary by surface?

Potential Scope: What will this project include and exclude?

1) Only include matches after 2000 (match statistics only available after 1991)
2) For following the careers of the Big 3:
    a) It may be helpful to only consider data from matches after the first of them turned Pro (Roger Federer)
    b) Might also be helpful to only consider data from tournaments in which one of them participated
    c) Observe the differences of serve statistics for each of the three surfaces: Grass, Clay and Hard

Part 3: Data Profile & Jupyter Notebooks (Scripts)

Data Preparation: Overview of Cleaning and Wrangling Steps

This section outlines the steps I have taken for cleaning, wrangling and preparing my data set. I only used the "Matches" CSV. The cleaning and wrangling steps were divided into two notebooks. I am focusing my analysis on matches played after the year 2000.

Overview of Each Notebook

Notebook 1a: "ATP Initial Exploration Part 1"

This is the first of scripts conducting an initial exploration of the ATP Dataset. This notebook focuses on exploring the raw data and steps to cleaning/wrangling (the second notebook will take the prepared data and create subsets for further analysis). In this script, the original "matches.csv" was cleaned and renamed, "df_post2000". The cleaning steps are outlined as follows:

- Renaming Columns
- Deriving New Variables
- Removing all matches prior to 2000
- Removing all data with missing/faulty match statistics*
- Changing Variables of Certain Data Types
- Exported "df_post2000" as a PKL file into "Prepared Data" Folder

* Some match statistics were not "missing" but were very likely faulty, or the result of incomplete matches (i.e. player retirement). For example:

Some data suggests that the winner of certain matches did not miss a single serve in an entire match, which is highly unlikely. In all of these cases, the loser also played 8 or fewer second serves, which is also highly unlikely. It is more likely that the match was incomplete or not properly collected. In such cases, these entries were removed.

Notebook 1b: "ATP Initial Exploration Part 2"

While Notebook 1a focused on cleaning and wrangling steps, Notebook 2b created subsets to use for analysis.

Using the prepared "df_post2000" from the first notebook, this notebook created a new dataframe (called "df_matchstats") and 12 subsets. The steps were followed:

1) Create new main data frame consisting of only match statistics: ‘df_matchstats’
2) Divide df_matchstats into 2 categories; Best of 3 or 5 Matches:
    a) ‘df_matchstats_BO3’
    b) ‘df_matchstats_BO5’
3) For the 2 categories, create the following subsets
    a) One for each surface: Clay, Grass, Hard (‘df_clay’, ‘df_grass’ and ‘df_hard’)
    b) Matches involving the Big 3 (‘df_big3_win’ and ‘df_big3_lose’)
4) For the last two categories (surface and Big 3), further subsets were created dividing each dataframe into matches played Best of 3 or Best of 5. These subsets were not used for the remainder of my project.
    a) e.g. df_clay_BO3, df_clay_BO5 etc.)

Notebook 2: “6.2 Exploring Relationships”

In this notebook, the same process is applied to the main dataframe, 'df_matchstats' and three of the subsets created in Notebook 2, based on court surface. The goal was to see if the court surface had any affect on any of the variables or the correlations between them. Using the prepared ‘df_matchstats’ from Notebook 2, the data was prepared (namely, irrelevant columns were removed) in order to make various correlation charts to explore the relationships between different variables. The following charts were created:

1) Correlation heatmap
2) Scatterplot
3) Pair Plot
4) Categorical Plot

Using the Correlation Heatmap: Through these maps, I focused my attention on the three aforementioned serve statistics using the correlation heatmap. Of the three statistics, winner’s percentage of points won on 1st serve (‘w_%1stWon’) and percentage of points won on 2nd serve (‘w_%2ndWon’) had strong negative correlations with the number of break points the winner faced (‘w_#bpFaced’), with the former being stronger than the latter (0.54 to 0.44). There was no relationship between a winner’s percentage of first serve points won (w_%1stIn) and the number of break points that they faced.

Using the scatterplot: One can visualize the negative correlation between a player’s (namely, the winning player) percentage of first serve points won and number of break points they faced.

Using the Pairplot: No important insights were gleaned from this plot

Using the Categorical Plot: Categorical plots were used to display the distribution of winning and losing players' percentage of first serve points won and 2nd serve points won ('w_%1stWon', 'w_%2ndWon', 'l_%1stWon' and 'l_%2ndWon'). Categorical plots were also used to display winning player's heights and percentage of first serves in play; no insights were gleaned for those two variables

The process was repeated for the following subsets:

1) df_grass
2) df_clay
3) df_hard
4) df_BO5
5) df_BO3

Insights

1) Across all surfaces, both the winner and loser won about 20% more points on their first serves than on their 2nd serve (76.3% to 56.2% for winners and 65.4% to 44.9% for losers).
2) Match winners on average won 10% more points on both their first and second serves than losers
3) Regarding court surface:
    a) Both winning and losing players had the highest averages of points won on first serve on grass courts (78.7% and 68.5%), followed by hard courts (77.2% and 66.5%), and the lowest averages on clay courts (73.6% and 62.3%)
    b) Both winning and losing players had the highest averages of points won on second serve on grass courts (56.9% and 46.3%), followed by hard courts (56.4% and 4.15%,) and the lowest averages on clay courts (55.8% and 43.9%).
    c) Court surface had a stronger impact on points won on 1st serve rather than 2nd serve.
4) There was no significant difference in the stats between matches played Best of 3 sets and matches played Best of 5 sets.

Notebook 3a: "Logistic Regression"

Using logistic regression, the three serve statistics are tested to see their correlation with the likelihood of winning a match on the ATP tour. Using 'df_matchstats', statistics for the winner and loser were created and a new variable was added (1 for a match won and 0 for a match lost). The two dataframes were then combined and are ready for the regression.

The following insights were gleaned:

1) % of points won on first serve: 0.13 correlation with likelihood of winning a match
2) % of points won on second serve: 0.10 correlation with likelihood of winning a match
3) % of first serves in play: 0.04 correlation with likelihood of winning a match

Notebook 3b: "Logistic Regression"

The steps from script 3a are repeated here, using the additional variable of "court surface". The goal is to see if the correlation between each serve statistic and likelihood of winning a match varied by court surface.

Notebook 4: "Big 3 Serve Stats at Their Favorite Slams"

Between the Australian Open 2000 and Wimbledon 2022, 90 Grand Slam Tournaments have taken place (Wimbledon was canceled in 2020 due to the Covid Pandemic). A member of the "Big 3" (Nadal, Federer and Djokovic) has won 63 of those 90 Grand Slam Titles. This notebook explores the average percentage of points won on 1st serve by the Big 3 at three Grand Slam Tournaments: Australian Open, Roland Garros and Wimbledon. Each one of the Big 3 has the record number of titles at one of these three tournaments. Using 'df_matchstats', this notebook will explore their serve statistics at each of these tournaments and later on, will be compared to the average match winner's serve statistics on the same surface.

This notebook displays and visualizes (through line and bar charts) the average percentage of points won on first serve by each of the Big 3 at each tournament. The first grand slam included in my data set is the Australian Open in 2000 and the last is Wimbledon 2022 (Novak Djokovic is the only member of the Big 3 who won additional Grand Slam Tournaments after Wimbledon 2022).

These three were taken as each of the Big 3 holds the record for most number of titles at one of these tournaments. As of Wimbledon 2022 (the last recorded grand slam tournament on this dataset), here is where the big 3 stands at the three slams
:

1) Australian Open:
    a) Novak Djokovic: 9 Titles
    b) Roger Federer: 6 Titles
    c) Rafael Nadal: 2 Titles
2) Roland Garros
    a) Rafael Nadal: 14 Titles
    b) Novak Djokovic: 2 Titles
    c) Roger Federer: 1 Title

3) Wimbledon
    a) Roger Federer: 8 Titles
    b) Novak Djokovic 7 Titles
    c) Rafael Nadal: 2 Titles

Part 4: Data Dictionaries

List of Subsets/Wrangled Dataframes Used and Created for this Project

The following subsets were created using 'df_matchstats', based on the following parameters:

1) **Prepared Data for Visualization Analysis (Notebook 2):** Removed all non-numerical and NaN values for correlation heatmap, categorical plot, pair plot and scatter plot:
    a) 'df_matchstats2'
2) **Subsets: whether a match was played best of 3 or 5 sets (Notebook 1b):**
    a) 'df_BO3'
    b) 'df_BO5'
3) **Subsets: one for each of the three main court surfaces, and further divided into matches played best of 3 or best of 5 sets (Notebook 1b):**
    a) df_grass; df_grass_BO3; df_grass_BO5
    b) df_clay; df_clay_BO3; df_clay_BO5
    c) df_hard; df_hard_BO3; df_hard_BO5
4) **Wrangled dataframe: Created to run a logistic regression (Notebook 3a.)**
    a) 'combined_df'
5) **Wrangled dataframe: Created to run a logistic regression and includes court surface (Notebook 3b.)**
    a) 'coefficient_df'
6) **Wrangled dataframe: the average first serve points won percentage by each of the Big 3 at one of the three grand slam tournaments, minus the US Open (Notebook 4).**
    a) 'result_df'

Main dataframe: 'df_matchstats'

	Type	Calculated	Description
tourney_id			Tournament's unique ID number
Year			The year the tournament took place
tourney_name			Name of tournament
surface			Court Surface (e.g. hard, clay, grass)
tourney_level			Level of the tournament
winner_id			Player id number
winner_name			Name
winner_age			Player's age

winner_rank			Winner’s ranking at time of match
winner_ht			Winner’s height
loser_id			Player id number
loser_name			Name
loser_age			Player’s age
loser_rank			Loser’s ranking at time of match
loser_ht			Loser’s height
best_of			Best of how many sets (3 or 5)
round			The round of the tournament (1st round, QF, etc.)
minutes			Match duration in minutes
w_#ServeGames			Number of service games played
w_#aces			Number of aces
w_#dfs			Number of double faults
w_#ServePoints			Number of points played on serve
w_#1stServesIn			Number of first serves in
w_#2ndServePoints			Number of points played on second serve
w_%1stServesIn			Percentage of first serves in play
w_#1stWon			Number of points played on first serve won
w_%1stWon			Percentage of points played on first serve won
w_#2ndWon			Number of points played on second serve won
w_%2ndWon			Percentage of points played on second serve won
l_#ServeGames			Number of service games played
l_#aces			Number of aces
l_#dfs			Number of double faults
l_#ServePoints			Number of points played on serve
l_#1stServesIn			Number of first serves in
l_#2ndServePoints			Number of points played on second serve
l_%1stServesIn			Percentage of first serves in play
l_#1stWon			Number of points played on first serve won
l_%1stWon			Percentage of points played on first serve won
l_#2ndWon			Number of points played on second serve won
l_%2ndWon			Percentage of points played on second serve won

‘combined_df’

	Type	Calculated	Description
1stServesIn			
1stWon			
2ndWon			
Win			1 = Match Won; 0 = Match Lost

‘coefficient_df’

	Type	Calculated	Description
Feature			One of the three serve statistics
Coefficient			Correlation between serve stat and match win
Surface			Court Surface

‘result_df’

	Type	Calculated	Description
Year			The year the tournament took place
tourney_name			Name of tournament
%1stWon			Percentage of First Points Won by Specific Player
Player			Nadal, Djokovic or Federer

Appendix: Data Dictionary for Original Datasets (Prior to Cleaning/Wrangling)

CSV 1: ATP Matches

	Type	Calculated	Description
tourney_id			Tournament's unique ID number
tourney_name			Name of Tournament
surface			Court Surface (e.g. hard, clay, grass)
draw_size			Size of the draw
tourney_level			Level of the tournament
tourney_date			Date of the tournament
match_num			Match number in a certain tournament
winner_id			Player id number
winner_seed			Player's tournament seeding
winner_entry			How did the player enter the tournament?
winner_name			Name
winner_hand			Right or Left-Handed
winner_ht			Player's height
winner_ioc			Player's country of origin
winner_age			Player's age
loser_id			Player id number
loser_seed			Player's tournament seeding
loser_entry			How did the player enter the tournament?
loser_name			Name
loser_hand			Right or Left-Handed
loser_ht			Player's height
loser_ioc			Player's country of origin
loser_age			Player's age
score			Match Score
best_of			Best of how many sets (3 or 5)
round			The round of the tournament (1st round, QF, etc.)
minutes			Match duration in minutes

w_ace			Number of aces
w_df			Number of double faults
w_svpt			Number of points played on serve
w_1stin			Number of first serves in
w_1stWon			Number of first serve points won
w_2ndWon			Number of second serve points won
w_SvGms			Number of service games played
w_bpSaved			Number of break points saved
w_bpFaced			Number of break points faced
l_ace			Number of aces
l_df			Number of double faults
l_svpt			Number of points played on serve
l_1stin			Number of first serves in
l_1stWon			Number of first serve points won
l_2ndWon			Number of second serve points won
l_SvGms			Number of service games played
l_bpSaved			Number of break points saved
l_bpFaced			Number of break points faced
winner_rank			Winner’s ranking at time of match
winner_rank_points			Winner’s number of ranking points at time of match
loser_rank			Loser’s ranking at time of match
loser_rank_points			Loser’s ranking at time of match

CSV 2: Rankings

	Type	Calculated	Description
ranking_date			Week and year of ranking (rankings are updated weekly)
rank			Player ranking
player			Player ID number
points			A player’s number of ranking points

CSV 3: Players

	Type	Calculated	Description
player_id			Each player’s unique Player ID Number
name_first			
name_last			
hand			Whether a player is right or left-handed
dob			Date of Birth
ioc			Country of Origin/Represented
height			A player’s height in centimeters
wikidata_id			