6.1 Sourcing Open Data
Christoph Schöneseiffen

Explanation:

I have chosen the below data set because I'm a big tennis enthusiast and tennis was probably my first contact with statistics when I was a little child. Even today in my tennis club we discuss a lot of matches. Thus working with this data set to gain insight and improve my Python skills at the same time was a perfect match.

Data Profile

1. Tennis_atp
   Link: JeffSackmann/tennis_atp: ATP Tennis Rankings, Results, and Stats (github.com)

Data Source & Collection:

The data is sourced from a github profle by Jeff Sackmann, who collects tennis data.

For purpose of this project the data sets used have been limited to the years 2002 – present (20-06-2022) and ATP singles matches only (Meaning, neither matches below ATP level nor doubles).

The match data itself is available on the internet, but he has enriched it gathering classical tennis statistics regarding service games and match duration. Additionally, he has enriched the data with personal players data that is publicly available (date-of-birth, height, playing hand). This data is collected manually and can be submitted by other users if incomplete data is found. For this he requires a certain proof/documentation (e.g. ATP-profile, Wikipedia entry).

Overall, the data appears reliable.

Content:

The data includes 49 columns and 60.699 rows.

Each row provides data for an individual match. This includes, tournament, players, personal information on players and match statistics.

A data dictionary has been provided by Jeff Sackmann to check meaning of columns (although most are self-explainatory)

Limitation to the data set:

- Not all data is available for all players e.g. height, hand played with, date of birth
- It cannot be verified that all data is correct
- Since data has been limited starting 2002 no insights can be generated for matched player in prior years
- Since only ATP-level matches are analyzed, lower-level matches are left out although players might have played matches on several levels (especially in young years or lower ranked players), therefore, the results might not give a complete picture for individual players

Ethics:

- The data is taken from another GitHub account and must be acknowledged
- The data is publicly available (including players data) and no restrictions apply

Data Cleaning and Descriptive Analysis:

6.1 Sourcing Open Data
Christoph Schöneseiffen

Please check the python code provided.

Some minor adjustments have been made to the data set. Overall, no rows have been deleted because they still contain match data although match statistics might be incomplete or not available.

After initial cleaning the data still has 49 columns and 60.699 rows.

Research-Questions:

- Which player has won the most matches since 2002?
- Which player has won the most titles since 2002?
- Which player has the best ace to number of service points played percentage?
- Does a greater height contribute to a high number of aces?
- Which factors contribute to winning a match (e.g. age, ranking, first serve percentage, prior tournament results…)
- Players from which countries/how many different countries have won tournaments?

NOTE:

The following data set has been downloaded and is considered to be used at a later stage, for now it has been disregarded and might be added to the analysis later on.

2. Tennis-data.co.uk
   Link: http://www.tennis-data.co.uk/alldata.php

Data Source & Collection:

The website uses other available data to compile the datasets. This links to official results pages as well as bookmaking websites. Hence the data is probably gathered by web scraping.

Overall the data seems reliable.

Content:

The Site provides Match Data for ATP Torunaments starting 2000 including betting odds from bookmakers.

The structure of the data is very similar to the first data set, however, it provides several columns that are not included in the other data set.

This includes:

- the results are structured in a different way (each set score per player in a different column)
- Tournament Name and Location are presented in different columns
- A Comment column that indicates if a match was completed or ended by retirement or walkover
- The number of sets won per player in a match

Limitation to the data set:

- Accuracy cannot be tested especially for betting odds (however, those will most likely not be used in the project unless deemed interesting or necessary later on)