LJ04062018

Project Proposal – Programming Project, June 2018

Problem statement

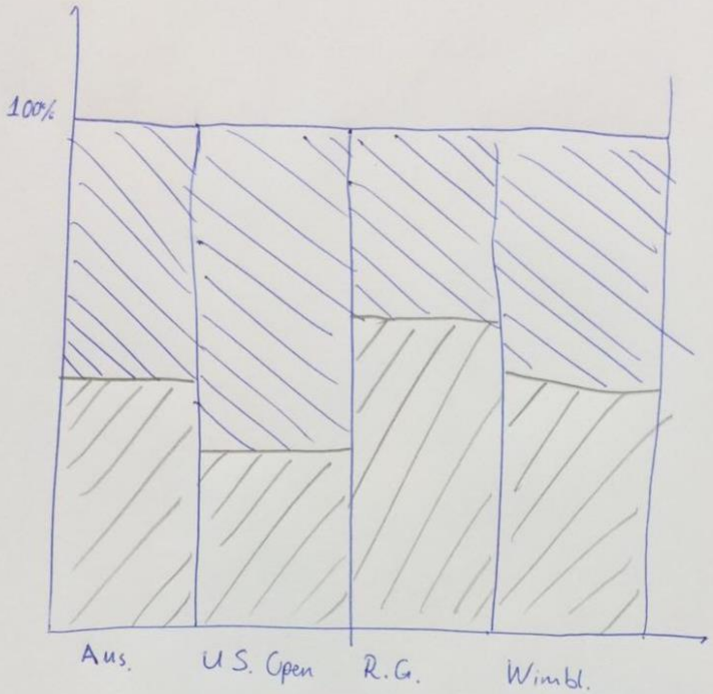
Answer:
The Women's Tennis Association (WTA) oversees a large number of yearly tournaments between an even greater number of tennis players, resulting in a sea of data and statistics regarding matches, rankings, player statistics (e.g. height, handedness, nationality, etc.). This abundance of information makes for difficult interpretation of statistics and detection of patterns.

Solution

My proposal for the final project constitutes a number of linked (interactive) visualizations on player- and match-statistics in the WTA. With these visualizations, I aim to clarify patterns and relationships found within the data (e.g. which country produces the most Grand Slam winners? Does left handedness increase one's probability of winning a match? Etc.)

** I plan to focus on Grand Slam tournaments, exclusively, in order to avoid intolerable computation times.

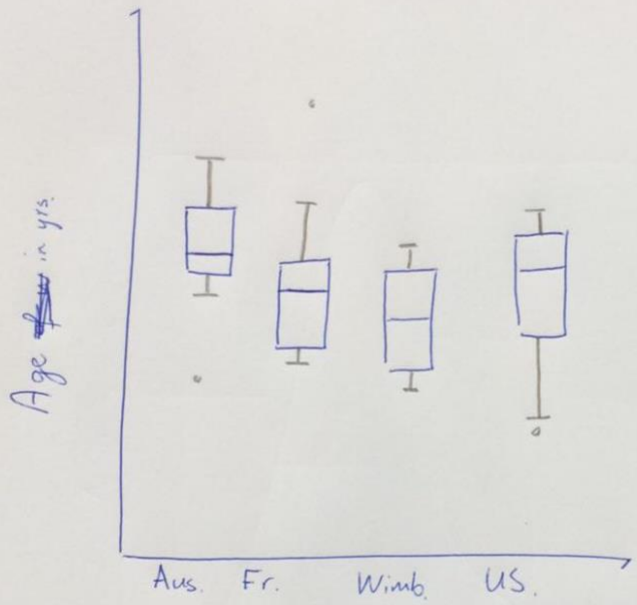
<u>Sketches:</u>



Percentage of left-or right-handed match winners of every grand slam.

▨ : right-handed

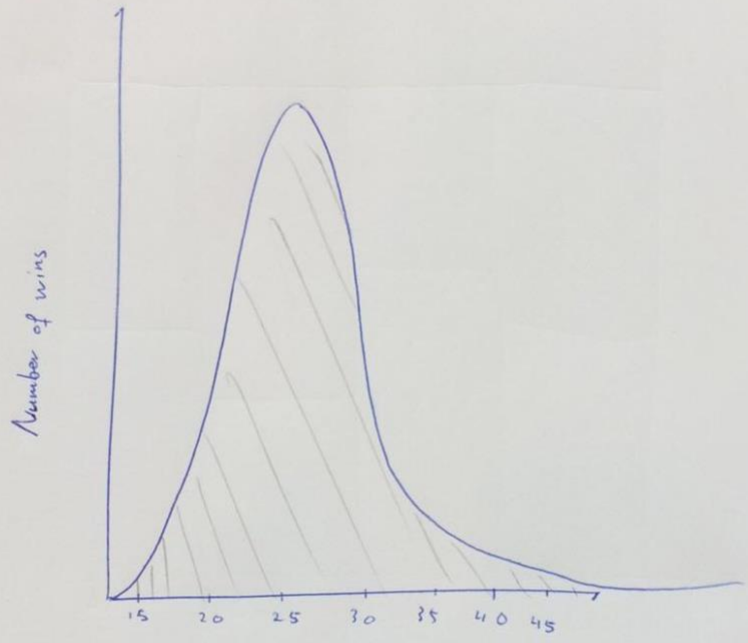
▨ : left-handed



Grandslam

Boxplots of ages of Grandslam winners.

- handed
handed



Number of grandslam wins based on age of player

Prerequisites

[list the data sources and required modifications (include links); list the external components (e.g. libraries); include a review of similar visualizations; identify the hardest parts]

Link to data source (Kaggle): https://www.kaggle.com/joaoevangelista/wta-matches-and-rankings

External components: D3.js, d3-tip

Similar visualizations:

- **https://www.kaggle.com/residentmario/exploring-wta-players**
- **https://www.kaggle.com/anuj8june/wta-matches-and-rankings-comprehensive-analysis**
- **https://www.kaggle.com/pcbaradhwaj/wta-data-understanding-and-basic-plots-notebook**

Hardest part: I expect the majority of technical difficulty to arise from data preprocessing and manipulation for a couple for reasons. Firstly, I'm planning on making rather simple and straight forward charts (i.e., bar charts, scatter plots, line graphs) as the complexity of the data does not call for more complex visualizations. Secondly, the size of the four csv-files is rather large, and they are very unbalanced. So I suspect it will take me a good amount of time to prepare the data for visualization and analysis.