**Exploring ATP Tennis Match Statistics**

**Description**

The use of data in professional sports has increased substantially over the past couple of decades, and this includes tennis which has been increasingly following this trend, with data being used extensively by players, coaches, tournament organisers, broadcasters, betting companies and spectators. This project involves exploring match data from professional tennis matches in the ATP World Tour over one or more years, using publically available statistics from across all tournament matches to determine how some of the key match statistics relate to match results. Analysis of these match statistics may be used to determine the factors most strongly associated with winning matches, identify characteristics of "upset" matches (where the favourite or highly ranked player loses to someone significantly lower rated), or compare aspects of the tennis tour across different periods.

**The Data**

The key data for this project can be found at the following link:

<https://github.com/JeffSackmann/tennis_atp>

The files “atp\_matches\_xxxx.csv” contain match statistics for all ATP Tour level men’s singles matches for that year. There are also files containing match statistics for men’s doubles, qualifying matches and matches from challengers and futures (lower tier tournaments than the main ATP tour).

This will provide the main data for the project. The suggestion is to begin with one season of data (e.g. atp\_matches\_2019) and explore the data, then consider additional seasons and / or some of the additional data for follow-up after identifying directions for further analysis.

In addition to this match level data, there may be other sources of data desired for further analysis depending on what is pursued. For example, from the same source there are Match Charting data which give detailed point by point data for selected matches:

<http://www.tennisabstract.com/charting/>

<https://github.com/JeffSackmann/tennis_MatchChartingProject>

These could be used to examine selected matches in more detail or to try to gather additional data to help with analysis.

**Objectives**

The project is internal, and within the scope of analysing the ATP tennis match data, there is a lot of room for the student to define the specific questions and approaches that they wish to consider in the project.

A starting objective will be to take one season of match data – e.g. the most recent full ATP season – and try to provide an overview of this season, including key statistics and visualisation of the data. Doing this should help come to understand what content is in the data, and potentially identify some interesting aspects that could be explored further. This may mean going back and analysing data from previous seasons or exploring other data (such as the Match Charting data) further.

Possibilities to consider investigating further include:

* Investigating “upset” matches: using the match data to try to identify key features or matches that are considered “upsets”, i.e. where a significantly less favoured opponent beats a significantly more favoured opponent. This would include defining exactly what counts as an “upset” and analysing to see if there are significant differences in any of the match statistics between upset matches and non-upset matches.
* Comparing match statistics across different periods: using match data from multiple seasons to compare the statistics from different periods or eras in the game.
* Match prediction: identifying the statistics most strongly associated with winning a match and using these to create a model for predicting the outcome of matches, using historical data of the players involved and e.g. logistic regression to determine the probability of the player winning the match.

**Useful Resources**

Below are a list of articles and resources to begin the literature review and generate ideas for exploring the data.

* Exploratory Data Analysis with R: <https://medium.com/@tobikasali/exploratory-data-analysis-with-r-f0b0a5163ecd>
* Understanding the Importance of First Serve in Tennis with Data Analysis: <https://towardsdatascience.com/understanding-the-importance-of-first-serve-in-tennis-with-data-analysis-4829ab088d36>
* Predicting the Outcome of a Tennis Tournament: Based on Both Data and Judgments: <https://www.researchgate.net/publication/330883807_Predicting_the_Outcome_of_a_Tennis_Tournament_Based_on_Both_Data_and_Judgments>
* Machine Learning for the Prediction of Professional Tennis Matches: <https://www.doc.ic.ac.uk/teaching/distinguished-projects/2015/m.sipko.pdf>
* Visualizing Professional Tennis Upsets: ATP 2012-2014 Men's Singles Matches: <https://nycdatascience.com/blog/student-works/visualizing-professional-tennis-upsets-atp-2012-2014-matches/>
* Exploratory Analysis of 2012-2017 ATP World Tour Dataset: <https://rpubs.com/patenish/atp>
* Predicting Tennis Matches Using Machine Learning: <https://libstore.ugent.be/fulltxt/RUG01/002/945/727/RUG01-002945727_2021_0001_AC.pdf>
* Data-driven analysis of point-by-point performance for male tennis player in Grand Slams: <https://revistas.rcaap.pt/motricidade/article/view/16370>
* The analysis and forecasting of tennis matches by using a high dimensional dynamic model: <https://rss.onlinelibrary.wiley.com/doi/full/10.1111/rssa.12464>
* The role of analytics in tennis is on a long, slow rise: <https://www.tennis.com/news/articles/the-role-of-analytics-in-tennis-is-on-a-long-slow-rise>
* The logit function and its applications in sports modelling <https://www.sportstradingnetwork.com/article/logit-function-applications-sports-modelling/>
* Logistic Regression: <https://christophm.github.io/interpretable-ml-book/logistic.html>
* The history of analytics and statistics in tennis: <https://medium.com/the-sports-scientist/the-history-of-analytics-and-statistics-in-tennis-e19aa206fdf0>
* Advantage, Analytics: How Tennis Players Are Using Science to Elevate Their Artform: <https://sporttechie.com/tennis-analytics-us-open-rafael-nadal/>