


The dataset problem domain: Analysis of Professional Tennis Match Results

Dataset source: <http://tennis-data.co.uk/alldata.php>

You will need to download this dataset from home since Massey's filter restricts access to this website due to its categorization as a gambling site. If you're unable to download it from home, please let me know.

**Tennis-Data.co.uk**
Results | Odds | Links

Network Sites

Data Updated: 26th February 2023

BeGambleAware

Follow @12Xpert

Home	Livescore	Odds	Free Bets	Casino	Poker	Football	Articles	Contact
------	-----------	------	-----------	--------	-------	----------	----------	---------

NEW CUSTOMERS
[£50 Free Bets](#)
[£40 Free Bets](#)
[£30 Free Bets](#)
[£25 Free Bets](#)

ATP Tennis Tournaments

FAVOURITE BOOKIES
[bet365](#)
[William Hill](#)
[Betfred](#)
[Betway](#)

Tennis Organisations

Tennis Betting Sites

























Tennis Results Sites

Tennis Portals

Tennis News Sites

General Betting Sites

Data Files: All Competitions
[Notes.txt](#)
(text file key to the data files and data source acknowledgements)
ATP Men's Tour
A complete Excel file (zipped) for each ATP season is available. Individual CSV files for each competition are available through the links in the data menu to the right.

 2023 (Match results and betting odds)
 2022 (Match results and betting odds)
 2021 (Match results and betting odds)
 2020 (Match results and betting odds)
 2019 (Match results and betting odds)
 2018 (Match results and betting odds)
 2017 (Match results and betting odds)
 2016 (Match results and betting odds)
 2015 (Match results and betting odds)
 2014 (Match results and betting odds)
 2013 (Match results and betting odds)
 2012 (Match results and betting odds)
 2011 (Match results and betting odds)
 2010 (Match results and betting odds)
 2009 (Match results and betting odds)
 2008 (Match results and betting odds)
 2007 (Match results and betting odds)
 2006 (Match results and betting odds)
 2005 (Match results and betting odds)
 2004 (Match results and betting odds)
 2003 (Match results and betting odds)
 2002 (Match results and betting odds)
 2001 (Match results and betting odds)
 2000 (Match results)

SITE RESOURCES
[Odds, Results Data](#)
[Live Tennis Scores](#)
[ATP 2023 Calendar](#)
[WTA 2023 Calendar](#)

ODDS & RESULTS
[All Tournaments](#)
[>Grand Slams](#)
[Australian Open](#)
[French Open](#)
[US Open](#)
[Wimbledon](#)
[>Masters Series](#)
[Masters Cup](#)
[Cincinnati](#)
[Hamburg](#)
[Indian Wells](#)
[Madrid](#)
[Miami](#)
[Monte Carlo](#)
[Montreal](#)
[Paris](#)
[Rome](#)
[Shanghai](#)
[Stuttgart](#)
[Toronto](#)

Task 1: Wrangling, reshaping, EDA (25 marks)

- Collect data covering 10 years (2013 - 2022) from the above website. Read each excel dataset using Python and combine into a single dataset.
- Check that all the data has been read. Check that all the data in the combined dataset is in order based on the 'Date'.

- What other data-checking operations could you perform to make sure that the data is ready for analysis? Use various approaches to perform sanity checking on the data, including some plotting.
- Create a new column that is based on the 'ATP' column. The 'ATP' column is an integer value which represents a sequence/ordering of all the tournaments in a given year. Call the new column 'OverallSequence' which represents the order of all the tournaments from 2013 to 2022.
- Create EDA 10 visualisations of the dataset and explain each one. Be curious and creative. Ensure that the plots are clean and interpreted.

Task 2: Analysis questions and plotting (25 marks)

- Who are the top 10 players by total wins in the dataset, and how many wins do they have? Plot this.
- Which players have won the most Grand Slam matches and titles in the last 10 years? Plot this.
- Who are the top 10 players according to the largest number of First Round tournament losses across all 10 years? Plot this.
- Identify the 5 biggest upsets for each year in the dataset based on ranking differentials. List player names, rankings, winner/loser, score, and tournament name and what the difference in the rankings was at the time.
- Who were the top 10 players at year-end in 2017? How have their rankings changed over the period of 2013 to 2022? Plot this.

Task 3: Advanced analysis questions (25 marks)

- Which tournaments have had on average the most upsets (where a lower-ranked player defeated a higher-ranked player)? List the top 10 and plot their averages.
- Determine who the top 10 ranked players were at the end of 2022. Then calculate their head-to-head win-loss record against each other for all the matches they played in 2022. Present this result.
- List the top 5 players who had the longest winning streaks between 2013 – 2022. List their names, the lengths of their winning streaks and the year(s) in which they occurred.
- In tennis, each set is played first to 6, but sometimes it is played to 7. A tiebreak is a set that someone wins 7-6 and is different to someone winning a set 7-5. Tiebreaks are stressful and some players perform better than others in tiebreaks. Count how many tiebreaks each player in the entire dataset has played. Then, calculate the percentage of tiebreaks that each player has won. List the top 10 players according to the percentage of tiebreaks won.
-

Task 4: Open questions and analyses (25 marks)

- Come up with 3 more questions of your own.
- Try to demonstrate the usage of more advanced data wrangling functionalities as you answer the questions like group by, pivots etc...
- Create several plots.

Hand-in: Submit a single zipped file via the Stream assignment submission link. It should contain one notebook with all the answers embedded, and an HTML version of your notebook also with its output showing in case we have issues running your code.