Data Narrative 3

Analyzing data of the four championships of the 2013 Grand Slam

Akshat Barnwal
Civil Engineering
Indian Institue of Technology
Gandhinagar, India
akshat.barnwal@iitgn.ac.in

Abstract*—This is a report on the following multivariate dataset provided by the UCI Machine Learning Repository: *https://archive-beta.ics.uci.edu/dataset/300/tennis+major+tournament+match+statistics
which is a collection of 8 files containing the match statistics for both women and men at the four major tennis tournaments of the year 2013.

Keywords—Python, Pandas, Seaborn, libraries, CSV, KDE, Heatmap, Correlation, Net Point, Ace, Double Fault, Unforced Error, Break Point.

I. Overview of Dataset

Each file consists of 42 attributes such as the Player Name, Aces, Result and Round of the Tournament. The following files are present in the dataset:

A. *CSV files*

The following two CSV files are present in the repository and they are the primary source of where we obtain our data from:

- **AusOpen-men-2013.csv** – data on the Australian Open Tournament for men for the year 2013.
- **AusOpen-women-2013.csv** – data on the Australian Open Tournament for women for the year 2013.
- **FrenchOpen-men-2013.csv** – data on the French Open Tournament for men for the year 2013.
- **FrenchOpen-women-2013.csv** – data on the French Open Tournament for women for the year 2013.
- **USOpen-men-2013.csv** – data on the US Open Tournament for men for the year 2013.
- **USOpen-women-2013.csv** – data on the US Open Tournament for women for the year 2013.
- **WimbledonOpen-men-2013.csv** – data on the Wimbledon Open Tournament for men for the year 2013.
- **WimbledonOpen-women-2013.csv** – data on the Wimbledon Open Tournament for women for the year 2013.

B. *Attribute Information*

Though not present in the data file explicitly, this section is available in the source website of the data and holds crucial information about each of the attribute/variable present in the .csv files. As all the attributes in the files are in short-hand notation, without this information, it would have been quite difficult to decipher the meaning of said attributes.

II. Scientific Questions/Hypotheses

A. ***Question A:*** *What is the average number of points scored in a tennis game?*

We are starting off with a rather basic question that can help the uninitiated understand the game. Keeping track of the score is the most important thing in a game and for someone who is not familiar with tennis, understanding the range and patter in which the score varies can help get a good initial understanding of the game.

B. ***Question B:*** *What is the density-distribution of the point-winning shots and error-shots and what is the difference in mens tournaments compared to womens tournaments?*

Another important question that we can ask is the of the point-winning shots and the error-shots. This gives us an idea of what percentage of shots played by the player result in a net positive or negative. And further, we can compare the men's and women's tournaments to get a understanding of the performance of players based on gender.

C. ***Question C:*** *What is the trend of Double Faults by the players in each round of the Tournament?*

A double fault is by far one of the most gravest mistakes that a player can make in a game by making two wrong serves in a row. It is one of the worst ways to lose a point in a game. So intuitively, we would expect the trend of committing a double fault to go down as the tournament progresses into higher rounds.

D. ***Question/Hypothesis D:*** *What are the winning percentages on the second serve compared to the first serve for a player? Hypothesis – the second serve is less likely to win a point (Backfire effect)*

This question aims to answer multiple things. First of all, we can see what is the average percentage of getting a serve right, be it the first serve or the second serve. Further, how likely it is to win a point on the first or second serve.
As for the confirming Backfire effect, a tendency to react to disconfirming evidence by strengthening one's previous beliefs, we are hypothesizing that if a player is attempting a second serve (after making a fault on their first serve), he/she might be low on confidence which backfires on their ability to get a better winning serve on the second try.

E. ***Question E:*** *What is the ratio of the Net Points won to the Net Points attempted and what is the trend for it progressing into the higher rounds of the tournament?*

Net Points are one of the most staple forms of winning a point in tennis, it happens when a player approaches the net while attempting a shot. As the tournament progress, one would expect the success ratio (win-attempt ratio) to increase as players are playing for higher stakes and the matches are played by statistically better players.

F. ***Question F:*** *What is the trend between the Total Points Won by Players 1 and 2 categorized by the outcome of the match?*

This is an interesting question, that should have a rather definite answer. We want to see what is the number of total points won by the players and what is the outcome of the match corresponding to which player has the higher number of points. It is also a good example for a clustering question.

G. ***Question G:*** *What is the correlation between the different attributes of Player 1 and the outcome of the match?*

We would like to know the attributes that play the most important factor in a player winning or losing a game. Be that the highest positive correlation or the highest negative correlation.

H. ***Question H:*** *How do the championship winners play, specifically the number of Aces hit by them?*

Trying to learn from the best, we track the trend for the number Aces (one of the most important and efficient ways to get a point) won by championship contendors in the tournament.

III. Details of Libraries used

A. *Pandas*

Pandas is a powerful and open-source python library for data manipulation and analysis in Python. It provides a fast, flexible, and expressive data structure for working with large structured data. The primary data structure used is the DataFrame, a two-dimensional table with labelled axes. It also includes Series, which is a one-dimensional labelled array.
Pandas provides a wide array of tools comprising of various functions and methods for manipulating and analysing data, including data cleaning, transformation, and aggregation. It can also be used to plot said data without the aid of the matplotlib library.
As such, it is no surprise that Pandas happens to be the de-facto standard Python library for working with data and is a popular choice for data professionals and researchers alike.

B. *Matplotlib*

Matplotlib is a popular data visualization library in Python that has a variety of tools for creating visualizations, both static and animated. It is widely used in data analysis to understand data visually.
Matplotlib provides a variety of plots such as line plots, bar plots, histograms, scatter plots, distribution plots and many more. These plots are also highly customisable, allowing users to change everything from axis limits, colours, font styles, titles, labels and much more. The plots are also publication-quality which is a great plus point.
Moreover, Matplotlib is a versatile library and can be used with a variety of data structures such as lists, arrays, as well as Pandas Series and DataFrames.

C. *Seaborn*

Seaborn is another data visualization library in Python that is built on Matplotlib. Seaborn however, provides a higher-level interface for creating attractive and informative statistical graphics.
Seaborn is primarily designed to work with Pandas DataFrames and provides functions for creating various types of plots, and has much more options than Matplotlib. It also provides functions for statistical data visualization such as regression, distribution, and categorical plots.
This is an advanced library that is really helpful in plotting advanced plots and as an added bonus, the plots are much more aesthetic and pretty.

D. *Numpy*

Although not used much in this data narrative, NumPy is one of the most essential libraries to use when getting started with data analysis. Useful libraries of the likes of Pandas and Scipy are built upon existing funcitons and methods of the NumPy library.

NumPy, short for Numerical Python, is a powerful library for scientific computing Python. One of the key features of it is the multidimensional array object, ndarray, and various derived object classes from it.

IV. Answers to the Questions (with appropriate illustrations)

A. *Question A:*

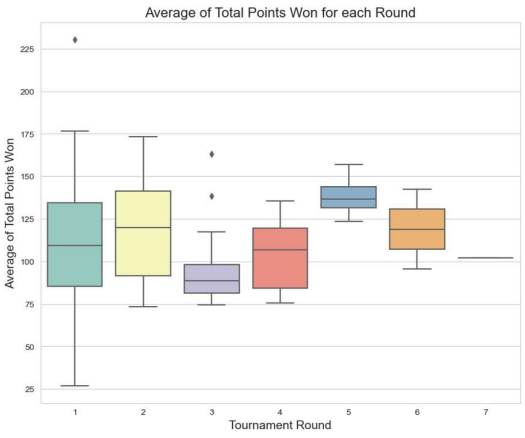

Fig. 1
Boxplot depicting the average of total points scored in a game as the rounds progress along with standard deviation for the same.
Dataset – Australian Open Men.

This boxplot shows that the average number of total points scored lies in the range of 100-125 without any considerable variation for the round in which the match is played. However, there are some outliers in the data as can be seen from the standard deviation being quite varied for each round. Another notable observation is that the variation is significantly lower in the higher rounds. This could be a result of two reasons: a) smaller dataset as a lesser number of matches take place in the higher rounds; and b) more consistent and better players that keep a tighter spread of the points scored.

B. *Question B:*

Comparing the winning shots and error shots for both men and women in the same tournament lets us determine the difference without having to consider external factors such as court surfaces. From the KDE plot of Fig. 2.1, we can see that women tend to have around 35 error shots and 20 winning shots per game.

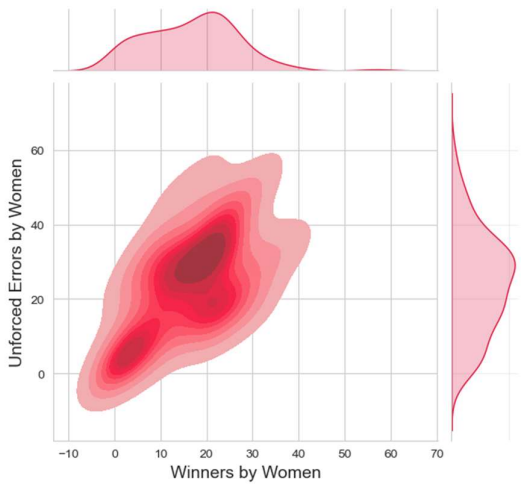

Fig. 2.1
KDE plot depicting the density distribution of the winning shots and error shots in the women's championship.
Dataset – Australian Open Women.

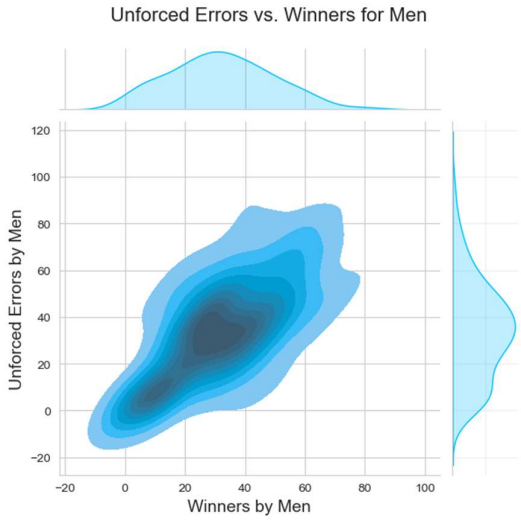

Fig. 2.2
KDE plot depicting the density distribution of the winning shots and error shots in the men's championship.
Dataset – Australian Open Men.

On the other hand, from Fig. 2.2, we can see that men also play an average of around 35 error shots but 33 winning shots which is higher than the those of the women in a game.

C. *Question C:*

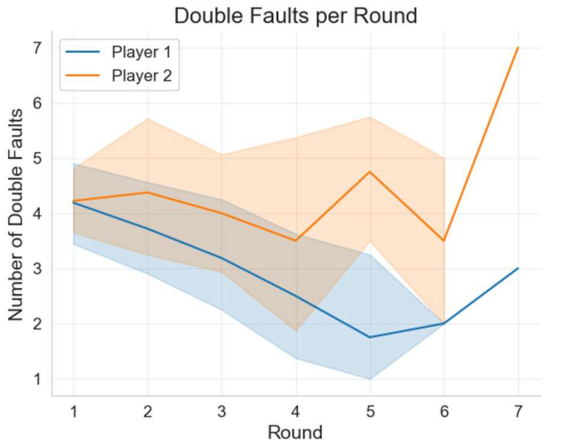

Fig. 3
Trend of the number of Double Faults as the tournament progresses into higher rounds.
Dataset – Australian Open Women.

From the plot, we can clearly see that number of Double Faults committed by both players of a game goes down. Except for the last round where there is a large spike, but that can be attributed as a sampling error as only one match takes place in that round. Coming back to the plot, the trend intuitively makes sense because as the tournament progresses, only the good players remain and as the stakes are higher, players tend to commit lesser number of faults.

D. *Question D:*

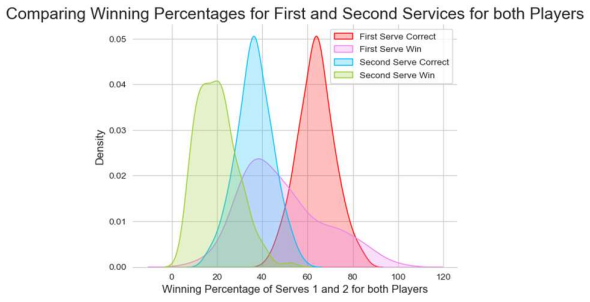

Fig. 4
A multi-KDE plot showing the relationships between the first and second serve percentages and wins.
Dataset – French Open Men.

Let us take the time to understand each plot individually here. The first-serve percentage, referring to the percentage of times when a player got the first serve correct, lies in the range of 60-65%. But the number of those serves which actually result into a win are lower, at around 40%. Now, if a player commits a fault in the first serve, they get a second try. But as a result of already losing the first serve, their confidence is less, and so the chance of getting the second serve right is drastically reduced to around 35% which half of the first-serve percentage. And further, the second-serve win percentage is even lower at just 20%. We can clearly see the Backfire effect in work here to confirm our initial hypothesis.

E. *Question E:*

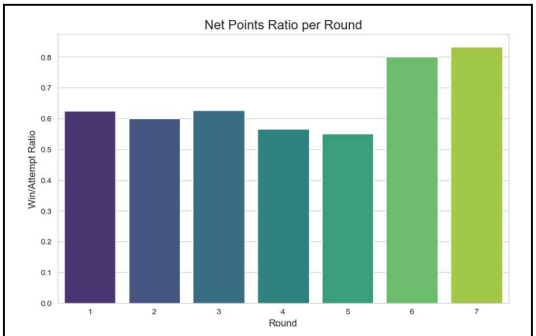

Fig. 5
A bar plot to show the ratio of Net Points that are won to the ratio of Net Points attempted as the rounds progress.
Dataset – French Open Women.

From the plot in Fig. 5, we can see that the win-attempt ratio of Net Points does not waver that much as the rounds progress, except for in the last round. This can, again, be attributed to the following two reasons: a) Sampling error as significantly less matches are played in the latter two rounds; and b) the matches are held being between better players who have already won previous rounds. This can also be considered as a classic example of the Survivorship Bias as we are only looking at the data of filtered and hardened players and is not a representation of the wider dataset.

F. *Question F:*

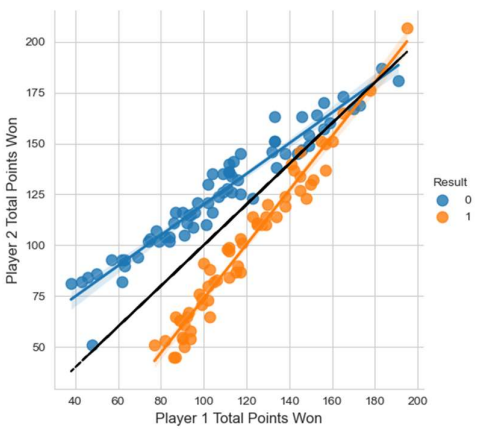

Fig. 6
Jointplot to show the number of total points won by players of a match, grouped by the result of the match.
Dataset – US Open Men.

The plot in Fig. 6 shows a clean clustering in the number of points won by each player. When Player 1 wins more total points, they are the one to win the match and vice versa for Player 2. The clustering in the plot is distinct and clear to imply that the player who wins the greater number of total points is also the one who wins the match.

G. *Question G:*

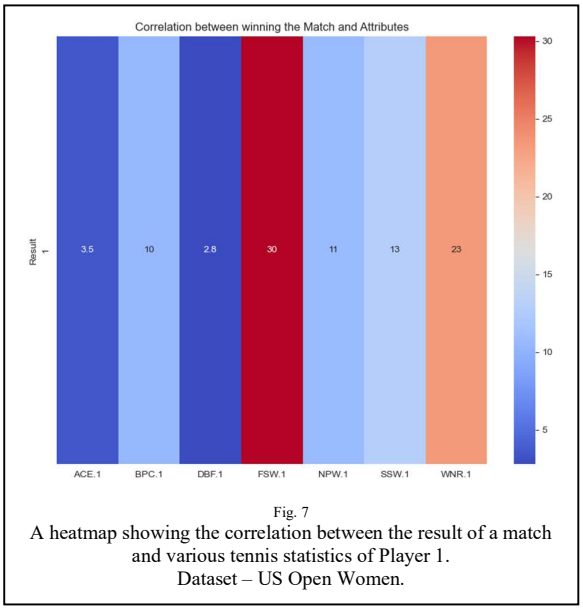

Fig. 7
A heatmap showing the correlation between the result of a match and various tennis statistics of Player 1.
Dataset – US Open Women.

Reading the heatmap in Fig. 7, we can clearly see that certain attributes are highly correlated to winning a match compared to others. With attributes like FSW.1 (first-serve win percentage) and WNR.1 (the number of winning shots) being having the highest correlation. On the other hand, attributes such as DBF.1 (number of double faults committed) have a strong negative correlation to not winning the match. Note that the one key takeaway from this data is that winning on the first serve is the most important thing to do in order to win a tennis match.

H. *Question H:*

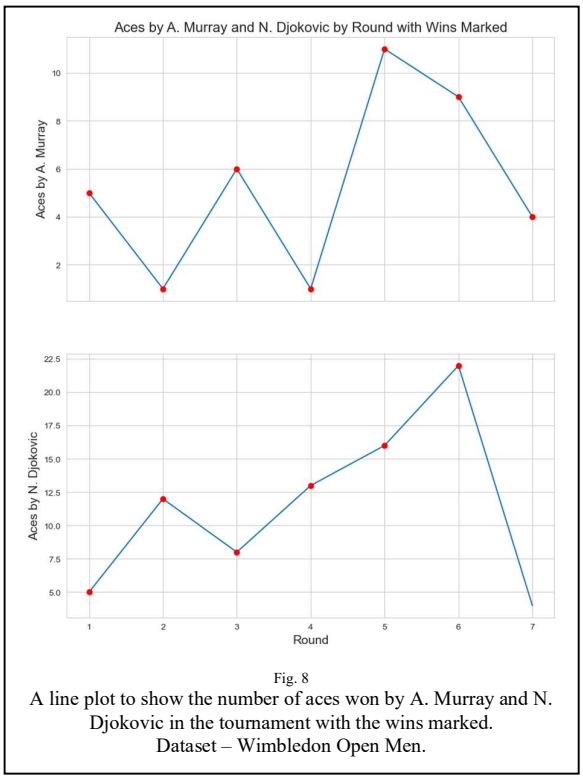

Fig. 8
A line plot to show the number of aces won by A. Murray and N. Djokovic in the tournament with the wins marked.
Dataset – Wimbledon Open Men.

From the line plot, we can see that both the player have a higher tendency to win Aces and not just that, but in the final match, the number of aces scored by Djokovic takes a rather sharp decline. This might be one of the reasons for him losing the match.

V. Summary of the Observations

The following are some key observations from the Tennis Statistics Dataset:

- The average points scored in a tennis match is in the range of 100-125.
- The average winning shots are around 30 and the average error shots are around 35 on average.
- Double Faults have a strong negative correlation with winning the match and, the number of double faults committed by players decreases as the rounds of the tournament progress.

- Second serves are tricky and less likely to result in a correct win compared to first serves, cue Backfire effect.
- First serves are very crucial, having the highest correlation to winning a match. And so are the number of aces.
- The win-attempt ratio for Net Points increases as the rounds progress, cue Survivorship Bias.

VI. Unanswerable Questions

While the given dataset contains a lot of information, 42 attributes to be precise, there are still some limitations to the kinds of question we can answer.

Following are some of the questions which I came across while thinking about the dataset that I believe could not be answered with the current data:

- How have the players improved compared to previous games?
  This question could not be answered as there is no information on previous championships.

- What is the age or nationality distribution of the players?
  This question was also not answerable due to a lack of sufficient data.

- How do players compare across the four different championships?
  The reason this question wasn't answered was because concatenating 4 large DataFrames using Pandas would become truly enormous and getting anything useful out of it would be cumbersome.

VII. References

Following are the websites that I could help from while working this report. Note that these are just the links to the parent website where the help was taken from, I could not reference every single question or doubt that I referred to these websites for.

[1] https://pandas.pydata.org/docs/
[2] https://seaborn.pydata.org/
[3] https://matplotlib.org/stable/index.html
[4] https://stackoverflow.com/
[5] https://www.geeksforgeeks.org/pandas-tutorial/
[6] https://www.usta.com/en/home/improve/tips-and-instruction/national/tennis-terms-definitions.html
[7] https://tenniscompanion.org/basics/

VIII. Acknowledgement

I would like to thank Professor Shanmuganathan Raman for all the data narrative assignments we have done. It was an excellent learning experience, getting to understand data and work with it, and discovering multiple Python libraries in the process. It gave us the opportunity to apply all that we had learnt in the class to something that was much more practical and required work from our side. It helped us improve not just our skill of working with data but also our skills of researching, writing reports and getting better at Python.