Data Narrative: Major Tennis Tournaments of 2013

Sriram Srinivasan
Electrical Engineering
Indian Institute of Technology, Gandhinagar
Gandhinagar, India
sriram.srinivasan@iitgn.ac.in, Roll No. 22110258

***Abstract*— The following report aims to raise significant scientific questions based on the eight datasets containing the match statistics for both women and men at the four major tennis tournaments of the year 2013. Each dataset has 42 columns and a minimum of 76 rows. The datasets consist of a variety of information like the first serve accuracy, number of aces, double faults committed, unforced errors and the set scores of each match.**

The raised questions will be answered by using the following statistical modules in Python – NumPy, Pandas and Matplotlib. All the statements will be backed by evidence using figures, tables and plots.

I. Overview of the Dataset

The datasets have been carefully curated by Shruti Jauhari, Aniket Morankar and Ernest Fokoue. They consist of a variety of well-compiled information about the statistics of the 2013 tennis tournaments – Australian Open, French Open, Wimbledon and US Open. The eight datasets can be found in the following Machine Learning repository: https://archive-beta.ics.uci.edu/dataset/300/tennis+major+tournament+match+statistics

II. Scientific Questions/Hypotheses

A. *The first service percentage directly reflects the confidence of the player as it is independent of the opponent. Consider the Big Three players - Roger Federer, Novak Djokovic and Rafael Nadal, and analyse the confidence of the players in the matches they won compared to the ones they lost. (Men's Australian Open 2013)*

B. *Analyse the evolution in the performance of the finalists of the Women's Australian Open 2013 - Dominika Cibulkova and Na Li with the passage of rounds based on the fraction of total points scored against the respective opponents.*

C. *Consider the finalists of the Men's French Open 2013 - Rafael Nadal and David Ferrer. Irrespective of the result of the final, people look up to the players who are consistent throughout the tournament. Analyse the consistency of the players based on the number of unforced errors and double faults committed.*

D. *Players who are the most skilled in hitting aces and winners do not necessarily win a tournament. They might lack consistency but are often the record creators. With respect to the Women's French Open 2013, find the top 3 players who have mastered hitting aces and winner shots.*

E. *Find the top 10 players who have won the maximum number of breakpoints in the Men's US Open 2013. This shows their resilience against an opponent who has a strong advantage of serving.*

F. *Tournaments tend to get more competitive as the games progress. With respect to the Women's US Open 2013 tournament, find the average number of sets played in a game with each passing round.*

G. *Is there a correlation between the number of aces served and the number of winners hit? Specifically, are the players who serve aces frequently equally skilled at hitting winner shots? Analyse this trend with respect to the Men's Wimbledon 2013 tournament.*

H. *Find the top 5 players who have attempted the maximum number of net points in the Women's Wimbledon 2013. This shows their fearlessness and a strong drive to snatch a point by taking some risks.*

III. Details of Libraries and Functions

The following libraries will be used to analyse the given dataset and answer the above questions:

A. Pandas

Pandas provides powerful data structures such as Series and DataFrame that allow for the manipulation and analysis of structured data. It has a variety of functions for reshaping, grouping, merging and plotting data that increase the efficiency and performance of data analysis tasks.

pandas.read_csv(): Reads data from a CSV file and creates a Pandas DataFrame for easy manipulation.

pandas.DataFrame.group_by(): Groups data based on one or more columns and allows for applying aggregate functions to each group.

B. NumPy

Numpy is a library used for numerical computing in Python. It provides a powerful array object that enables one to perform mathematical operations on large datasets. Numpy has a wide range of mathematical functions for operations such as linear algebra, Fourier transforms, and random number generation.

np.mean(): It is used to calculate the average or mean of an array or a portion of an array along a specified axis.
np.median(): It is used to compute the median value of an array or a portion of an array along a specified axis. The median is the middle value of a sorted array, or the average of the two middle values if the array has an even number of elements.

C. Matplotlib

Matplotlib is a plotting library in Python that provides various functions for creating visualizations such as line charts, scatter plots, bar graphs and histograms.

xticks(): It is used to set or get the x-axis tick locations and labels of a plot.

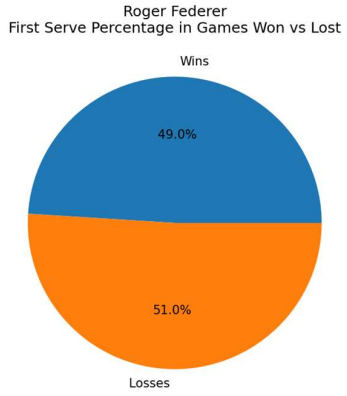
twinx(): It can create a twin y-axis that shares the same x-axis as the original plot.

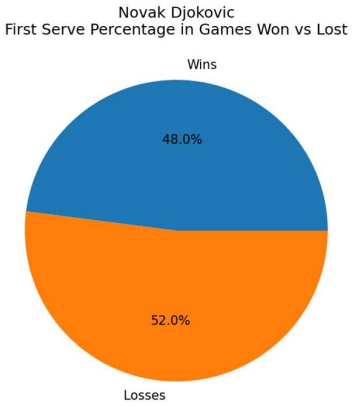
gca(): It is used to get the current axes instance of the current figure.

IV. Answers to the Questions (with Appropriate Illustrations)

A. Question 1: The first service percentage directly reflects the confidence of the player as it is independent of the opponent. Consider the Big Three players - Roger Federer, Novak Djokovic and Rafael Nadal, and analyse the confidence of the players in the matches they won compared to the ones they lost. (Men's Australian Open 2013)

Solution: The code analyzes the first serve percentage (FSP) of the "Big Three" tennis players (Roger Federer, Novak Djokovic, and Rafael Nadal) during the Australian Open 2013. The program uses a for loop to iterate over each player, and for each player, it calculates the average FSP during the games that the player won and lost. The results are stored in a dictionary with each player's name as the key. Finally, the code creates a pie chart for each player, showing the percentage of first serves won during games won vs lost. The pie charts are displayed using Matplotlib's Pyplot module.





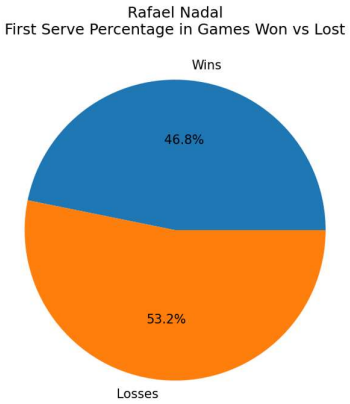


Figure 1: First serve percentage of the 'Big 3' in wins vs losses

B. Question 2: Analyse the evolution in the performance of the finalists of the Women's Australian Open 2013 - Dominika Cibulkova and Na Li with the passage of rounds based on the fraction of total points scored against the respective opponents.

Solution: The code loads tennis match statistics data of Australian Open Women's 2013 from a CSV file and extracts the total points won percentage by two finalists, Na Li and Dominika Cibulkova, for each round of the tournament. It calculates the fraction of total points scored by each finalist in each round and plots it as a bar chart, with different colors for each player. The x-axis shows the rounds of the tournament and the y-axis shows the fraction of total points scored by each player.

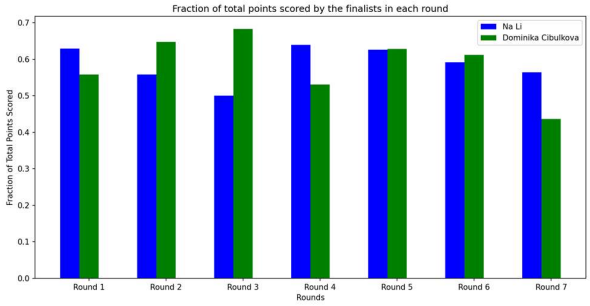


Figure 2: Fraction of total points scored by the finalists by round

C. Question 3: Consider the finalists of the Men's French Open 2013 - Rafael Nadal and David Ferrer. Irrespective of the result of the final, people look up to the players who are consistent throughout the tournament. Analyse the consistency of the players based on the number of unforced errors and double faults committed.

Solution: The code uses Pandas to load tennis match statistics from a CSV file for the French Open 2013 men's tournament. It loops through the seven rounds of the tournament and uses conditional statements to filter the data for each round and the finalists. It then creates two bar plots using Matplotlib to show the number of unforced errors and double faults committed by each finalist in each round.

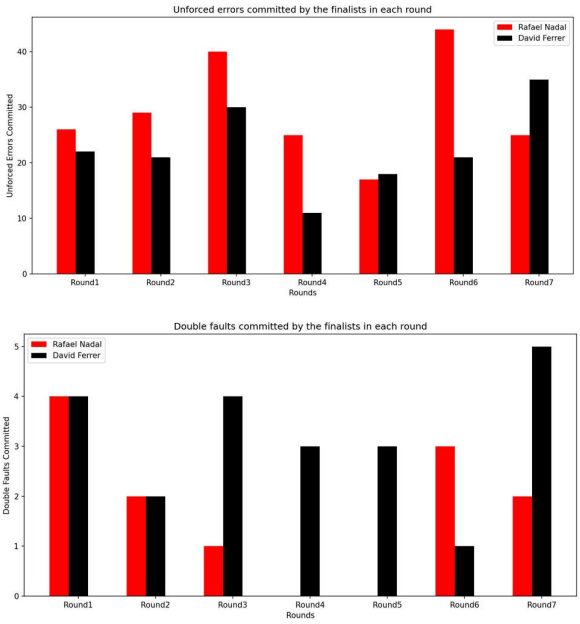
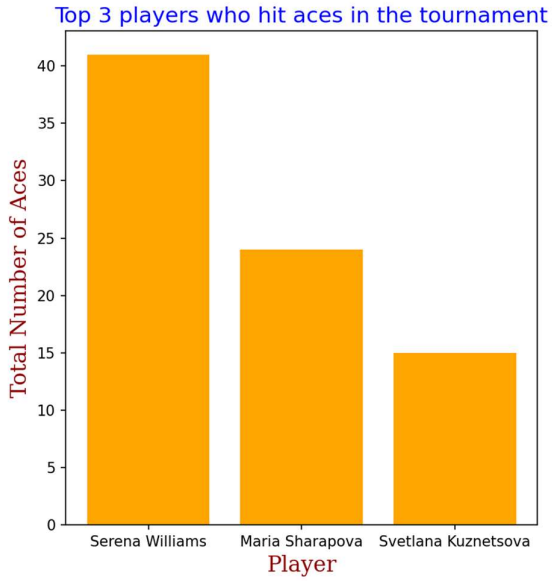


Figure 3: Unforced errors and double faults by the finalists by round

D. *Question 4: Players who are the most skilled in hitting aces and winners do not necessarily win a tournament. They might lack consistency but are often the record creators. With respect to the Women's French Open 2013, find the top 3 players who have mastered hitting aces and winner shots.*

Solution: The code loads a CSV file of match statistics from the 2013 French Open women's tournament into a Pandas dataframe. The code then creates a new dataframe with columns for the number of aces and winners hit by each player. The data is then grouped by player, and the number of aces and winners for each player is summed. The top 3 players for aces and winners are then plotted using a bar chart created with Matplotlib.



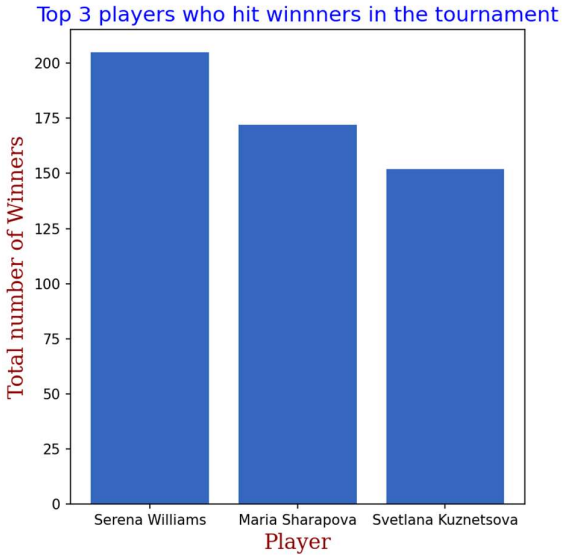


Figure 4: Top 3 ace servers and winner shot hitters

E. *Question 5: Find the top 10 players who have won the maximum number of breakpoints in the Men's US Open 2013. This shows their resilience against an opponent who has a strong advantage of serving.*

Solution: This code reads in a CSV file of men's US Open tennis match statistics from 2013 using the pandas library. It extracts the players' names and the number of break points they won in each match. It then creates a new dataframe which groups them by player. The resulting dataframe is sorted in descending order based on the total number of break points won by each player, and the top 10 players are displayed. The output shows the name of the player and the total number of breakpoints they won in the tournament.

Player	Break Points Won
David Ferrer	84
Novak Djokovic	83
Richard Gasquet	80
Rafael Nadal	75
Mikhail Youzhny	66
Stanislas Wawrinka	66
Milos Raonic	60
Lleyton Hewitt	56
Roger Federer	54
Tomas Berdych	50

Figure 5: Top 10 players who won the maximum number of breakpoints

F. *Question 6: Tournaments tend to get more competitive as the games progress. With respect to the Women's US Open 2013 tournament, find the average number of sets played in a game with each passing round.*

Solution: This code loads data on the 2013 US Open women's tennis tournament into a Pandas DataFrame. It then groups the matches by round and calculates the average number of sets played in each round. Finally, it creates a lollipop graph of the average sets played per round, with the x-axis showing

the round and the y-axis showing the average number of sets played.

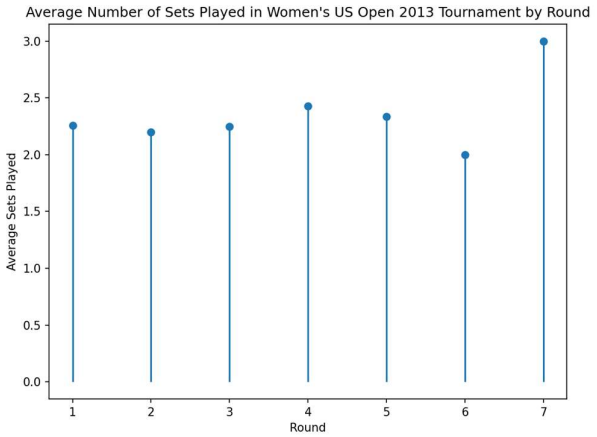


Figure 6: Average number of sets played in Women's US Open 2013 by round

G. *Question 7: Is there a correlation between the number of aces served and the number of winners hit? Specifically, are the players who serve aces frequently equally skilled at hitting winner shots? Analyse this trend with respect to the Men's Wimbledon 2013 tournament.*

Solution: The code imports necessary libraries and loads the Wimbledon Men's 2013 tennis match data from a CSV file. It then creates a scatter plot with the x-axis showing the number of aces served by each player and the y-axis showing the number of winners hit by each player.

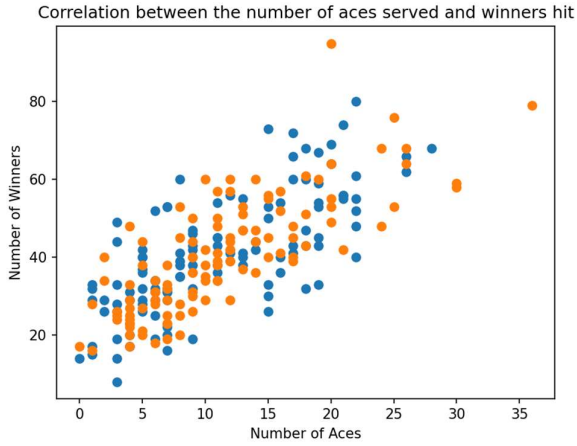


Figure 7: Correlation between the number of aces served and winners hit

H. *Question 8: Find the top 5 players who have attempted the maximum number of net points in the Women's Wimbledon 2013. This shows their fearlessness and a strong drive to snatch a point by taking some risks.*

Solution: This code reads a Wimbledon women's tennis match statistics CSV file using pandas. It selects the net points attempted by each player and groups them by player, then sorts the result in descending order. Finally, it prints the top 5 players who have attempted the most net points in the tournament.

Player	Net Points Attempted
A.Radwanska	170
K.Flipkens	151
S.Lisicki	140
N.Li	133
R.Vinci	115

Figure 8: Top 5 players who attempted the maximum number of net points

V. Summary of the Observation

A. *Confidence of the 'Big 3' player in games won vs lost (Men's Australian Open 2013)*

Roger Federer, Novak Djokovic and Rafael Nadal are considered the greatest players in the world. This can be seen in the way they handle their mental game. The first serve percentage does not change, irrespective of whether they win or lose a game. This might not be the case with other emerging players who might miss their serves in high-pressure tournaments. Moreover, all of them have a marginally worse first-serve percentage in the games they won. This might be because they try to serve aces when they are winning to finish off a game sooner.

B. *Evolution in the performance of the finalists of the Women's Australian Open 2013*

From the graph, it appears that Na Li has mostly scored a greater fraction of points against her points compared to Dominika Cibulkova, except in round 3 where Dominika Cibulkova scored a significantly higher fraction of points. This reflects Na Li's consistency and momentum, which ultimately paid off in the finals against Dominika Cibulkova.

C. *Consistency of the finalists: Rafael Nadal and David Ferrer (Men's French Open 2013)*

Rafael Nadal has committed a significantly higher number of unforced errors in rounds 1 to 6 compared to David Ferrer. David Ferrer has been the most consistent player, considering unforced errors. On the other hand, David Ferrer has made a huge number of double-fault mistakes. This weakness, which is clearly visible in the trends, turned out to be fatal in the finals. Rafael Nadal, being one of the greatest players, perfomed well under pressure in the finals despite having an inconsistent form.

D. *Top 3 ace servers and winner shot hitters in Women's French Open 2013*

Serena Williams, Maria Sharapova and Svetlana Kuznetsova hold a common record in the exact same order of serving the maximum number of aces and hitting the maximum number of winners. This introduces a possible question of correlation between serving aces and hitting winners, which will be addressed later.

E. *Top 10 players who won the maximum number of breakpoints*

As displayed in the table, David Ferrer has won 84 breakpoints and is closely followed by Novak Djokovic. It is no doubt that all three – Novak Djokovic, Rafael Nadal and Roger Federer feature in the top 10. They are known for their fighting back skills in all kinds of situations.

F. *Competitiveness of Women's US Open 2013 based on the average number of sets played by round*

Unlike the expectation of an increasing trend in the number of sets as the rounds progress, they seem to lie around mostly between 2 and 3 in all the rounds. This might be because the competition might be quite high even when two players of lower levels play in the earlier rounds. The number of sets is directly related to the closeness of the level of the two players.

G. *Correlation between the number of aces served and winners hit*

Looking at the scattered plot, there seems to be a direct correlation between the number of aces served and winners hit. A line can be fitted to the scatter points. This shows that the skills of serving aces and hitting winners are similar, and they improve simultaneously.

H. *Top 5 players who attempted the maximum number of net points*

A.Radwanska has attempted 170 net points and is closely followed by K.Flipkens and S.Lisicki. They have attempted 151 and 140 points, respectively.

ACKNOWLEDGMENT

I would like to thank Professor Shanmuga Raman for giving me the opportunity to work on the dataset provided and use my data analysis skills to extract valuable information and trends from the source.

REFERENCES

[1] "API Reference — Pandas 1.5.3 Documentation," n.d. https://pandas.pydata.org/docs/reference/index.html#api.

[2] "NumPy Documentation," n.d. https://numpy.org/doc/.

[3] "API Reference — Matplotlib 3.7.0 Documentation," n.d. https://matplotlib.org/stable/api/index.html#.