ES-114

Data Narrative

Hardik Jain
Mechanical Engineering
Indian Institute of Technology
Gandhinagar
Roll Number - 22110091

I. Overview of the Dataset

The given datasets contain information about every match of the four Grand Slam tournaments that happen each year. These datasets have data about the matches of Men's Singles and Women's Singles formats that take place. It provides us with information about statistics like the number of aces, number of winners, number of unforced errors, number of double faults, number of net points, number of break points, score and result of a match and more. Using these datasets, lot of insightful things can be derived which can help in studying a player.

II. Scientific Questions/Hypotheses

Question 1 (Men's Australian Open 2013)

Who among the men won the Australian Open in 2013? Was he consistent with his strong serve or good shots throughout the tournament? Did his strong serve or good shots contribute more to his win?

Question 2 (Women's Australian Open 2013)

Is the performance index of players in the first two rounds a good measure to predict the players who will reach the final stages of the tournament (quarterfinals and further)? If not, what could be the reason?

Question 3 (Men's French Open 2013)

What is the aggression level among the players? Are there some players who play very aggressively or are there some who are not so aggressive?

Question 4 (Women's French Open 2013)

Is there a relation between the number of faults and unforced errors made by a player in a match and them winning or losing that match? Is it true that less number of faults and unforced errors increase the chance of a player winning a match?

Question 5 (Wimbledon Men 2013)

Which players completely dominated their opponent (won the match in straight sets) the greatest number of times? What factor contributed most towards them dominating their opponent?

Question 6 (Wimbledon Women 2013)

In general, do players having higher number of break points in a match go on to win the game? If not, what could be the reason?

Question 7 (Men's US Open 2013)

How has the dynamics of the players who played the other Grand Slams as well changed here? Are there any players who played consistently on all surfaces? Did the consistency help them get far in the tournament?

Question 8 (Women' US Open 2013)

Is there a variation among the playing style of the players as the tournament progresses and the opponent changes? Is this change of their playing style effective in them winning?

III. Details of Libraries and Functions

Pandas - Pandas is a Python library that is widely used for working with data sets. It provides us with various features which enable us to analyse and manipulate data efficiently [1]. It has data structures like series and dataframe which store data in a one- dimensional and two-dimensional form respectively. By utilising the different features of the Pandas library, we can derive a lot of inferences from the dataset.

Matplotlib - Matplotlib is a Python library that enables us to plot data visually, thus making it easier for us to interpret it. It has various types of plots available such as line, bar, histogram, pie, scatter, area, boxplot et cetera. The plot is chosen according to the type of data that is available.

Seaborn - Seaborn is a Python library that is built on top of the Matplotlib library and it also helps in visual interpretation of data. It also enables us in plotting line, bar, scatter et cetera. Another feature of this library is that it helps in making the plots visually attractive [2].

Numpy – Numpy is a Python library that is used to deal with numerical data, arrays and matrices.

Sklearn – Sklearn is a Python library that is used for implementing machine learning algorithms. It includes various classification, regression , clustering and dimensionality reduction algorithms that are very useful. This library is built on top of the libraries like Numpy, Pandas and Matplotlib [3].

IV. Answers to the Questions

Question 1

Stanislas Wawrinka won the Australian Open in 2013.

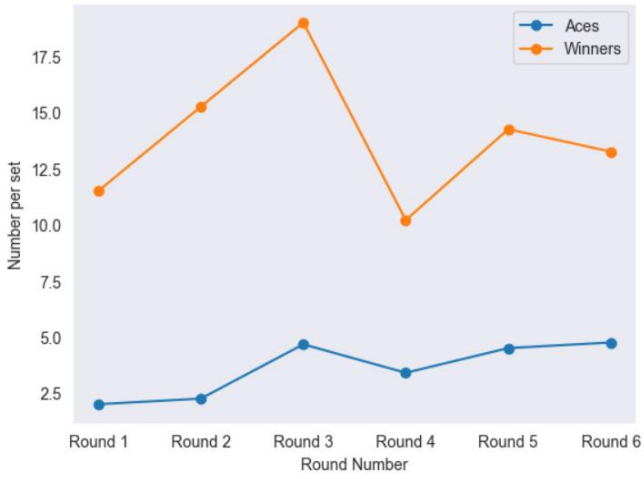

Fig. 1 – Line plot showing the variation in the number of aces and winners Stanislas Wawrinka hit in different rounds of the tournament

The line plot clearly shows that there has been some variation in the number of aces and winners per set in the different rounds that Stanislas Wawrinka played. The number of aces per set has relatively varied lesser than the number of winners per set. So, he was not very consistent with the number of winners he played per set but was fairly consistent with the number of aces.

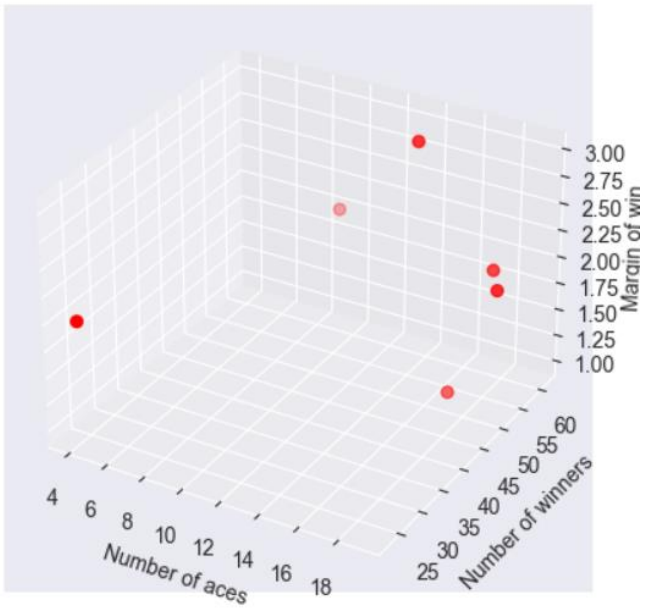

Fig. 2 – 3D scatter plot showing the distribution of Stanislas Wawrinka's number of aces and winners against the margin of win in that round

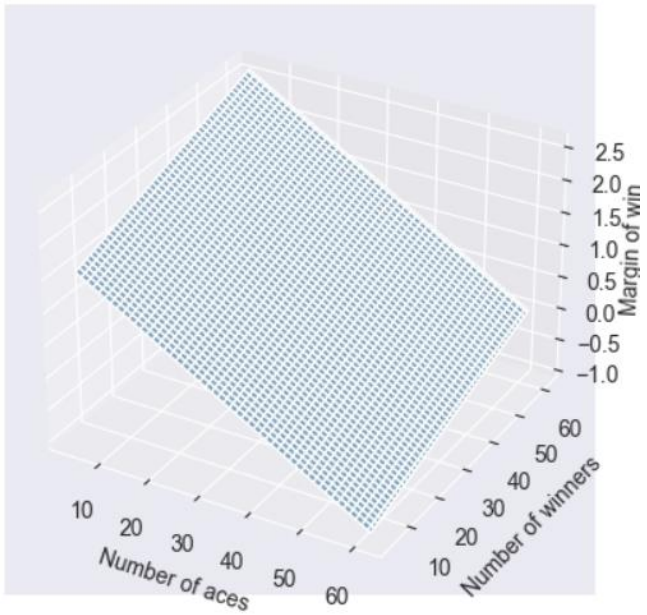

Fig. 3 – Regression plane

The slope of the regression plane against number of aces is -0.04560131 and the slope of the plane against number of winners is 0.01874743. This shows that the number of winners has a positive relation with the margin of win and the number of aces is not related to the margin of win. Hence, Stanislas Wawrinka's good shots (winners) contributed more to his win than his good serve (aces).

Question 2

Player	Performance index
Kirsten Flipkens	5.8859
Agnieszka Radwanska	5.8009
Serena Williams	5.575
Samantha Stosur	4.6864
Angelique Kerber	4.1788
Flavia Pennetta	3.6168
Monica Niculescu	3.5175
Magdalena Rybarikova	3.3178
Na Li	3.2321
Elina Svitolina	3.1869

Fig. 4 – Dataframe showing the top 10 players with highest performance indices in round 1

Player	Performance index
Jelena Jankovic	5.9795
Angelique Kerber	4.4401
Alison Riske	4.3994
Zarina Diyas	4.2329
Samantha Stosur	4.1476
Serena Williams	3.9686
Flavia Pennetta	3.5643
Agnieszka Radwanska	3.4287
Dominika Cibulkova	3.3098
Casey Dellacqua	3.1539

Fig. 5 – Dataframe showing the top 10 players with highest performance indices in round 2

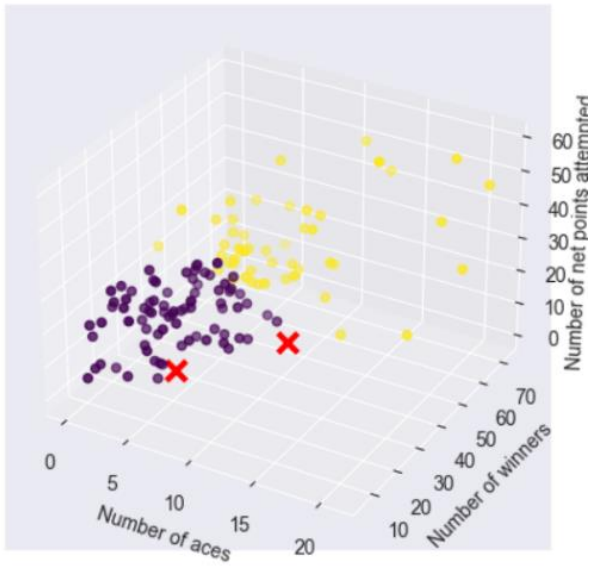

Fig. 6 – Clusters showing the distribution of players with different parameters that decide on an aggressive play style

The performance indices of these players were calculated by taking a weighted sum of various parameters which affect performance like number of winners, number of aces, number of unforced errors, number of double faults, number of break points and number of net points. The coefficients in the weighted sum were taken from the correlation matrix between these parameters and the result of the match.

Agnieszka Radwanska, Angelique Kerber, Flavia Pennetta, Samantha Stosur and Serena Williams are the women who were in the top ten players with the highest performance indices in both rounds one and two.

Out of these, the players who reached quarterfinals or above were Agnieszka Radwanska and Flavia Pennetta. The others got eliminated before even reaching the quarterfinals. Thus, the performance index of the players in the first two rounds is not a good criterion to predict the players reaching the final stages of the tournament. Some players having high performance index in the initial round may play a bad match in the third or fourth round and get eliminated. Hence, it is better to consider the performance index in all four rounds to predict the players reaching quarterfinals or further.

Question 3

The two clusters formed in the scatter plot clearly depict the aggression of the players while playing. The players in the purple cluster are less aggressive than the players in the yellow cluster.

Some players in the yellow cluster have a highly aggressive playing style as they have a very high number of aces, winners and net points.

Some players in the purple cluster are not at all aggressive and play safe as the values of all the parameters in their case are lesser.

Question 4

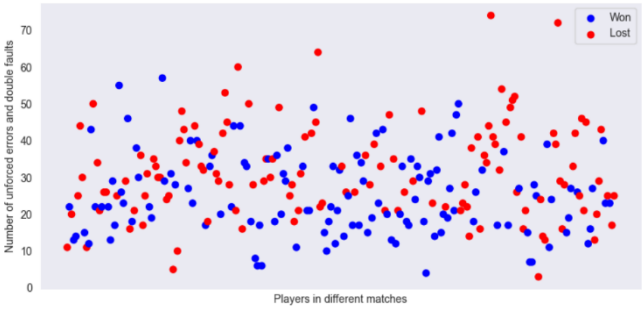

Fig. 7 – Scatter plot showing the distribution of players in different matches with the number of unforced errors and double faults they made in a match

	Player	Number of double faults	Number of unforced errors	Won/Lost	Total errors
0	Su-Wei Hsieh	3.0	8	L	11.0
1	Eugenie Bouchard	1.0	21	W	22.0
2	Eugenie Bouchard	2.0	18	L	20.0
3	Jie Zheng	0.0	13	W	13.0
4	Jie Zheng	0.0	14	W	14.0
...	...	...	...	...	...
249	Sorana Cirstea	1.0	22	W	23.0
250	Sorana Cirstea	6.0	19	L	25.0
251	Caroline Garcia	2.0	21	W	23.0
252	Caroline Garcia	3.0	14	L	17.0
253	Anna Tatishvili	3.0	22	L	25.0

Fig. 8 – Dataframe showing the number of faults made by different players in different rounds of the tournament and the result of that match

The scatter plot shows that the majority of the blue dots are concentrated in the bottom half of the graph. There are also some blue dots in the upper half of the graph which shows the players who won despite having higher number of errors. This implies that generally the people who won a match made lesser number of errors. So, making lesser number of errors and faults increases the probability of a player winning that match.

The red dots are scattered more than the blue dots but are more in number above 25-30 errors and faults. This shows that a player with having higher number of faults and errors has a lesser probability of winning.
Implementing Random Forest Classifier on the data which shows the number of faults and errors against the result of the match gives us a prediction with 87.79% accuracy. This shows that the prediction made by this algorithm was pretty accurate as the data is consistent with the players having more errors generally losing and players having fewer errors generally winning.

Question 6

	Number of break points created	Number of break points won	Won/Lost
Number of break points created	1.000000	0.748996	0.409229
Number of break points won	0.748996	1.000000	0.623851
Won/Lost	0.409229	0.623851	1.000000

Fig. 9 – Correlation matrix showing the correlation coefficients between a player creating and winning break points and the outcome of that match

The correlation matrix shows that the correlation coefficient between number of break points won and the player winning that match is 0.748996 which depicts a fairly strong positive correlation.

Fig. 10 – Scatter plot showing the distribution of players in different matches with the number of break points they won in that match

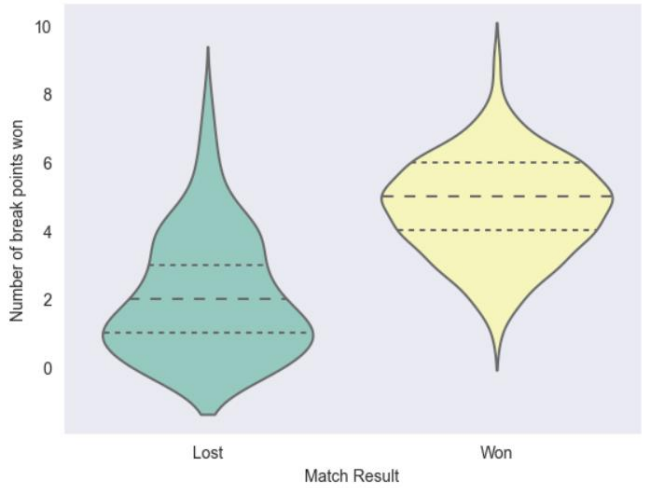

Fig. 11 – Violin plot showing the distribution of the players who won and lost against the number of break points they won

The scatter plot clearly shows the concentration of red and blue dots in the lower and upper half of the plot respectively. This is in accordance to the result derived from the correlation matrix that players having higher number of break points mostly win the game and players having lesser number of break points mostly lose the game.
The players having higher number of break who went on to lose the game could be because the opponent had a higher number of break points than this player implying that this player had a weak serve.

V. UNANSWERABLE QUESTIONS

Question 5

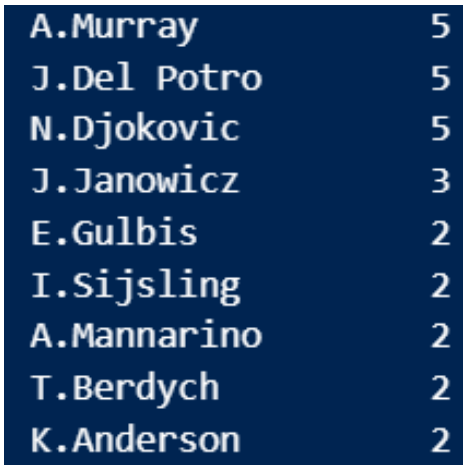

Fig. 12 – Figure showing the number of times some players won in straight sets

N. Djokovic, J. Del Potro and A. Murray won in straight sets the maximum number of times (5 times).

	Player	First Serve Percentage	First Serve Won	Second Serve Percentage	Second Serve Won	Aces	Double Faults	Winners	Unforced Errors	Break Points Created	Break Points Won	Net Points Attempted	Net Points Won	Margin of win
0	A.Murray	57	39	43	20	11	2.0	38	16	10	5	23	17	0
1	J.Del Potro	73	45	27	12	10	1.0	34	10	8	6	11	9	0
2	N.Djokovic	63	41	37	26	5	0.0	40	20	11	3	26	20	0
3	A.Murray	63	44	37	20	11	1.0	41	14	15	4	23	19	0
4	J.Del Potro	69	46	31	16	13	0.0	37	13	11	4	10	6	0
5	N.Djokovic	59	39	41	22	12	0.0	41	12	18	4	20	14	0
6	A.Murray	65	44	35	16	9	0.0	40	14	7	5	28	20	0
7	J.Del Potro	61	47	39	20	10	4.0	29	12	9	4	13	7	0
8	N.Djokovic	74	40	26	12	8	1.0	38	3	16	5	21	18	0
9	A.Murray	63	55	37	16	15	4.0	45	16	10	5	42	31	0
10	J.Del Potro	72	55	28	18	9	1.0	36	15	11	2	28	15	0
11	N.Djokovic	65	44	35	17	13	3.0	40	16	13	6	22	15	0
12	J.Del Potro	72	55	28	11	12	0.0	42	11	8	3	21	17	0
13	N.Djokovic	61	42	39	21	16	1.0	36	13	10	4	17	11	0
14	A.Murray	64	48	36	16	9	2.0	36	21	17	7	37	26	0

Fig. 13 – Dataframe showing the values of each parameter which helped them in dominating their opponent in the matches these three players played

In the dataset containing containing information about Wimbledon Men 2013, the columns having the labels ‘TPW.1’ and ‘TPW.2’ are empty. These columns should have contained the information about the total number of points player 1 and player 2 won in each match. Since these columns are empty, we cannot find the margin of win of that match and hence we cannot decide which factor contributes more towards a player winning in straight sets. Therefore, I was unable to answer this question.

Question 7

	WNR.1	UFE.1	WNR.2	UFE.2
0	NaN	NaN	NaN	NaN
1	NaN	NaN	NaN	NaN
2	NaN	NaN	NaN	NaN
3	NaN	NaN	NaN	NaN
4	NaN	NaN	NaN	NaN
...	...	...	...	...
121	NaN	NaN	NaN	NaN
122	NaN	NaN	NaN	NaN
123	NaN	NaN	NaN	NaN
124	NaN	NaN	NaN	NaN
125	NaN	NaN	NaN	NaN

Fig. 14 – Dataframe showing the empty values of the parameters 'WNR.1', 'UFE.1', 'WNR.2', 'UFE.2'

To understand the dynamics of the players in this tournament, the number of unforced errors and number of winners are important criteria to be taken into consideration. Since this dataset does not contain these values, it is not possible to understand the player dynamics in this tournament. Therefore, I was unable to answer this question.

Question 8

	Player	Number of aces	Number of net shots attempted	Won/Lost
0	S Williams	9	13.0	W
1	S Williams	4	11.0	W
2	S Williams	6	13.0	W
3	S Williams	6	13.0	W
4	S Williams	4	8.0	W
...	...	...	...	...
147	M Keys	1	13.0	L
148	L Dominguez Lino	6	5.0	L
149	L Arruabarrena	3	8.0	L
150	K Date-Krumm	2	14.0	L
151	A Medina Garrigues	8	14.0	L

Fig. 26 – Dataframe showing the playing style parameters of different players in different rounds of the tournament and the outcome of that match

To visualize the variation in the playing style of different players across the tournament, a line plot is the best. But, since there are many players, it was not practical to make a line plot for everyone. I couldn't find an alternative solution to this, and hence I couldn't answer this.

VI. References

[1] "pandas (software)," Wikipedia. Accessed Apr 24, 2023 [Online]. Available: https://en.wikipedia.org/wiki/Pandas_(software)

[2] GeeksforGeeks, "Introduction to Seaborn – Python." Accessed: Apr 24, 2023 [Online]. Available: https://www.geeksforgeeks.org/introduction-to-seaborn-python/

[3] "scikit-learn," Wikipedia. Accessed Apr 24, 2023 [Online]. Available: https://en.wikipedia.org/wiki/Scikit-learn