Data Narrative -3

ES-114

Name : Aryan Sahu

Roll no. : 22110038

I. Overview of the Dataset

The given dataset contains the match statistics for both women and men at the four major tennis tournaments of the year 2013. It includes various parameters such as the players' names, the round of the tournament, the number of games won, the number of aces, double faults, and unforced errors committed, the first and second-serve win percentage, the number of net points attempted and won, and the first serve percentage. The dataset provides insights into the players' performance based on various factors, which can help one know the properties of Tennis tournaments.

II. Scientific Questions/Hypotheses

We can understand many inferences by analysing the dataset. Following are some valuable inferences we would want to get from the given data:

A. *How does the first serve percentage (FSP) affect the outcome of a tennis match? Do players with higher FSP tend to win more matches?*

B. *How does the round of the tournament (e.g., early rounds vs. later rounds) impact the performance of tennis players?*

C. *Is there evidence in the given data to suggest that men and women handle performance pressure differently in tennis matches?*

D. *Does more experience in the game of tennis uplift the ability to handle failure? Derive a conclusion by plotting an approximate curve of the First-serve vs Second-serve win percentage.*

E. *The type of ground may affect the overall outcome of the game, but it has very little effect on the faults committed during a service in a tennis match. Support this statement using the dataset.*

F. *What are the chances that an attempt to win a point in the match succeeds? Use the given dataset to find the above statistically.*

G. *How can we show that in a tournament, most of the matches are one-sided? What can be the reason for this?*

H. *Find all the possible ways in which a match ends in a tournament. Compare the result with our proconceptions.*

III. Important Libraries and Functions Used

Following are some of the libraries and functions used to mine the dataset effectively and thoroughly:

A. *Pandas*

Pandas is a library offered by the Python programming language, which helps analyse tabular data. It offers to create data types consisting of rows and columns and helps analyse large amounts of data.

B. *Read CSV*

In the Pandas library, we are using the read_csv function to import the CSV file to read the data from and analyse it for our benefit.

C. *Dataframe*

We are creating dataframes as a medium to represent the data inside the Python program. Dataframes provide various in-built functions which are useful to derive important inferences from the given dataset.

D. *Matplotlib*

It is a library that offers plotting methods and data visualisation options. We have used this library to plot the diverse data and visualise it with the help of graphs.

E. *Numpy*

It provides options for convenient data management, and we have used it to perform regression techniques to visualise graphs accurately.

F. *Seaborn*

Seaborn is similar to the matplotlib library and we have used this as well to plot the various graphs.

IV. Solutions to the Questions

We are using the above-specified tools to mine the dataset and obtain various inferences. Following are the answers got from this process.

A. *How does the first serve percentage (FSP) affect the outcome of a tennis match? Do players with higher FSP tend to win more matches?*

We have made a scatter and box plot to visualise the contribution of FSP towards the match outcome in the Wimbledon Men's tennis tournament.

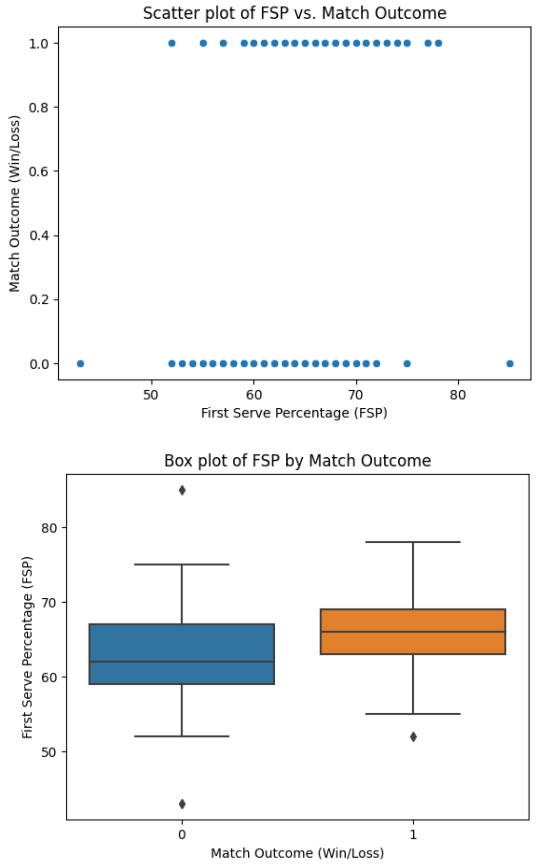

Fig. 1 Scatter plot and box plot of the First serve percentage vs Match Outcome.

From the plot, it is visible that the won matches tend to have higher values of First serve percentage. This is evidence pointing towards FSP being a positive contributing factor for the match outcome.

B. *How does the round of the tournament (e.g., early rounds vs. later rounds) impact the performance of tennis players?*

We have plotted the Average Unforced errors and the Average Double faults committed vs the round of the tournament to infer the performance standard of players according to the rounds.

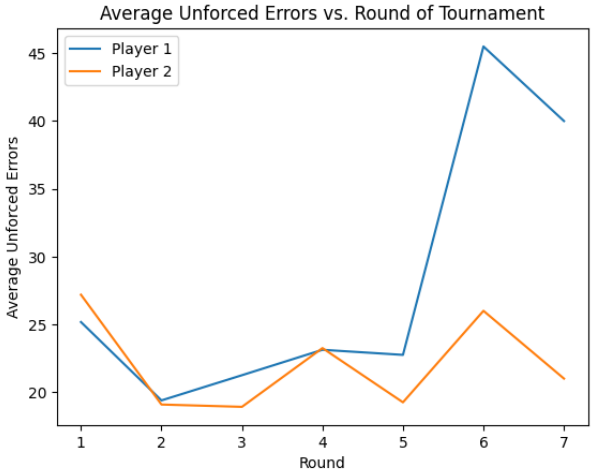

Fig. 2 Average Unforced Errors vs The Round of the Tournament.

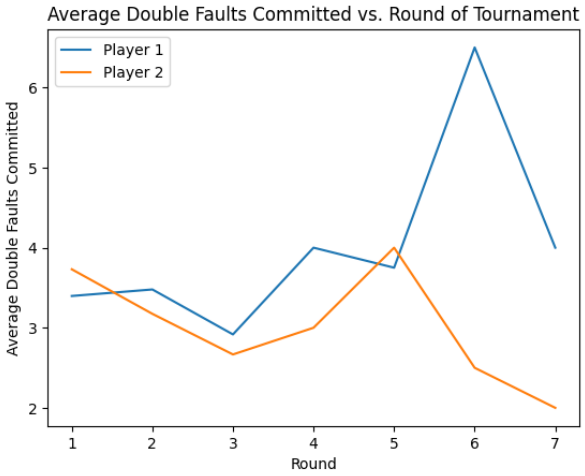

Fig. 3 Average Double Faults committed vs The Round of the Tournament.

We can see that as the rounds have progressed, the faults and errors committed have spiked up. This shows that the performance index has decreased slightly as we advance in the tournament. This could prove that the increase in performance pressure because of the greater importance of the round is a reason to lead to faults.

NOTE: Even though the last round i.e. 7th round, is the Average Unforced Errors vs The Round of the Tournament.most important, the players have committed fewer faults and errors. This shows that in the final rounds, players are more cautious about the game.

C. *Is there evidence in the given data to suggest that men and women handle performance pressure differently in tennis matches?*

In the last question, we saw how men had underperformed due to the pressure in the more important rounds of the tournament. Thus, we plot the same curves for women to compare their behaviour.

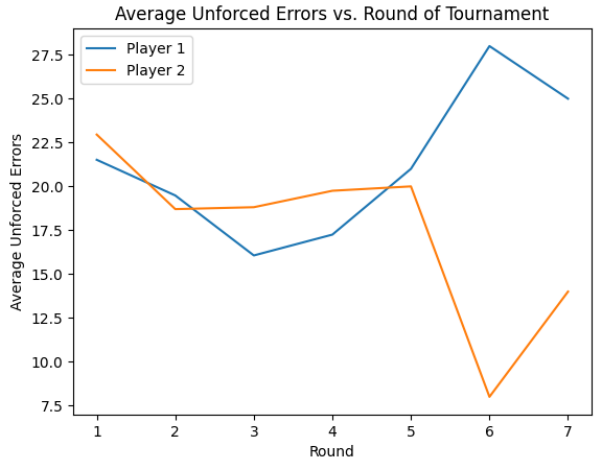

Fig. 4 Average Unforced Errors vs The Round of the Tournament.

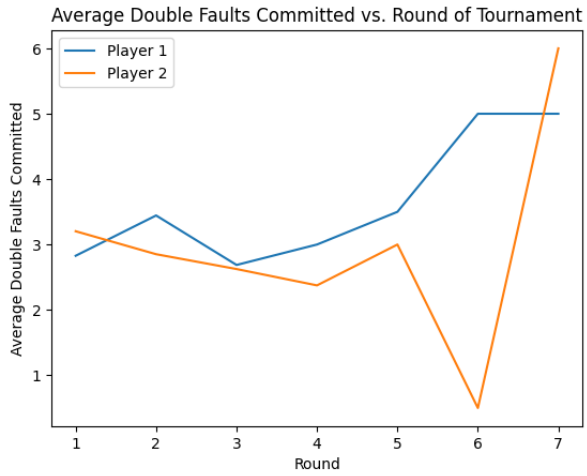

Fig. 5 Average Double Faults committed vs The Round of the Tournament.

We observe that the number of errors and faults in the case of women's matches was very similar to that of men. But unlike men, the faults and errors in the final round were much higher than they were in the previous round. Also, their faults in the semifinals were very few compared to other rounds. This shows that women can perform better under pressure compared to men in the case of tennis matches.

D. Does more experience in the game of tennis uplift the ability to handle failure? Derive a conclusion by plotting an approximate curve of the First-serve vs Second-serve win percentage.

We have plotted a scatter plot of the first and second-serve win percentage win percentage and used the regression method to approximate a relation between them.

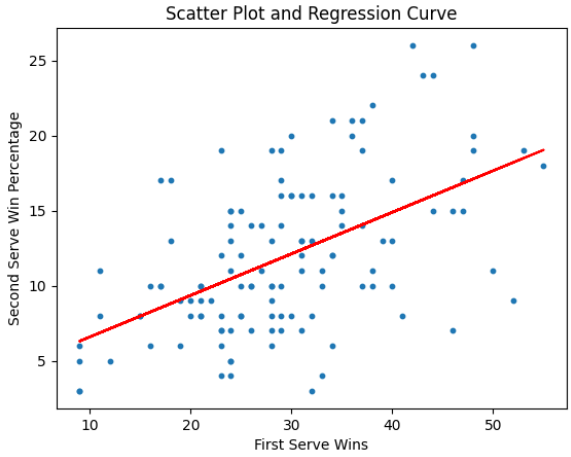

Fig. 6 An average relation between the first-serve and second-serve win percentage.

From the plot, we can see that the second-serve win percentage is very low compared to first-serve wins. This shows that failing to serve initially sways the player's confidence leading to them losing that rally. Though the value for second-serve wins is low for all players, the second-serve win percentage has increased for those having higher first-serve wins. These players have more experience in the game and have mastered their serves. This means that experienced players can counter their feelings and handle failure differently, and their experience has increased their ability to handle failure.

E. The type of ground may affect the overall outcome of the game, but it has very little effect on the faults committed during a service in a tennis match. Support this statement using the dataset.

We can plot the average number of Double Faults committed in different tournaments:

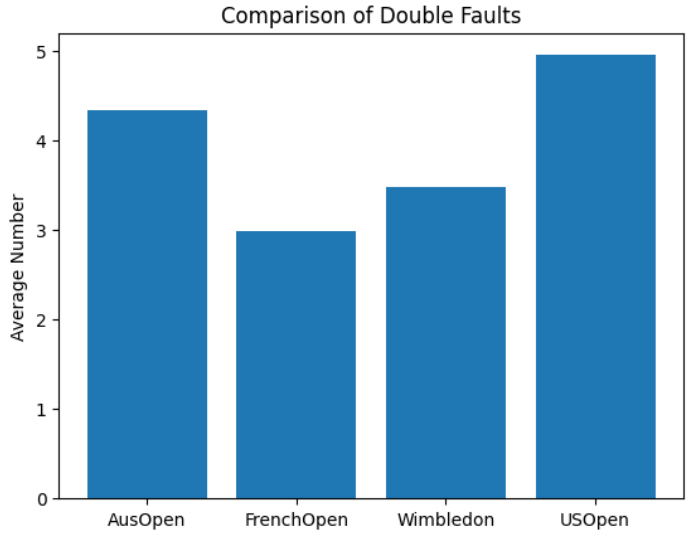

Fig. 7 Average number of double faults committed in different tournaments.

Observing the graph, we can interpret that there is only a small difference in the number of double faults committed across different tournaments. This supports the fact that double faults occur only due to the players' mistakes and not due to the type of ground.

F. What are the chances that an attempt to win a point in the match succeeds? Use the given dataset to find the above statistically.

We can plot a graph between the Net points attempted by players and those won by them in a particular tournament.

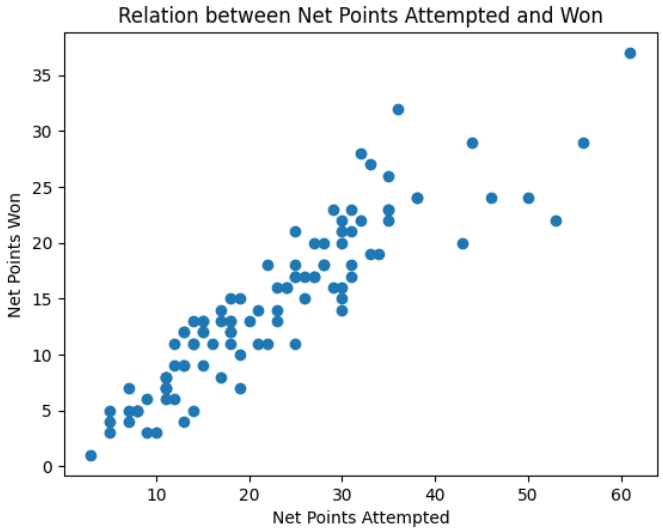

Fig. 8 Scatter plot between Net points attempted and net points won.

The graph is linear with almost a slope of ½. This means the average success rate when attempting points is only 50%. This means that attempting for a point is mostly a play of chances, and one could end up either way. Thus, this also proves that putting an effort to keep attempting for points is favourable over attempting carefully but only a few times.

G. How can we show that in a tournament, most of the matches are one-sided? What can be the reason for this?

We have plotted histograms of the final number of games won by each player in a certain match.

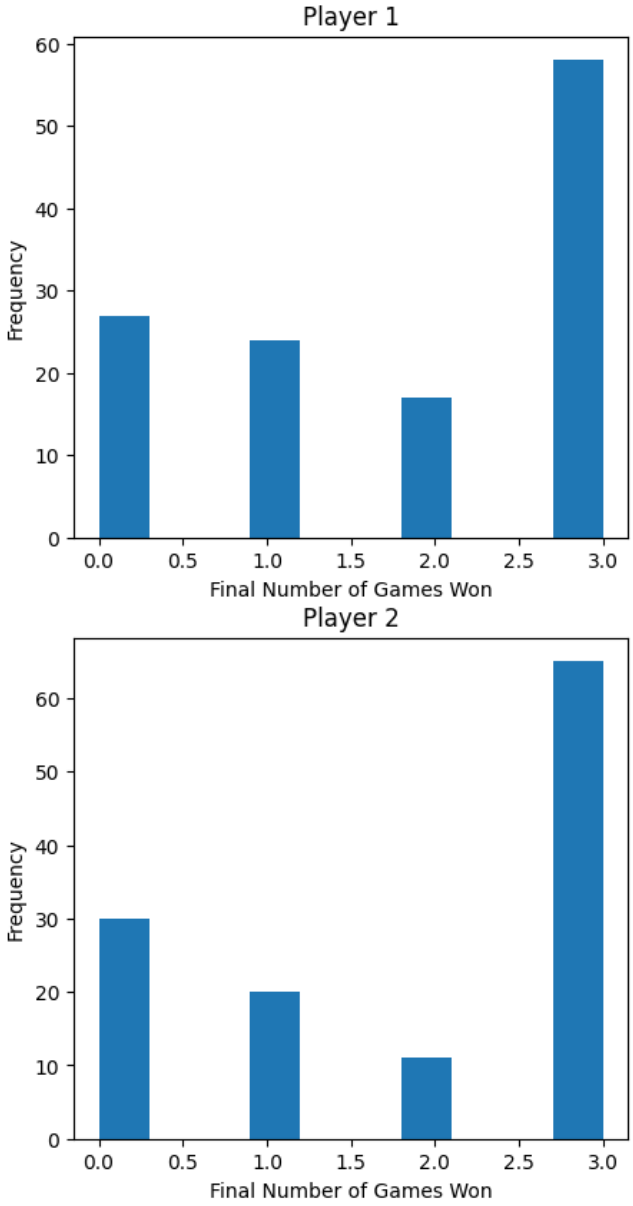

Fig. 9 Histogram of final number of games won by players in matches across the tournament.

We can see that the trend for the number of games won is the same for both players. The trend shows that most of the time, the match settles at one winning three games while the other not winning any. This shows that in a tournament, most matches are one-sided. This can be because, in tournaments, the number of matches for the initial rounds is the highest, and to keep the final rounds more interesting, the strong players are not paired up in the initial rounds.

H. Find all the possible ways in which a match ends in a tournament. Compare the result with our proconceptions.

Here, we have plotted the number of games won by player 1 vs those won by the other player.

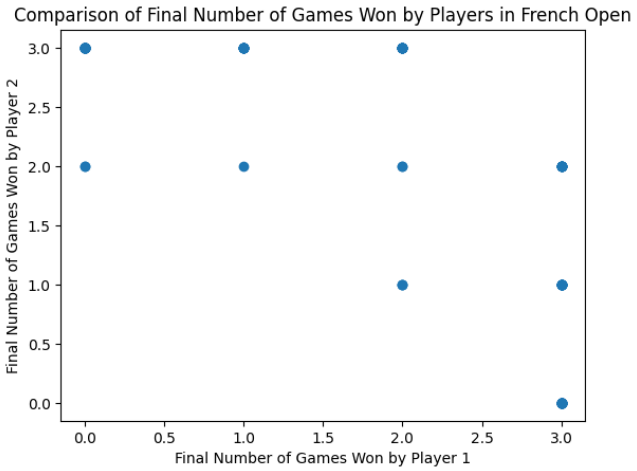

Fig. 10 Graph showing the number of games won by players 1 and 2 in any match of the tournament.

From the graph, we observe that a match has the following possible outcomes (no. of games won by player 1, no. of games won by player 2): (0,2), (0,3), (1,2), (1,3), (2,1), (2,2), (2,3), (3,0), (3,1), (3,2). We might observe that a few other cases that seem to be possible are not present in this dataset i.e. (2,1) and (3,3). We think that if (2,2) appears in real data, then (3,3) should also be possible. But it isn't because a player wins the match once anyone has won 3 games. Another preconception we have is that all the cases that are possible should occur in real data also. But in our data, the case of (2,1) doesn't happen. This is because there are not enough matches to make every case happen yet.

V. Summary of the Observations

From the analysis of the Tennis tournament dataset, several interesting observations were made. First, the First Serve Percentage (FSP) appears to be a positive contributing factor to the match outcome, with matches won having higher values of FSP. Secondly, as the rounds progress, the number of unforced errors and double faults committed by the players increases. This suggests that the pressure of the more important rounds leads to decreased performance and an increase in mistakes.

Furthermore, when comparing the behaviour of men and women in the tournament, it was observed that women tend to perform better under pressure compared to men. While both men and women had similar numbers of errors and faults, women had fewer faults in the semifinals and more in the final round, while men had fewer faults in the final round.

The dataset also revealed that experienced players can handle failure better and have a higher second-serve win percentage. Additionally, the number of double faults committed across different tournaments remains consistent, indicating that double faults occur mainly due to the players' mistakes and not the type of ground. Finally, most matches in the tournament were one-sided, with one player winning three games while the other did not win any. Overall, these observations provide insights into the behaviour and performance of players in tennis tournaments.

VI. Acknowledgment

I would like to express my hearty gratitude to our course instructor, Prof. Shanmuga, for allowing me to work on this project and to have this wonderful experience. I also thank all the Teaching Assistants for their guidance across this project and for helping me grasp all the learnings.

Aryan Sahu
Roll no. 22110038

VII. References

This narrative has used the following references to gain knowledge of using the libraries and functions:

1. Pandas Documentation-

   https://pandas.pydata.org/docs/

2. Stack Overflow-

   https://stackoverflow.com/

3. Geeks for Geeks-

   https://www.geeksforgeeks.org/

4. McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56)

5. J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.