Data Narrative 3

Hitesh Kumar
Roll No.- 22110098
Branch - Computer Science and Engineering
Professor- Shanmuga Nathan Raman, IITGN

Aim— This report aims to explore the dataset of Tennis Major Tournaments, analyze it, ask questions or develop hypotheses, find the answers, or try to prove or disprove them. This is for developing a better understanding of analyzing data and getting familiar with the python libraries like numpy, pandas, matplotlib, seaborn, etc.

Overview— This dataset contains match statistics of Tennis Major Tournament Matches from the 2011 to 2016 seasons. The data includes various attributes such as Player Names, the results of the match (win or loss), the number of Aces, Double Faults, First and Second Serve Percentages, Break Points created and converted, the total number of points won, etc. This dataset can be used for various analytical purposes, such as predicting the winner of a tennis match, identifying patterns and trends in players' performances, and analyzing the relationship between various match statistics and their impact on the match's outcome.

I. Theory

A. Python

Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including structured, object-oriented, and functional programming [1].

II. Libraries

A. Numpy

It is a software library primarily used to make and analyse arrays and plots and perform different operations.

B. Pandas

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

C. Matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

- Create publication-quality plots
- Make interactive figures that can zoom, pan, update
- Customise visual style and layout
- Export to many file formats
- Embed in JupyterLab and Graphical User Interfaces.
- Use a rich array of third-party packages built on Matplotlib

D. Seaborn

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics [2].

III. Scientific Questions and Hypotheses

A. Compare how player 1 behaves on his increasing set score and his opponent's increasing set score.

In this particular question, firstly, we observed the set score of each player for all matches. Then we grouped data according to the average set results of each player. We found the number of Aces scored by the players and plotted the graph. We have found that when the average set result of player 1 increases, his average number of Aces also increases, implying his confidence. But, when the average set result of player 2 increases, player 1 feels less confident, and his average number of Aces remains approximately the same. You can observe the same in Fig. 1

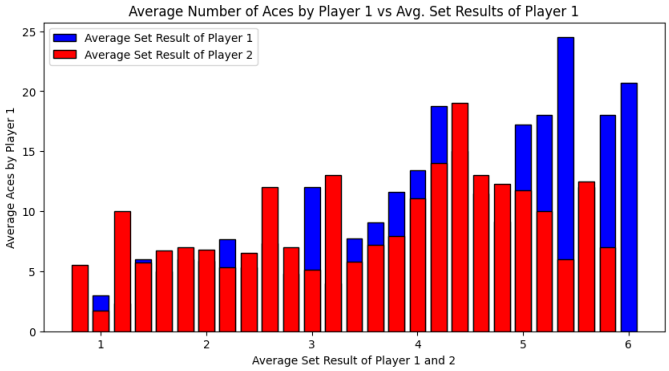

Fig. 1.

B. What is the efficiency of the player having the maximum total points in converting the break-point to win?

First, we found the maximum of total points for both players along the column. Secondly, we observed the player for which the maximum of total points is greater. Then lastly, we calculated the conversion rate for the player having the maximum total points. And we found that," The conversion rate or efficiency of Madison Keys in converting the breakpoint

into a win is high, i.e., 76.92%, which is why she was able to score the maximum total points in Australia Open(Women)."

C. Do you think that the final number of wins by player 2 depends on the number of first serves won by player 2?

We approached this question by plotting a scatter plot between the final number of wins by player 2 VS the number of first-serve wins by player 2.

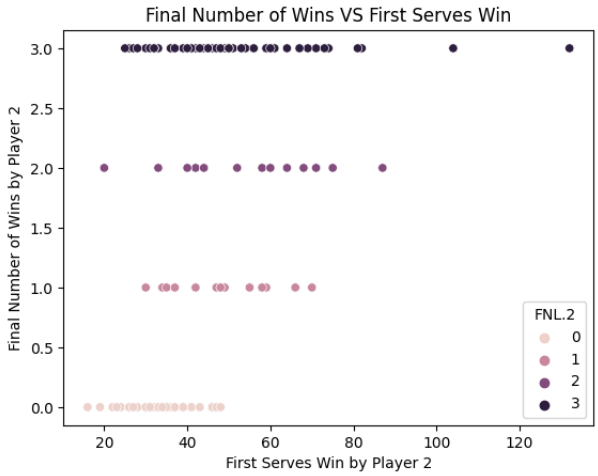

Fig. 2.

After observing the plot as shown in Fig.2, we get to the conclusion that there is some correlation between them. And this can be proven by the fact that the player will feel more powerful and efficient when he will continuously score on his first serves, resulting in more final wins.

D. Find how much a player does the Double Faults usually? Also, calculate the standard deviation of Double Faults done by players 1 & 2.

We calculated the mean error of Double Faults by both players and concluded that a player, on average, does 2 double faults in a match. If women players start to improve and make fewer errors, then there is a possibility that women's tennis will get more popular than men's tennis. Here is the code snippet for this calculation:

```
data = FOw[['DBF.1','DBF.2']]
mean_error = (np.mean(data['DBF.1'])+np.mean(data['DBF.2']))//2
print(f"On an average a player does '{int(mean_error)}' Double Faults in a match.",
      "\nThat implies that there is still a scope of improvement for women to do less errors.",
      "\nThen, may be the popularity of Women Tennis increases because the players are putting efforts for an ideal match.")
X = data.to_numpy()
X.reshape(-1,1)
StdDev = np.std(X)
print(f"\nThe Standard Deviation of Double Faults is turns out to be: {StdDev}")
```

Fig. 3.

E. Find an linear approximation of the Total points won based on the First serve wins by player 1 using Linear Regression and plot the respective graph.

We used the LinearRegression() function from sklearn.linear_model. Using this function we fitted the data and predicted the approximated equation of a line for the data. The respective graph is shown in Fig 4.

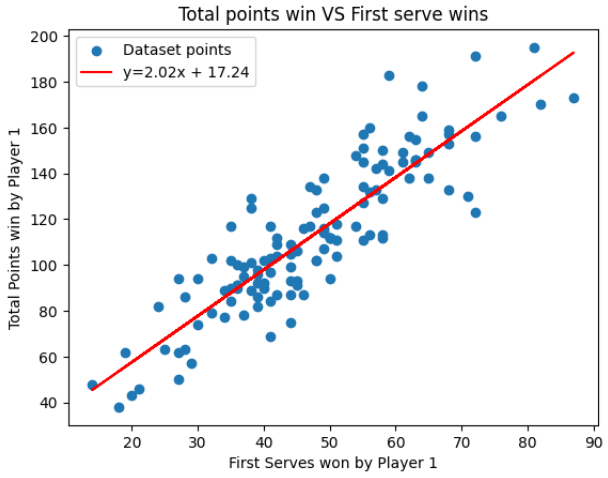

Fig. 4.

F. What is the correlation coefficient of number of Aces scored and number of final games won by player 2 and how does it influence the game of player 1?

We want to find out if there is some correlation between the number of Aces scored, and the number of games won. But, after some calculations to find the correlation coefficient and observing the plotted graph (Fig.5), we concluded that there is no correlation between the number of Aces scored and the number of games won as the coefficient of correlation turned out to be approximately 0.192329.

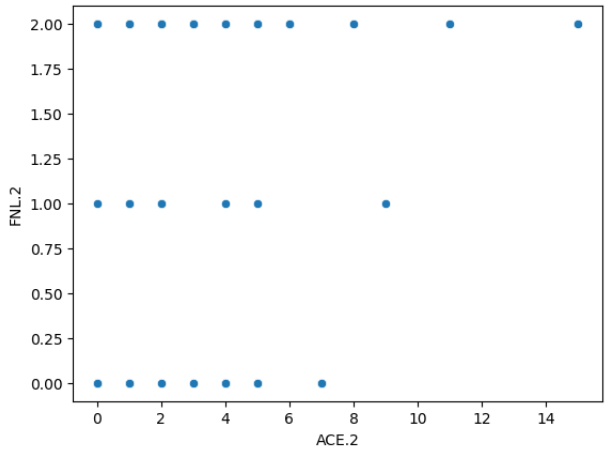

Fig. 5.

G. Find the player which has least unforced errors and then calculate the number of Winner titles he has won?

For this question, firstly, we found the indices for the unforced errors by both players are minimum. On comparing them, we found the index which will have the least value of the number of unforced errors. Then we found that the least value corresponds to which player we calculated the number of winner titles for him. On observing them, we get to know

that the player named R. Federer has the least unforced errors, and that is why he won about 32 games.

H. What is the accuracy percentage of winning net points by player 1? Also plot the percentage graph for player 1?

We plotted a graph as shown in Fig 6 using the stripplot() function of seaborn and plotted the percentage of accuracy in winning net points by player 1.

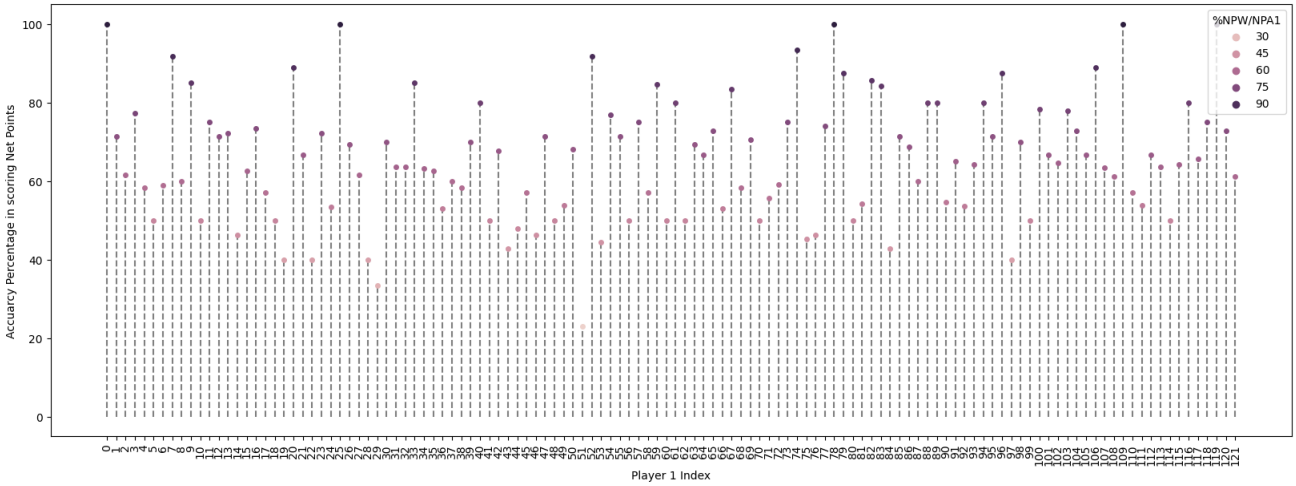

Fig. 6.

On observing Fig. 6, we can see that some peaks correspond to some players with higher accuracy in winning net points than others. This small difference sometimes becomes the win or lose in a match. So the other players should focus more on their accuracy if they want to win upcoming tournaments.

ACKNOWLEDGMENT

I heartily thank my teachers and batchmates who helped me understand some tennis terminology to analyze the data efficiently and how to cite in Chicago style.

REFERENCES

[1] Wikipedia. "Python (programming language)." Wikipedia. Last modified March 30, 2023. Available at: https://en.wikipedia.org/wiki/Python_(programming_language) (Accessed: April 22, 2023).

[2] Waskom, M.L. (2021) Seaborn: Statistical data visualization, Journal of Open Source Software. Available at: https://joss.theoj.org/papers/10.21105/joss.03021 (Accessed: April 22, 2023).