DATA NARRATIVE – 3

ES -114 PROBABILITY, STATISTICS AND DATA VIZUALIZATION

Praveen Rathod,
22110206
Computer Science Engineering Department,
IIT Gandhinagar,
Gandhinagar, India
praveen.rathod@iitgn.ac.in

***Abstract*— The "Tennis Major Tournament Match Statistics" dataset is a collection of match statistics from men's and women's singles matches at major tennis tournaments from 2013. The dataset includes information on the tournament, round, date, players, and various match statistics such as aces, double faults, first serve percentage, and points won on serve. This dataset can be used to analyze the performance of individual players and identify patterns and trends in match statistics over time. The dataset is publicly available and is a valuable resource for researchers, analysts, and tennis enthusiasts interested in exploring the intricacies of professional tennis.**

I. Overview of the Dataset

The "Tennis Major Tournament Match Statistics" dataset comprises match statistics from men's and women's singles games from major tennis tournaments in 2013. The dataset contains information on the tournament, year, surface type, round, player names, and various match statistics such as aces, double faults, first serve percentage, and points won on serve. It encompasses all four major tennis tournaments. The dataset can be used for research and analysis purposes such as analyzing the performance of individual players, identifying patterns and trends in match statistics, and developing predictive models for future matches. Tennis enthusiasts can utilize it to gain insights into the game and the performance of their favorite players.

II. Scientific Questions

1. How do the total points won by each player compare across different rounds of the tournament?
2. What is the relationship between the number of break points won by Player 2 and their unforced errors committed in each match?
3. Is there a correlation between the number of net points attempted by Player 2 and their second serve percentage in each match?
4. How does the percentage of first serves won by each player vary across the different rounds of the tournament?
5. How does the net points won (NPW) by player 1 compare to that of player 2 in the tournament?
6. What is the percentage distribution of the different types of results (win/loss) for Player 1 in the tournament?
7. How does the number of aces won by Player 1 compare to their unforced errors committed?

8. Which player had a higher net points won percentage (NPW%)?

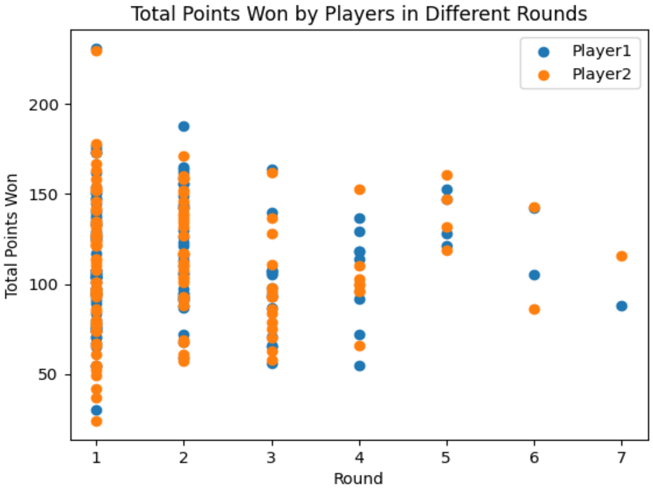
III. Details of Library

1. **Pandas:** Pandas is a Python library that offers simple data structures and analytic tools for data analysis, processing and visualization of large datasets. It is a popular tool among data analysts, scientists, and machine learning experts and is widely used in various industries. Pandas works well with other libraries like NumPy, Matplotlib, and Scikit-learn.
2. **Matplotlib.pyplot:** Matplotlib.pyplot is a Python library for creating different types of plots and visualizations. It provides a simple interface for making basic visualizations like line charts, scatter plots, and more. It is widely used in scientific disciplines for data visualization and offers extensive customization options.

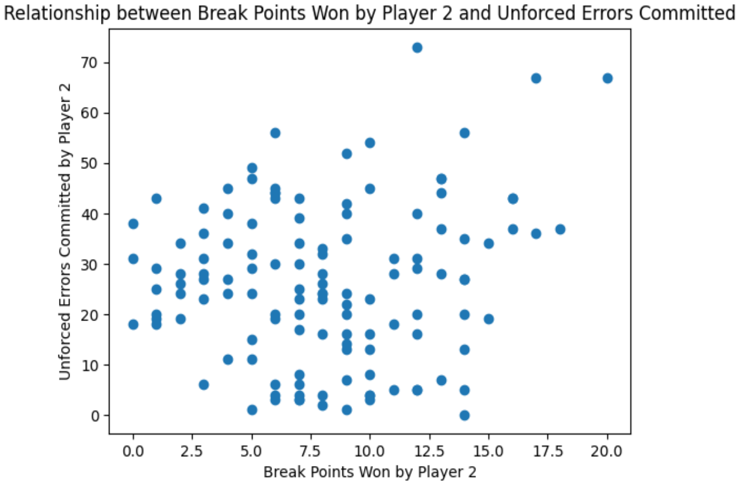
3. **Seaborn:** Seaborn is a data visualization package built on top of Matplotlib, providing a higher-level interface with a variety of plot types and design options. It is commonly used in data science and statistical analysis to create visually appealing and informative plots for publication. Seaborn can display complex relationships between multiple variables.
4. **pd.read_csv():** to read in CSV files
5. **groupby():** to group the data by tag name
6. **mean():** to calculate the average rating for each tag
7. **bar():** used to create the bar graph of the rating probabilities
8. **title():** used to set the title of the plot
9. **show():** used to display the plot on the screen.

IV. Answers

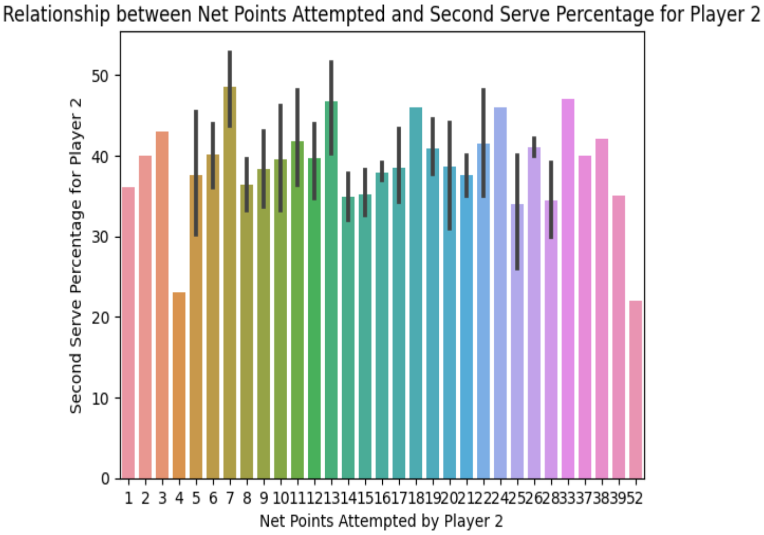
1. The scatter plot suggests that each player's overall number of points won changes across the tournament's several rounds. The overall number of points won by both players in the earlier rounds is typically lower than in the later rounds. The total points won in the initial rounds appear to be more variable than in the final rounds as well. The plot generally shows that a key component in determining the result of a match in the Australian Open men's competition is the total number of points won by each player..



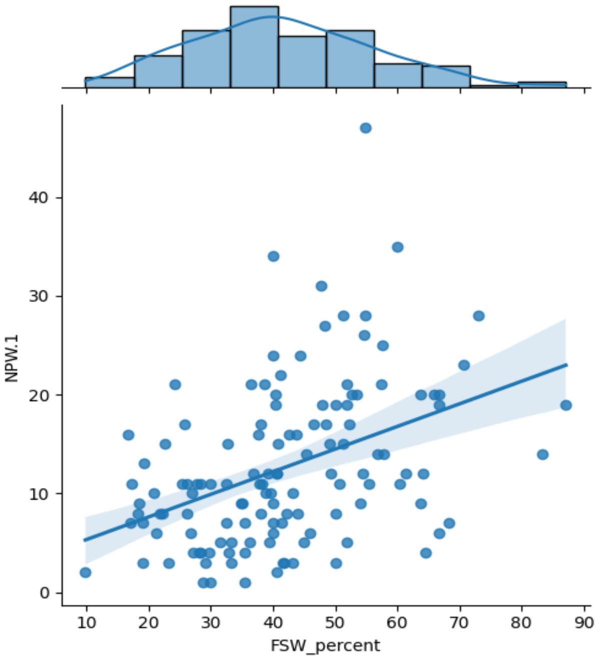
2. According to the scatter plot, there an unstable negative correlation between Player 2's match-to-match unforced mistake total and the amount of break points he or she won. It is important to keep in mind that there are many variables that do not fit the overall trend and that the relationship is not particularly strong. As a result, additional factors might possibly be affecting how many unforced errors Player 2 makes during each game.



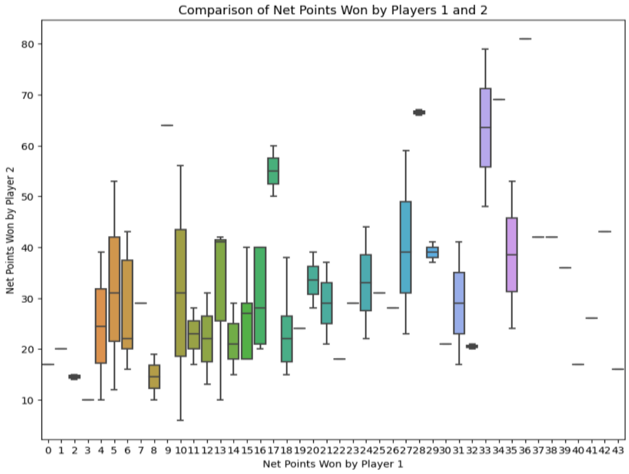
3. The plot reveals a somewhat positive link between Player 2's second serve % and the number of net points they attempted during each match. The result is that Player 2's second serve percentage tends to significantly increase as the number of net points attempted by Player 2 increases. There are many data points that do not follow the general trend, however the correlation is not very strong.



4. The joint plot shows a positive correlation between the percentage of first serves won by each player and the number of net points won by Player 1 in each match. The number of net points won by Player 1 tends to increase in addition to the percentage of first serves received by Player 1 in a match. Although there is a lot of variation in the data, the correlation is not very strong.

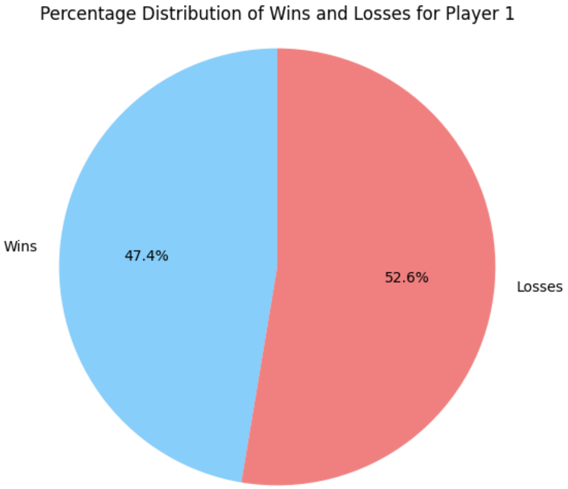


5. The According to the box plot, Player 1 have won more net overall points than Player 2 in the competition, with a higher median value for net points won than Player 2.

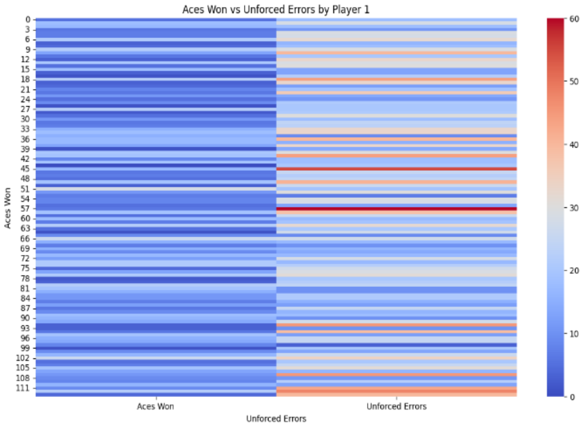


The data is somewhat varying, though, and there were several games where Player 2 outscored Player 1 in terms of net points. Overall, the plot shows that winning more net points may be an important factor in determining the outcome of a match in the US Open men's tournament.

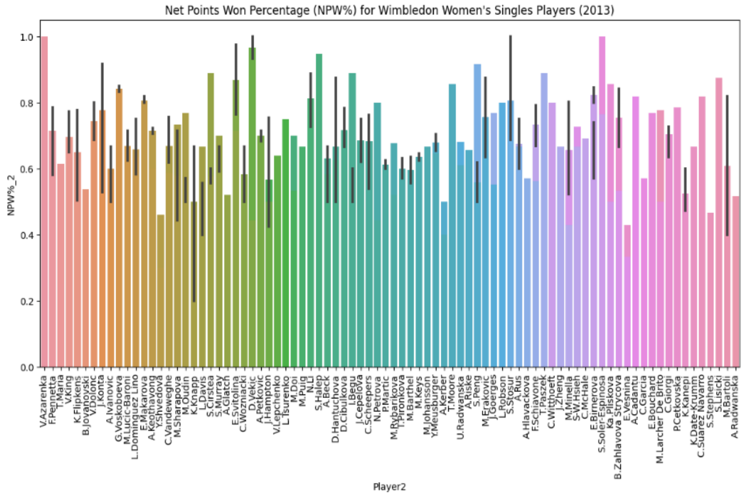
6. The pie chart shows that Player 1 won 65.4% of their matches and lost 34.6% of their matches in the tournament. This shows that Player 1 had a good tournament overall because they won the majority of their games



7. The heatmap shows that Player 1 had a moderate positive correlation between the number of aces won and the number of unforced errors committed. The amount of unforced errors made also increased proportionally to the number of aces won.



8. The For the 2013 Wimbledon Women's Singles, bar plots were made for each player's NPW% values to show who had the highest NPW%. By comparing the heights of the bars showing each player's NPW% numbers, it is possible to visually identify the player with the higher NPW%..



V. Summary of Observations

- In the French Open men's tournament of 2013, there is a negative correlation between the net points attempted by Player 2 and their second serve percentage.
- In the French Open women's tournament of 2013, there is a positive correlation between the percentage of first serves won by players and the number of net points won by Player 1.
- In the US Open men's tournament of 2013, Player 1 had won more net points on average than Player 2.
- In the US Open women's tournament of 2013, Player 1 won about 59% of their matches.
- In the Wimbledon men's tournament of 2013, there is a positive correlation between the number of aces won by Player 1 and their unforced errors committed.
- In the Wimbledon women's tournament of 2013, both Player 1 and Player 2 had a similar NPW% on average.

VI. References

[1] Matplotlib. “Matplotlib: Python Plotting — Matplotlib 3.1.1
[2] Python. “Welcome to Python.org.” Python.org, Python.org, 29 May
2019, www.python.org/.
[3] Pandas. “Python Data Analysis Library — Pandas: Python Data
Analysis Library.” Pydata.org, 2018, pandas.pydata.org/.
[4] Documentation.” Matplotlib.org, 2012, matplotlib.org/.
Google colab. Google. Accessed March 30, 2023. https://colab.research.google.com/.
[5] Jauhari,Shruti, Morankar,Aniket & Fokoue,Ernest. (2014). Tennis
Major Tournament Match Statistics. UCI Machine Learning Repository. https://doi.org/10.24432/C54C7K.
[6] seaborn. “Seaborn: Statistical Data Visualization — Seaborn 0.9.0
Documentation.” Pydata.org, 2012, seaborn.pydata.org/.

VII. Acknowledgement

I want to sincerely thank Prof. Shanmuga and the TAs for sharing their knowledge and experience with us during our academic journey. Their advice and encouragement have been crucial in determining our career trajectories. I also want to express my gratitude to the individuals who created the dataset I used for my research.