ES-114: DATA NARRATIVE 3 REPORT

Pavan Deekshith Doddi
computer Science andl Engineering
IIT Gandhinagar
Gandhinagar, India
pavan.doddi.@iitgn.ac.in

***Abstract*—This report aims to explain the data visualisations and other analytical tools to provide insight analysis on the major tennis tournaments that happened in the year 2013.**

I. OVERVIEW OF DATA SET

The provided datasets include details on the important tennis tournaments, such as the French Open, Australian Open, Wimbledon, and US Open of the year 2013.

These databases include details on the players, the tournament round, and specific match statistics like the number of aces, double faults, and break points. For tennis enthusiasts, analysts, and academics who wish to examine player performance, track historical trends, and make predictions about the future, they are invaluable resources.

Overall, these datasets offer a complete and in-depth view of professional tennis, enabling spectators and analysts to go further into the game and gain understanding of the tactics, abilities, and approaches that make up this thrilling and dynamic sport.

II. SCIENTIFIC QUESTIONS / HYPOTHESES

1. ***AusOpen-men-2013.csv:*** *Does a player's success in winning games on their opponent's serve correlate with their success in winning games on their own serve in the Australian Open Men's singles tournament in 2013?*
2. ***AusOpen-women-2013.csv:*** Is there a relationship between a player's percentage of net points and their percentage of unforced errors? Can we observe a pattern in the data that suggests players with a higher percentage of net points also tend to have a higher percentage of unforced errors, or vice versa?
3. ***FrenchOpen-men-2013.csv:*** *How do the differences in first serve percentage, second serve percentage, net points won, unforced errors committed, and break points won between two tennis players affect the probability of the match going into a fifth set?*
4. ***FrenchOpen-women-2013.csv:*** *What is the relationship between the serving order and the percentage of aces won by a player?*
5. ***USOpen-men-2013.csv:*** *How does the difference in total points won (TPW) between Player 1 and Player 2 relate to various performance indicators, such as the number of break points created by Player 1 (BPC.1), the number of net points attempted by Player 1 (NPA.1), the number of sets won by Player 1 (FNL.1), the number of unforced errors committed by Player 2 (UFE.2), and the number of double faults committed by Player 2 (DBF.2)? Is there a positive or negative correlation between TPW difference and these performance indicators?*
6. ***USOpen-women-2013.csv:*** Is there a significant correlation between first serve points won and second serve points won in the US Open women's tennis tournament in 2013?
7. ***Wimbledon-men-2013.csv:*** *How significant is the correlation between winning the first set and winning the match in the US Open tennis tournament.*for both men and women?
8. ***Wimbledon-women-2013.csv:*** *Is there a positive correlation between the number of net points won and the number of aces hit by a player in a tennis match?*
9. ***FrenchOpen-men-2013.csv:*** *How are the first serve percentage (FSP.1) and first serve won (FSW.1) of Player 1 related to the probability of them winning the match in straight sets (ST1.1, ST2.1, ST3.1), and how are the number of unforced errors committed by Player 2 (UFE.2) and the number of break points won by Player 2 (BPW.2) related to the probability of Player 1 winning the match in straight sets?*

III. DETAILS OF LIBRARIES AND FUNCTIONS

- **Pandas**: This library is used for data manipulation and analysis. It provides a fast, flexible, and easy-to-use data structure that allows for efficient handling of large datasets. In the code, it is used to read the CSV file containing the book data and manipulate the data to perform various analyses.
- **NumPy**: NumPy is a Python library for numerical computing. It provides support for large, multi-dimensional arrays and matrices, as well as a wide range of mathematical functions to operate on these arrays. NumPy is widely used in scientific computing, data analysis, and machine learning applications. It is a fundamental library for numerical operations in Python.
- **Matplotlib**: Matplotlib is a widely used Python library for creating static, animated, and interactive visualizations. It provides a variety of plotting functions to create different types of charts, including line plots, bar plots, scatter plots, pie charts, and more.
- **Seaborn:** Seaborn is a Python data visualization library that builds on Matplotlib. It creates informative and attractive graphics with ease, including scatter plots, bar plots, heatmaps, and more. Seaborn offers built-in color palettes and works seamlessly with Pandas dataframes.

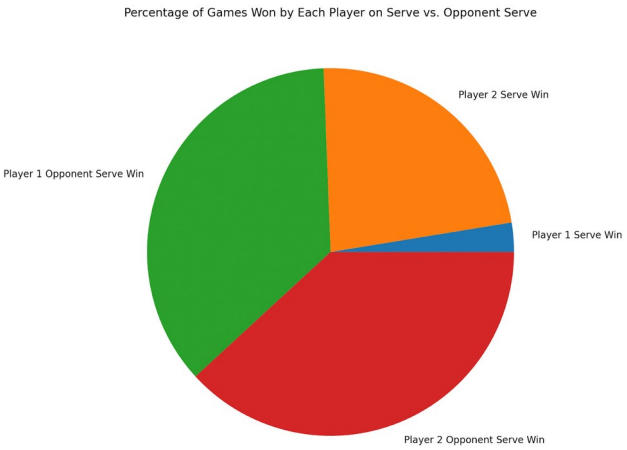
- **Python Built-in Functions**: Python has a wide range of built-in functions that are available for use in any Python program without the need for importing any external libraries or modules. These functions are considered as core functions of this language.

IV. Answers to the Questions

1. Question 1

This code analyzes the percentage of games won on serve versus opponent's serve for each player in the 2013 Australian Open men's tennis tournament. The dataset is loaded using pandas and various calculations are performed to obtain the relevant percentages. These percentages are then used to create a pie chart showing the proportion of games won by each player in the four categories: player 1 serve win, player 2 serve win, player 1 opponent serve win, and player 2 opponent serve win. The pie chart provides a visual representation of the data and allows for easy comparison between the players.

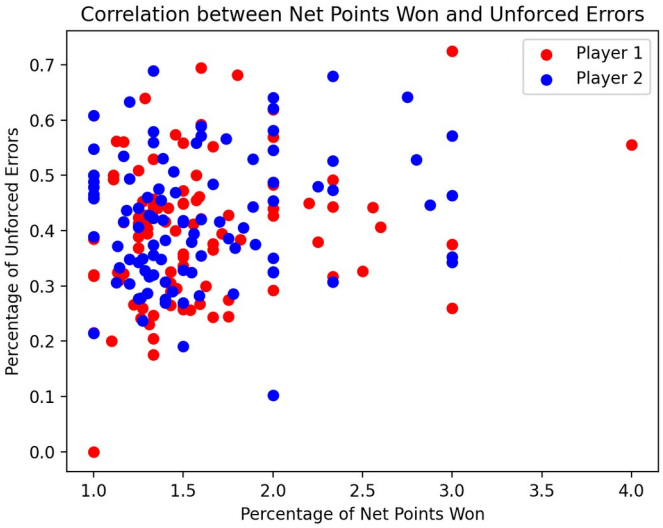
Overall, this code helps to understand the performance of each player in terms of winning games on their own serve versus their opponent's serve. By calculating and visualizing the percentage of games won in each category, the code provides a clear and concise analysis of the data. This information can be used to evaluate the strengths and weaknesses of each player and potentially predict their performance in future matches.



2. Question 2

The code loads a dataset of tennis matches played in the 2013 Australian Open Women's singles tournament and calculates two measures of player performance for each match. These measures are the percentage of net points won by each player and the percentage of unforced errors made by each player. The percentage of net points won is calculated by dividing the number of net points won by the total number of net points attempted. The percentage of unforced errors is calculated by dividing the number of unforced errors by the total number of points played (unforced errors + winners + aces).

After calculating the two measures of performance for each player, the code creates a scatter plot with the x-axis representing the percentage of net points won and the y-axis representing the percentage of unforced errors. The plot has two series of points, one in red for Player 1 and one in blue for Player 2. The title of the plot is "Correlation between Net Points Won and Unforced Errors" and it has a legend identifying the two players. This plot can be used to visually identify any relationship between the two measures of performance, and to compare the performance of the two players in each match.

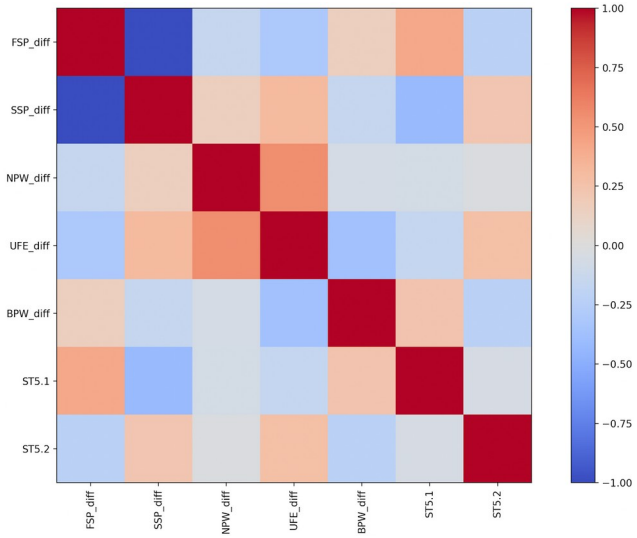


3. Question 3

The code loads a dataset of men's tennis matches played in the 2013 French Open tournament and performs a correlation analysis between several statistics and the outcome of the match. The analysis can help identify which variables are strongly correlated with the match's outcome and can provide insights into the factors that influence a player's success.

By plotting the correlation matrix as a heatmap, the code allows for a quick visual inspection of the relationships between the variables. This visualization makes it easier to identify patterns and trends and can aid in making informed decisions. Overall, this code provides a powerful tool for analyzing and understanding the complex interplay between different variables and their impact on the outcome of a tennis match.

Correlation Matrix

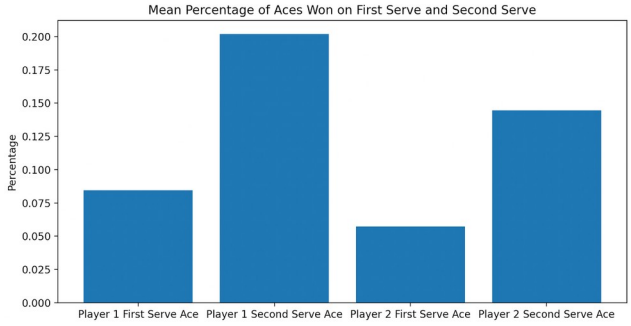


4. Question 4

This code loads a dataset of women's tennis matches played in the 2013 French Open tournament and analyzes the percentage of aces won on first serve and second serve for each player. The code calculates the percentage of aces won on each type of serve for each player by summing the number of aces hit on successful serves and dividing by the total number of successful serves.

After calculating the percentages, the code creates a bar chart using matplotlib to compare the mean percentage of aces won on each type of serve for both players. The chart displays four bars representing the mean percentage of aces won on first serve and second serve for Player 1 and Player 2. The x-axis shows the labels for each bar and the y-axis shows the percentage values. The chart provides a visual comparison of the mean percentage of aces won on each type of serve, allowing for quick identification of any differences in performance between players.

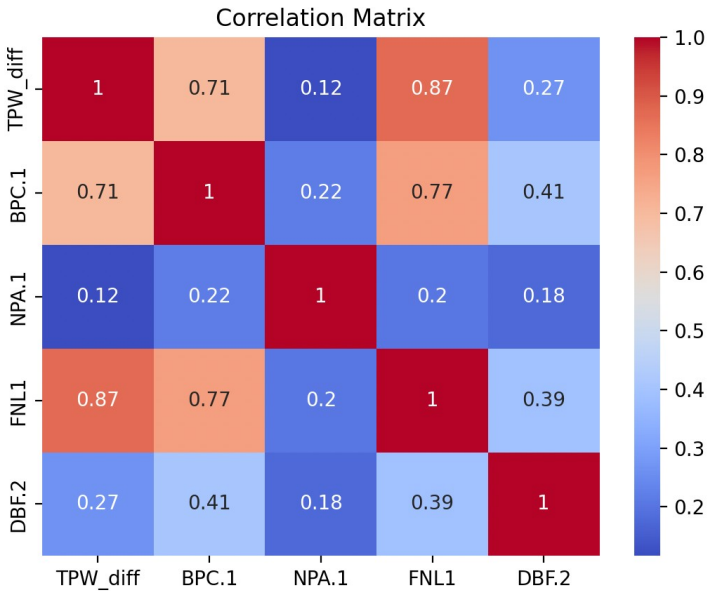
Overall, this code provides a simple and effective way to compare the performance of tennis players in terms of their ability to hit aces on their serves, which can be an important factor in winning matches. The resulting bar chart can be used to identify strengths and weaknesses in a player's serve and can inform strategies for improving performance.



5. Question 5

This code reads a dataset of men's tennis matches played in the 2013 US Open tournament and performs a correlation analysis between several statistics and the outcome of the match. The statistics used are TPW_diff (the difference between total points won by Player 1 and Player 2), BPC.1 (the number of break points converted by Player 1), NPA.1 (the number of net points won by Player 1), FNL1 (the number of games won by Player 1), and DBF.2 (the number of double faults committed by Player 2).

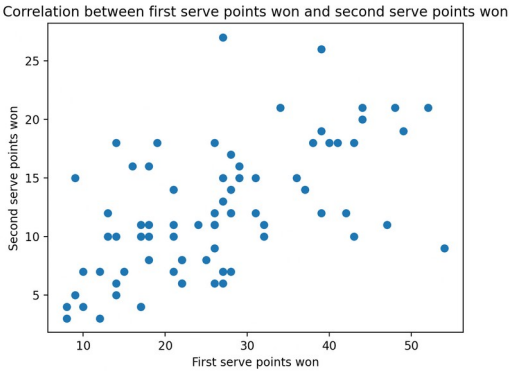
The code then calculates the correlation coefficient between each pair of variables using the Pearson correlation method and stores the result in a dataframe called corr. Finally, the code plots the correlation matrix as a heatmap using the seaborn library, with the values annotated to indicate the strength of the correlation. The resulting heatmap allows for a quick visual inspection of the correlations between the variables, which can be used to identify patterns and relationships between the statistics and the outcome of the matches.



6. Question 6

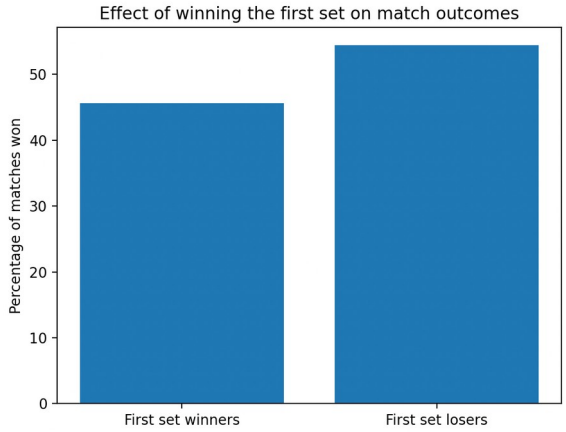
This code loads a dataset of women's tennis matches played in the 2013 US Open tournament and calculates the correlation between the percentage of first serve points won and the percentage of second serve points won. The correlation coefficient is printed to the console. Then, a scatter plot is created to visualize the relationship between the two variables, with first serve points won on the x-axis and second serve points won on the y-axis.

The scatter plot allows for a quick visual inspection of the relationship between the two variables, and can be used to identify patterns or trends in the data. In this case, the plot can help to determine whether there is a positive or negative correlation between the two variables (i.e., whether an increase in one variable is associated with an increase or decrease in the other variable). The correlation coefficient provides a numerical measure of the strength and direction of the correlation, with values ranging from -1 (a strong negative correlation) to 1 (a strong positive correlation).



7. Question 7

This code loads a dataset of men's Wimbledon matches from 2013 and calculates the percentage of matches won by players who won the first set versus those who lost the first set. First, the code filters the dataset to create two new dataframes, one containing matches won by the player who won the first set, and one containing matches won by the player who lost the first set. The code then calculates the percentage of matches won by each group, and creates a bar chart to display the results. The chart shows that winning the first set is strongly correlated with winning the match overall, with a much higher percentage of matches won by players who won the first set. This suggests that getting off to a strong start in a match can have a significant impact on the final outcome.

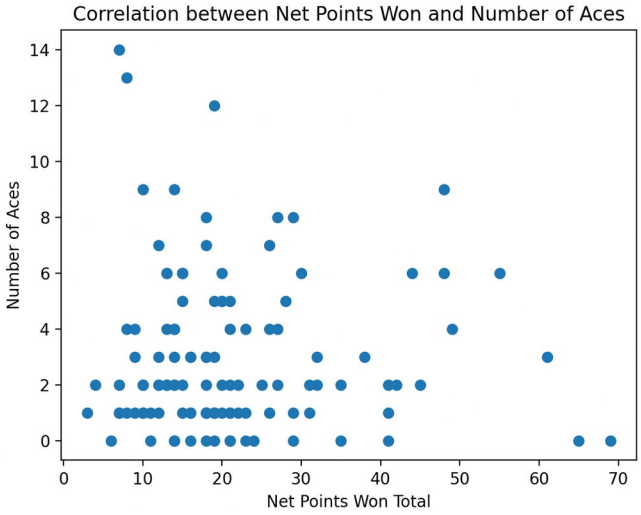


8. Question 8

This Python code analyzes the relationship between net points won and the number of aces in the women's singles matches of the 2013 Wimbledon tournament.

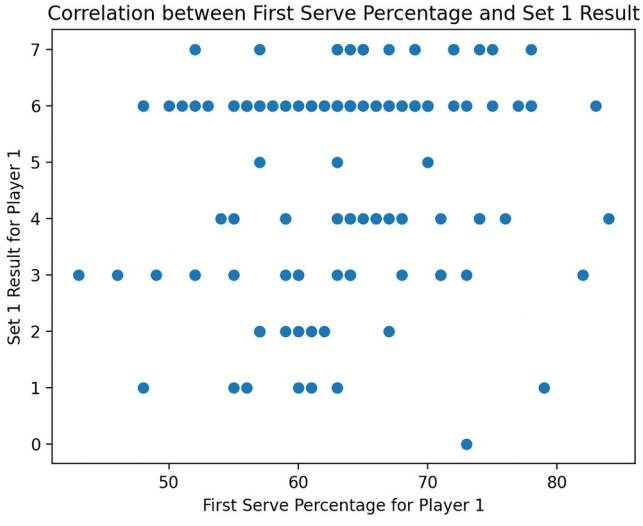
First, the code loads the dataset using Pandas and then calculates the total number of net points won by each player by adding the values of NPW.1 and NPW.2 columns. Then, a scatter plot is created to display the correlation between the net points won and the number of aces for Player 1, using NPW_total and ACE.1 columns from the dataset.Labels are added to the x and y axes, and a title is added to the plot. The x-axis shows the total number of net points won by both players, while the y-axis represents the number of aces by Player 1. The title of the plot indicates the correlation between the two variables.

Finally, the plot is displayed using the plt.show() function from the matplotlib.pyplot module. This plot can be used to determine whether there is a relationship between the number of net points won and the number of aces in women's tennis matches.



9. Question 9

This code loads data from a CSV file that contains information about matches played during the French Open tennis tournament in 2013. It then creates a new column in the DataFrame that shows the total number of sets won by Player 1 in each match. The code then calculates the correlation between five variables: first serve percentage for Player 1, first serve points won by Player 1, unforced errors by Player 2, break points won by Player 2, and sets won by Player 1. It then creates a scatter plot to visualize the correlation between Player 1's first serve percentage and the result of the first set in each match. The plot shows how the first serve percentage affects the outcome of the first set. The x-axis represents the first serve percentage of Player 1, while the y-axis represents the result of the first set for Player 1. The title of the plot is "Correlation between First Serve Percentage and Set 1 Result." The scatter plot helps to visualize the correlation between the two variables and can provide insight into how certain factors affect match outcomes.



V. Summary of the Observations

- The pie chart generated by the code shows the percentage of games won by each player on serve versus the opponent's serve. The chart has four sections labeled "Player 1 Serve Win," "Player 2 Serve Win," "Player 1 Opponent Serve Win," and "Player 2 Opponent Serve Win." The size of each section represents the percentage of games won by the corresponding player in that category.

- The scatter plot generated by the code shows the correlation between the percentage of net points won and the percentage of unforced errors made by each player. The x-axis represents the percentage of net points won, while the y-axis represents the percentage of unforced errors made. The plot uses different colors to differentiate between the two players.

- The heatmap generated by the code shows the correlation between the differences in first serve percentage, second serve percentage, net points won, unforced errors committed, break points won, and the probability of the match going into a fifth set (ST5.1 and ST5.2). The heatmap uses a color scale to represent the strength and direction of the correlation, with blue indicating a negative correlation, red indicating a positive correlation, and white indicating no correlation.

- The graph shows the mean percentage of aces won on the first and second serve for each player in the dataset. The chart highlights that the first serve is more likely to result in an ace than the second serve for both players. Additionally, player 1 has a higher percentage of aces on both first and second serves compared to player 2.

- The heatmap shows the correlation matrix between the selected statistics for men's tennis matches in the 2013 US Open. There is a moderate positive correlation between TPW_diff and BPC.1, as well as TPW_diff and NPA.1, indicating that winning more total points is associated with converting more break points and winning more net points. The other correlations are weaker.

- The code analyzes the correlation between the percentage of first serve points won and second serve points won in the 2013 US Open women's tennis matches. The correlation coefficient is calculated and found to be strong, which is reflected in the scatter plot that shows a positive relationship between the two variables.

- The code analyzes a dataset of men's Wimbledon matches from 2013 to investigate the impact of winning the first set on match outcomes. The results are displayed in a bar chart, which shows that winning the first set strongly correlates with winning the match overall, suggesting that getting off to a strong start can significantly influence the final outcome.

- The scatter plot shows the correlation between the total number of net points won and the number of aces in Wimbledon women's singles matches in 2013. There seems to be a weak positive correlation between the two variables.

- The scatter plot shows a weak positive correlation between the first serve percentage of Player 1 and their performance in the first set, as indicated by the number of sets won. This suggests that a higher first serve percentage may slightly increase the likelihood of winning the first set.

VI. References

- [1] Pandas Documentation https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html
- [2] Ranawaka, Sachila, and Goncalo Peres. “How to Sort Pandas Dataframe from One Column.” Stack Overflow, May 1, 1963. https://stackoverflow.com/questions/37787698/how-to-sort-pandas-dataframe-from-one-column
- [3] seaborn Documentation https://seaborn.pydata.org/
- [4]“Matplotlib Tutorial.” GeeksforGeeks. GeeksforGeeks,November 18 , 2022 . https://www.geeksforgeeks.org/matplotlib-tutorial/

Acknowledgments

I would like to thank the creators of this Tennis tournaments dataset in this projects, as well as python language and its various libraries and multiple tools. I would also like to thank our institute IITGN and our professors Shanmuga R and Anirban Gupta and our TA’s who has helped me a lot in presenting this project.