DATA NARRATIVE(ES114)

Aditya Kumar*, 22110015*

*Data Analysis*

*Abstract –* The objective of this study was to perform a data analysis task on a given dataset containing information about the Tennis tournament held in 2013 in 4 countries: Australia, the USA, France, and Wimbledon. The dataset was cleaned, processed and analysed to extract meaningful insights and observations. The analysis includes tasks such as merging datasets, filtering data, grouping, aggregating data, and visualizing data using graphs. The results obtained were used to draw conclusions and make recommendations for future studies.

1. Introduction

The dataset used in this study contains information about the Tennis tournament held in 2013 in 4 countries: Australia, the USA, France, and Wimbledon.

1. Overview of the Dataset

The dataset contains mainly eight files about the Tennis tournament in 2013 in 4 countries: Australia, the USA, France, and Wimbledon.

The Tennis Major Tournament Match Statistics dataset contains information on Tennis matches held in 2013 in 4 different countries. The dataset comprised of Men and Women’s matches with detailed information of statistical data such as the number of rounds, result, first serve percentage and number of Aces won etc. The result of the first player winning the match is 1, and the second player winning the match is 0. Each dataset consists of 7 rounds. This dataset can be used to analyse the performance of players and other factors, such as the cognitive motor skills of players across various countries

III. Scientific Questions/Hypotheses

Questions related to the given dataset “**Tennis Major Tournament Match Statistics”.**

**1.** How does the first-serve percentage (FSP) distribution vary by round?

**2.** Is there a significant difference in the number of double faults per match between the quarterfinals and the semi-finals?

3. Are there significant differences in the percentage of first serves and second serves won between players who won their first-round match and those who lost?

4. Are there significant differences in the number of unforced errors made by players in the quarterfinals compared to the first round? Also, Are there significant differences in the percentage of net points won by players who reach the final compared to those who don't?

5. Is there a correlation between the number of breakpoints won and the number of unforced errors made by a player in a match? Also, Do players who win more points on their second serve also tend to have a higher percentage of aces in the match?

**6.** Do players who win more break points also tend to have a higher percentage of first-serve points won?

7. Correlation between first serve percentage and percentage of break points saved.

8. Correlation between first serve percentage and percentage of break points saved.

**IV. Details of Libraries and Functions**

* Pandas: Used for data manipulation and analysis, providing data structures for efficiently storing and querying large datasets.

Functions used from pandas:

- read\_csv(): Used for reading data from a CSV file and returning a pandas DataFrame.

- groupby(): Used to group the data in the DataFrame based on a specified column(s).

- nunique(): Used to count the number of unique values in a pandas Series or DataFrame.

- reset\_index(): Used to reset the index of a DataFrame.

- merge(): Used to combine two DataFrames based on a specified column(s).

* Matplotlib: Used for data visualization and creating charts and plots.

Functions used from matplotlib:

pyplot.bar(): Used to create a bar plot. pyplot.xticks(): Used to set the x-axis tick labels. pyplot.xlabel(): Used to set the x-axis label. pyplot.ylabel(): Used to set the y-axis label. pyplot.show(): Used to display the plot.

* Numpy: Used for numerical computations and mathematical operations.

Functions used from numpy:

• nanmean(): Used to calculate the mean of a numpy array, ignoring any NaN values.

* Seaborn: Used for data visualization and creating charts and plots.

Functions used from seaborn:

1. sns.scatterplot(x,y) - creates a scatter plot between two variables x and y.
2. sns.histplot(x) – create a histogram of variable x.

**V. Answers to the Questions**

Australia Open Men

Question 1: How does the first-serve percentage (FSP) distribution vary by round?

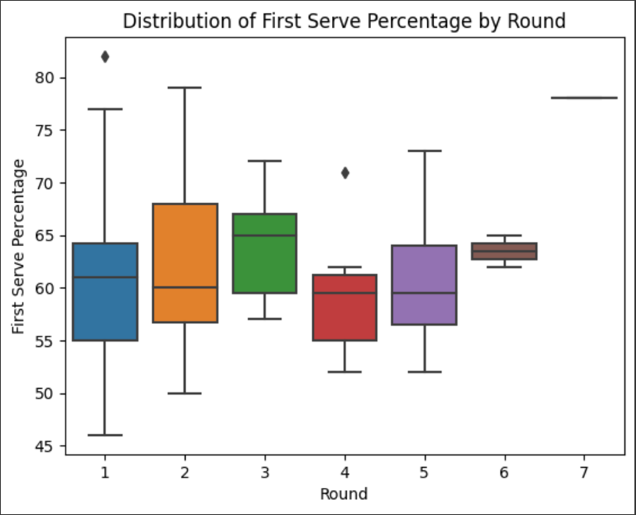


Fig 1. First Serve Percentage vs Round

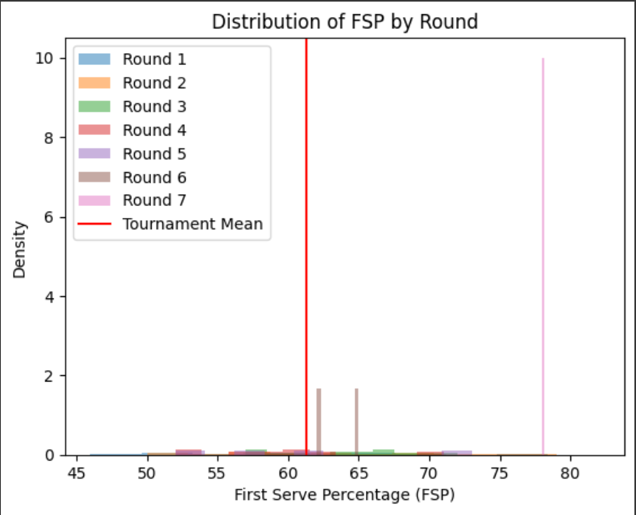
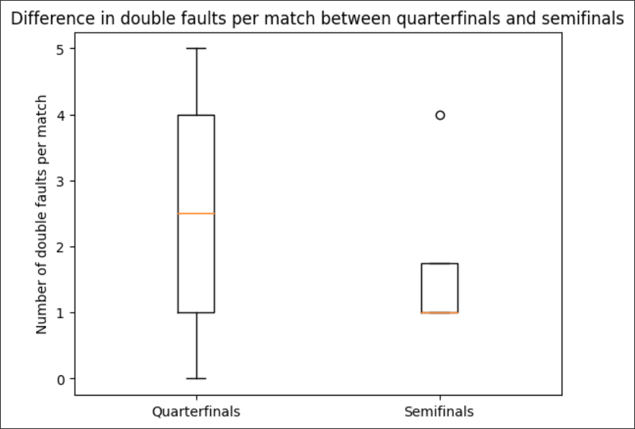


Fig 2. First-serve percentage density for each round

From the above figures and data, we can observe that for the finals the FSP is the maximum and as the playoffs start the FSP declines, which can denote the pressure felt by the players in the playoffs.

Australia Open Women

Question 2: Is there a significant difference in the number of double faults per match between the quarterfinals and the semi-finals?

Fig 3. Number of double faults in Quarterfinals and semi-finals

From the above figure, it can be concluded that the number of double faults decreased drastically from the quarterfinals to the semi-finals. The number of double faults a player makes can be an indicator of their skill level and consistency on the serve.

French Open Men

Question 3: Are there significant differences in the percentage of first serves and second serves won between players who won their first-round match and those who lost?

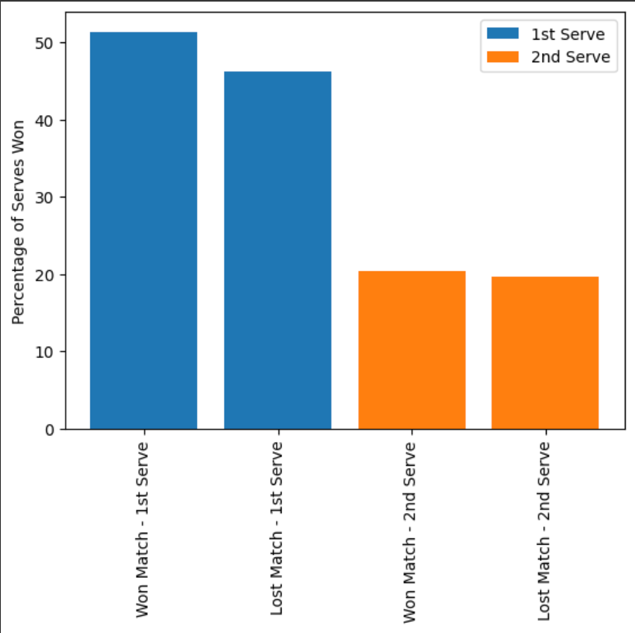


Fig 4. First and Second Serve percentage for the first and second round

The above figures show the mean of winning and losing Game first serve percentage for players. It can be concluded that there is not much difference in the winning and losing first-serve percentages, but still, the losing percentage is lower than the winning percentage. Also, from the data, we cannot predict for sure who will win the game based on first and second-serve percentages alone.

French Open Women

Question 4: Are there significant differences in the number of unforced errors made by players in the quarterfinals compared to the first round? Also, Are there significant differences in the percentage of net points won by players who reach the final compared to those who don't?

There is no significant difference in the number of unforced errors made by players in the quarterfinals compared to the first round.

There is no significant difference in the percentage of net points won by players who reach the final compared to those who don't.

US Open Men

Question 5: Is there a correlation between the number of breakpoints won and the number of unforced errors made by a player in a match? Also, Do players who win more points on their second serve also tend to have a higher percentage of aces in the match?

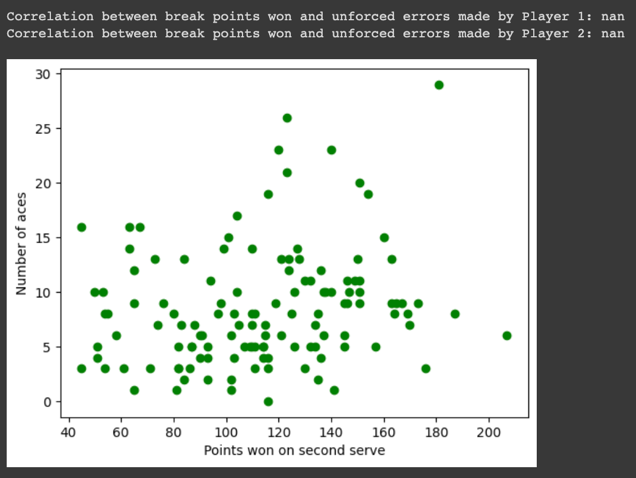


Fig 5. Number of Aces vs points won the second ace

From the above figure, we can see that there is no correlation between the number of breakpoints won and the number of unforced errors made by a player in a match, as the unforced error does not involve the other player. Also, the number of aces increases as second serve points increase, which shows that the confidence of the player increases, and his game gets aggressive if he tends to have a nice serve.

USA Open Women

Question 6: Do players who win more break points also tend to have a higher percentage of first-serve points won?

* Correlation between break points won and percentage of first serve points won by Player 1: 0.043630292576604196
* Correlation between break points won and percentage of first serve points won by Player 2: 0.2522382286104928

From the above data, we can observe that there are not any conclusive results as the variation is very high in both data. So, more data is required to draw some meaningful conclusions about the relation between break points and First-serve percentage.

Wimbledon Open Men

Question 7: Correlation between first serve percentage and percentage of break points saved.

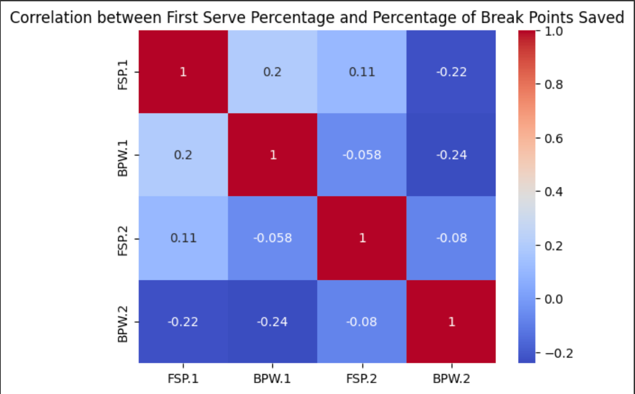


Fig 6. Correlation between First Serve percentage and Percentage of Break Points Saved for both Player 1 and Player 2

From the above figure, it can be concluded that indeed First Serve percentage and Percentage of Break Points Saved are related as higher first serve percentage can lead to a higher percentage of break points saved, as a strong first serve can put the opponent on the defensive and make it more difficult for them to break serve. Additionally, a higher percentage of break points saved can lead to a higher first-serve percentage, as players who are confident in their ability to save break points may be more aggressive on their first serve. Therefore, a positive correlation exists between the first serve percentage and the percentage of break points saved in tennis.

Wimbledon Open Women

Question 8: Correlation between first serve percentage and percentage of break points saved.

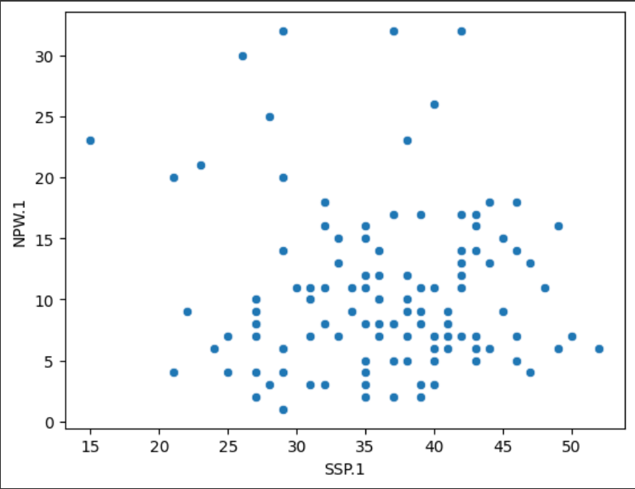


Fig 7. First-serve percentage vs percentage of break points saved

Correlation between second serve percentage and net points won:

-0.050241087830483366

From the above data, the graph is scatter throughout the values of first-serve percentage and Break points percentage, that shows there is a no relation between the first-serve percentage and Break points percentage.

VI. References

* 1. [To plot bar graph from dataframe.](https://stackoverflow.com/questions/70784315/bar-chart-using-dataframe-matplotlib)
  2. [How to sort data in a dataframe](https://www.geeksforgeeks.org/how-to-sort-pandas-dataframe/)
  3. [How to merge two dataframes.](https://pandas.pydata.org/docs/user_guide/merging.html)
  4. [How to take mean of data in dataframe.](https://sparkbyexamples.com/pandas/pandas-get-column-average-mean/" \l ":~:text=To%20get%20column%20average%20or%20mean%20from%20pandas%20DataFrame%20use,values%20for%20the%20requested%20axis.)
  5. [How to group data from a dataframe with some columns of the dataframe.](https://realpython.com/pandas-groupby/)

VII. Acknowledgement

I would like to express my special thanks of gratitude to Prof. Shanmuga for his guidance and support in my Data Narrative analysis by reviewing and helping me improve the questions.