Data Synthesis Narrative

Lavanya (22110130)
Department of Computer Science and Engineering
Indian Institute Of Technology Gandhinagar
Gujarat, India
lavanya.lavanya@iitgn.ac.in

***Abstract*— This report analyzes datasets of tennis matches, providing insights on player performance and strategic considerations.The analysis aims to explore patterns and trends in the data and provide assertions regarding the importance of successful serving and the impact of aces, winners, and unforced errors. The report employs data visualization tools , such as scatterplots, histograms, pie charts and box plots to present the findings in a clear and intuitive manner**.

I. OVERVIEW OF THE DATASET

The tennis major tournament match statistics dataset in the provided link is a comprehensive and structured collection of data from professional tennis matches. The dataset includes detailed information on various match statistics, such as serve percentages, aces, double faults, winners, unforced errors, break points, and net points. Additionally, the dataset provides information on players, their names, and the results of each match. The data is collected from a wide range of major tournaments, including Wimbledon, French Open, Australian Open, and US Open.

This dataset is valuable for a range of analytical purposes, such as identifying patterns and trends in player performance, studying the effectiveness of different strategies, and predicting match outcomes.

.

II. DETAILS OF LIBRARIES AND FUNCTIONS

There are various libraries in python, which were of great help to perform different operations on our Dataset.

Libraries used:

A. Pandas (I): Pandas is a python library used for working with datasets. It has functions for analyzing, cleaning, exploring and manipulating data.

Some of the functions I used from Pandas are:

- Read_csv: It is used to load a CSV file into a dataframe.
- Index: It returns the index information of a dataframe.
- Value_counts: It is used to get a series containing counts of unique values.
- Isin: It checks if the dataframe contains the specified values and can also be used to display or extract information common to two dataframes.
- Sort_values: It sorts the dataframe by the specified column name in either ascending or descending order.
- Len: used to return the number of rows in a dataframe.
- Replace: It is used to replace a string, list, dictionary, series, number from a Pandas Dataframe in Python.
- Astype: It is used to cast a Pandas object to a specified datatype.
- Groupby: It is used for grouping the data according to the categories and apply a function to the categories, which helps to aggregate the data efficiently.
- Head: It returns the first n rows from the specified dataframe.
- tolist(): It is used to convert a series type object into a list.
- Reset_index: It allows to reset the index back to the default or original indexes of a dataframe
- Notna: It detects non missing values for an array-like object.
- Unique: It gives unique values in the specified column name.
- Corr: It returns the correlation factor between the two specified columns.
- Kmeans: It is used for clustering all the data points.

B. Matplotlib (II) : Matplotlib is a visualization library in python for 2D plots of arrays. It is a multi-platform data visualization library built on Numpy arrays

Some of the functions I used from Matplotlib are:

- Plot: It is used to visualize a dataframe by plotting by plotting different columns into a graph.
- Set_xticklabels: It is used to set the xtick labels with a list of string labels.
- Xlabel and ylabel: It is used to provide labels to the x-axis and y-axis respectively.
- Title: It is used to add a title to a plot.
- Scatter: IIt is used to create a scatter plot for the given data values.
- Legend: A legend is an area describing the elements of a graph. The legend function in Matplotlib is used to place a legend on the axes.

C. Numpy (III) : It is a python library used for working with arrays. It also has functions for working in the domain of linear algebra, fourier transform and matrices.

III. HYPOTHESES .

A. AUSOPEN-MEN-2013*: Is there a pattern or trend in the frequency of unforced errors and double faults committed by a player as they move from Round 1 to Round 7 of a tournament, given the increased gravity of the matches?*

B. AUSOPEN-WOMEN-2013: *What is the likelihood that a player who wins the first set of a match will go on to win the entire match?*

C. FRENCHOPEN-WOMEN-2013: *Can a scenario arise where a player, despite having a higher number of winners (wnr), could still end up losing the match due to other factors such as committing a higher number of unforced errors?*

D. FRENCHOPEN-MEN-2013: *What is the relationship between a player's percentage of break points won and their chances of winning the match? Can statistical data be used to establish a correlation between these two factors, and if so, how strong is this relationship?*

E. USOPEN-MEN-2013:*Can the number of aces won by a player be considered a crucial factor in determining their chances of winning a match, considering that a considerable number of points in the match can be earned through service alone?*

F. WIMBLEDON-MEN-2013: *Is there any correlation between the number of winners gained by Player 2 and the number of unforced errors committed by Player 1 in a tennis match?*

G. USOPEN-WOMEN-2013: *Can a disparity be observed in the distribution of number of first serve won (FSW) between players who emerge as winners and those who lose in a tennis match?*

H. WIMBLEDON-WOMEN-2013:*What trends can be observed in various parameters such as first serve won, second serve won, aces won, winners earned, break points won, and net points won over the seven rounds for the winner of the tournament?*

IV. ANSWER TO THE HYPOTHESIS

A. AUSOPEN-MEN-2013: *Is there a pattern or trend in the frequency of unforced errors and double faults committed by a player as they move from Round 1 to Round 7 of a tournament, given the increased gravity of the matches?*

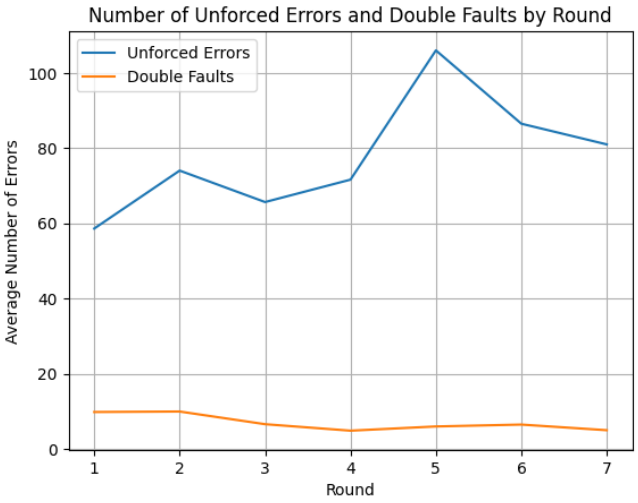


Figure 1.

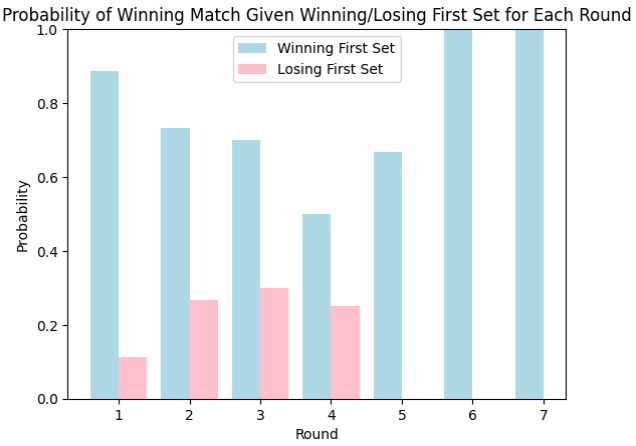
It was initially hypothesized that as the rounds of a tennis tournament progress from the first to the seventh, the number of unforced errors and double faults committed by players would decrease. This is due to the increase in the level of play and seriousness of the tournament, which should result in a decrease in the margin for errors.

However, upon analyzing the data presented in Figure 1, it is evident that the expected trend is not observed in a significant manner. While there is a slight decrease in the number of double faults committed by players, the decrease is not substantial. Conversely, the number of unforced errors exhibits a general increase from the first round to the seventh round of the tournament.

To conclude, the notion that the number of mistakes made by players would decrease as the tournament progresses due to the increase in the level of play and the tournament's seriousness is not necessarily

valid. This might be because the pressure on players also intensifies as the tournament progresses, which can be another factor that contributes to the occurrence of mistakes.

B. AUSOPEN-WOMEN-2013: *What is the likelihood that a player who wins the first set of a match will go on to win the entire match?*



. Figure 2.

Figure 2 presents a bar graph that displays the probability of winning a tennis match for player 1, depending on whether they won or lost the first set. The first bar in the graph represents the probability of winning the match after winning the first set, while the second bar indicates the probability of winning the match after losing the first set. Upon examination of the graph, it becomes apparent that the probability of winning the match after winning the first set is consistently higher than the probability of winning the match after losing the first set. Also, the graph indicates that in the last two rounds of the tournament, player 1 won all the matches in which they won the first set.

These findings suggest that winning the first set is an essential factor in determining the outcome of the match. The higher probability of winning after winning the first set could be attributed to the psychological advantage that winning the first set provides to the player. It gives them a boost in confidence and a sense of control over the match, which can positively impact their performance.

C. FRENCHOPEN-WOMEN-2013: *Can a scenario arise where a player, despite having a higher number of winners (wnr), could still end up losing the match due to other factors such as committing a higher number of unforced errors?*

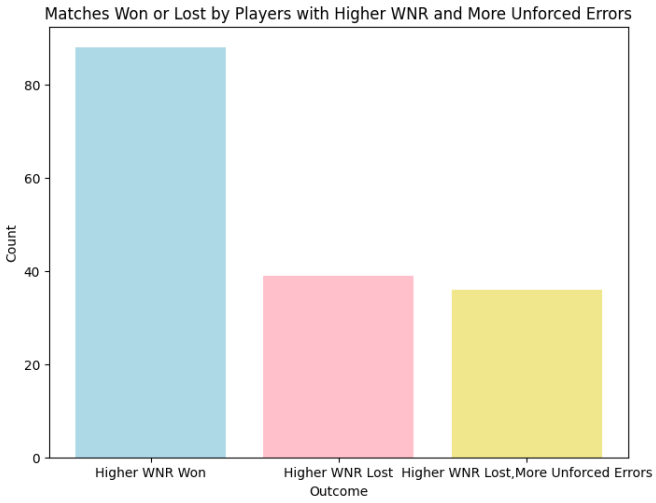


Figure 3.

Fig.3 displays a bar graph that illustrates the number of matches won and lost by the player with the higher winner count (wnr). The first bar represents the number of matches won by the player with the higher wnr, while the second bar represents the number of matches lost by the player with the higher wnr. The third bar represents the matches in which the player with the higher wnr lost due to committing a high number of unforced errors.

Upon analyzing the data presented in the graph, it can be observed that in most cases, the player with the higher wnr wins. However, there are still a significant number of instances in which the player with the higher wnr loses the match. This outcome could be attributed to various factors, with the most significant factor being a higher number of unforced errors committed.

In conclusion, while having a higher wnr count is a positive indicator of a player's performance, it is not the sole determinant of victory in a tennis match. The number of unforced errors committed is an equally crucial factor that must be minimized to increase the chances of winning the match.

D. FRENCHOPEN-MEN-2013: *What is the relationship between a player's percentage of break points won and their chances of winning the match?*

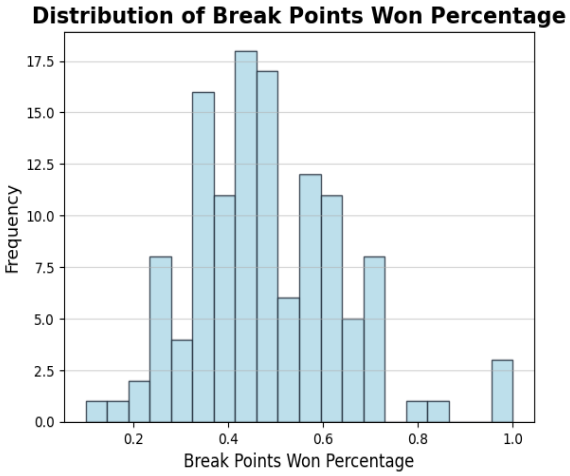


Figure 4.

Initially, it was expected that the winners of each match would have a higher percentage of break points won, which is calculated by dividing the number of break points won by the number of breakpoints created.

However, from the frequency distribution plot shown in Fig. 4, it can be inferred that having a very high break points won percentage is not necessarily a determining factor in winning a match. The peak of the graph lies in the range of 0.4 to 0.5, indicating that most of the winners had a break points won percentage of 40% to 50% only.

While some of the winners had a break points won percentage close to 100%, this was not the case for the majority of the winners.

Therefore, it can be concluded that having a high break points won percentage is not an essential criterion for winning a tennis match. Other factors, such as the ability to handle pressure situations and minimize unforced errors, also play a crucial role in determining the outcome of the match.

E. USOPEN-MEN-2013: *Can the number of aces won by a player be considered a crucial factor in determining their chances of winning a match, considering that a considerable number of points in the match can be earned through service alone?*

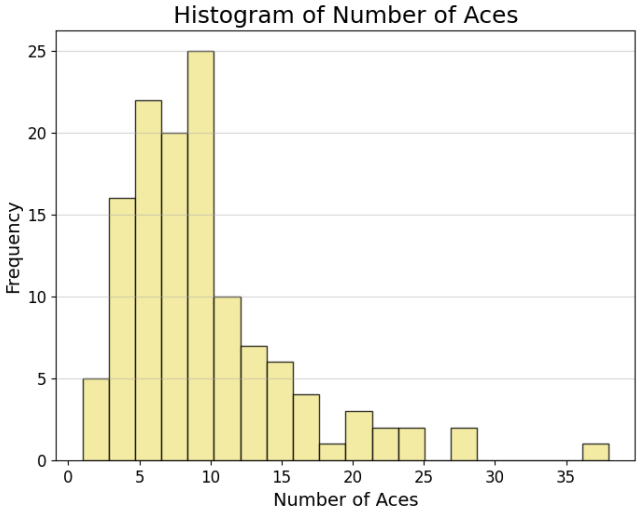


Figure 5.

It was initially expected that the winners will have a high number of aces won, as a significant number of points are scored through service.However, the histogram in Figure 5 reveals that this assumption was not accurate. The peak of the histogram is located towards the left of the graph, indicating that for most winners, the number of aces won is not a crucial factor contributing to their victory. Therefore, it can be concluded that although the number of aces won may play a role in determining the outcome of a match, it is not one of the primary factors.

F. WIMBLEDON-MEN-2013: *Is there any correlation between the number of winners gained by Player 2 and the number of unforced errors committed by Player 1 in a tennis match?*

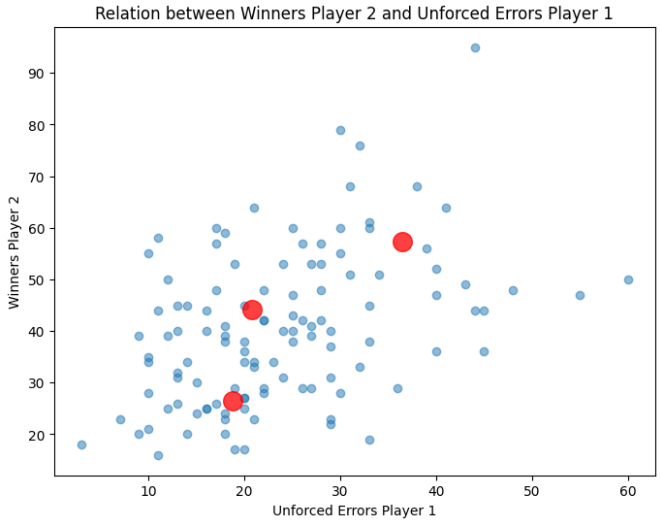


Figure 6.

It was initially assumed that there exists a direct relationship between the number of winners a player gains and the number of unforced errors committed by the other player. This is because a player earns a winner due to their own proficiency or the other player's lack of skill, and an unforced error is committed due to carelessness of a player. Therefore, if player 1 commits a high number of unforced errors, player 2 is expected to earn a greater number of winners.

Upon finding the correlation factor between the wnr of player 2 and unforced errors by player 1, it came out to be `0.41781230609196346`.

Since, the correlation factor came out to be positive, it shows a relation of direct proportionality between the two columns. But the value is not that close to one, showing that the relationship is not that strong.

Also, upon analyzing the scatter plot in figure 6, it is observed that there are three main clusters where the scatter points are concentrated. The majority of points are located in the region where player 1 commits an average number of unforced errors and a low number of winners are earned.

Hence, the expected trend is not strongly observed. It can be inferred that while the number of unforced errors committed by the other player does have an impact on the number of winners earned by the player, there are other determining factors that come into play.

G. USOPEN-WOMEN-2013: *Can a disparity be observed in the distribution of first serve won (FSW) between players who emerge as winners and those who lose in a tennis match?*

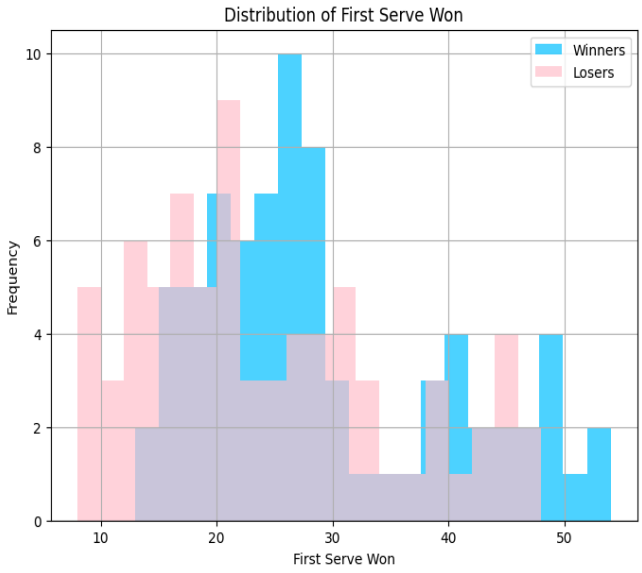


Figure 7.

The initial assumption is that the number of first serves won should be greater for the players who win the match compared to those who lose. From the frequency plot in fig.7, it can be observed that the peak for winners is more towards the right than the losers. This suggests that most of the winners won a higher number of first serves than most of the losers. Also, the highest number of first serves won are scored by the winners only.

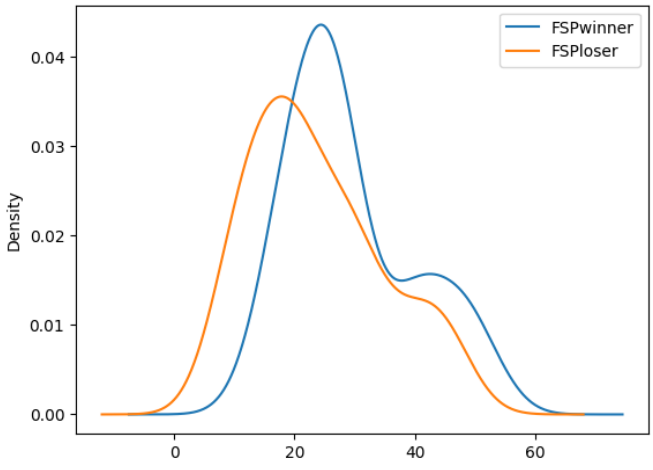


Figure 8.

The same observation can be made from the graph in fig.8. It is seen that the peak for winners is shifted more towards the right. This indicates that the probability of a winner having a high number of first serves won is greater than the probability of a loser having a high

number of first serves won.

H. WIMBLEDON-WOMEN-2013: What trends can be observed in various parameters such as first serve won, second serve won, aces won, winners earned, break points won, and net points won over the seven rounds for the winner of the tournament?

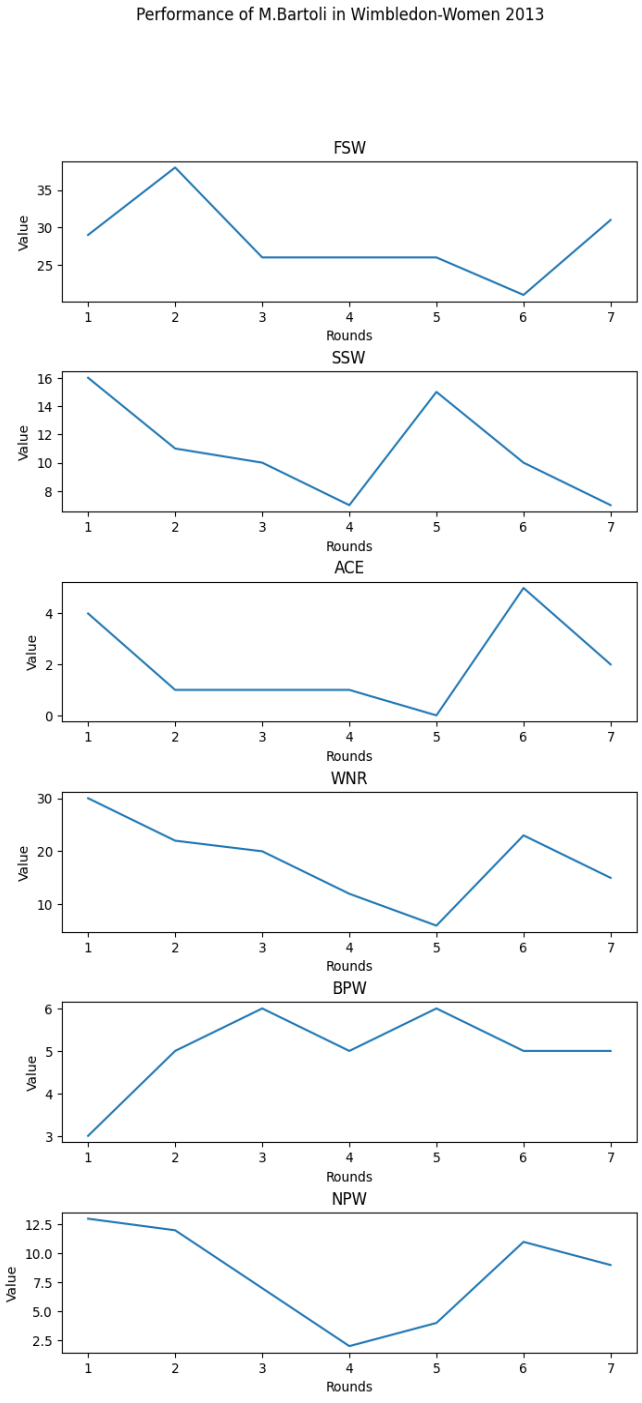


Figure 9.

Upon analyzing the graphs in figure 9, it can be observed that there is no distinct trend in the variation of the mentioned parameters for the winner of the tournament over the rounds. For some of the parameters, there is a decline followed by an increase in values from round 1 to round 7. On the other hand, for the remaining parameters, an increase in values followed by a decrease from round 1 to round 7 is evident.

V. SUMMARY OF THE OBSERVATIONS

The analysis of various graphs and data related to tennis matches has provided several interesting insights into the factors that determine the outcome of a match. Contrary to the expectation, the number of unforced errors committed by players does not necessarily decrease as the tournament progresses, indicating that the pressure on players increases as the tournament proceeds. Winning the first set is a crucial factor in determining the outcome of the match as it provides a psychological advantage to the player. While having a higher winner count is a positive indicator of a player's performance, minimizing unforced errors is equally crucial. Having a very high break points won percentage is not necessarily a determining factor in winning a match, indicating that handling pressure situations and minimizing unforced errors play a crucial role in determining the outcome. The number of aces won is not a primary factor contributing to victory. While the number of unforced errors committed by the other player does impact the number of winners earned by the player, there are other determining factors that come into play. The number of first serves won is not a reliable predictor of the outcome of the match. Overall, to perform well in a tennis match, a number of factors have to be considered.

VI. UNANSWERABLE QUESTIONS

WIMBLEDON-WOMEN-2013: What trend can be observed in the difference between the total points won by the winners and losers of the match as the tournament progresses from round 1 to round 7?

Unfortunately, there is no data available in the columns TPW.1 and TPW.2 in the dataset WIMBLEDON-WOMEN-2013, which makes the question unanswerable.

VI. ACKNOWLEDGEMENTS

I thank Professor Shanmuganathan Raman, department of Computer Science and Engineering and Electrical Engineering, Indian Institute of Technology, Gandhinagar for your invaluable guidance and support. Your dedication to teaching and research has been a source of inspiration to me.

VII. REFERENCES

(I) “Pandas Introduction,” n.d.

https://www.w3schools.com/python/pandas/pandas_intro.asp.

(II) GeeksforGeeks. “Python Introduction to Matplotlib,” May 14, 2018.

https://www.geeksforgeeks.org/python-introduction-matplotlib/.

(III) “Introduction to NumPy,” n.d.

https://www.w3schools.com/python/numpy/numpy_intro.asp.