Data Narratives-3 (April 2023)

Birudugadda Srivibhav -22110050, Computer Science and Technology Department, Prof. Shanmuga R, IIT Gandhinagar

***Abstract*- This data narrative project focuses on the analysis of tennis datasets to gain insights into the performance of professional tennis players. The project explores various aspects of the sport, such as player rankings, match statistics, and performance trends over time. The analysis uses various visualisation techniques to illustrate patterns and trends within the data. The aim of the project is to provide a deeper understanding of the factors that influence a player's success in tennis and to provide valuable insights for players, coaches, and fans of the sport.**

I. Aim

To analyse and compare the performance of male and female tennis players in the four Grand Slam tournaments (Australian Open, French Open, Wimbledon, and US Open) in the year 2013 and to investigate the factors that may influence player performance in these tournaments.

II. Overview of the datasets

Tennis Major Tournament Match Statistics:

The datasets used in this study consist of match statistics for male and female tennis players in the four Grand Slam tournaments (Australian Open, French Open, Wimbledon, and US Open) in the year 2013. There are eight separate datasets, with two for each tournament (one for men and one for women). Each dataset contains information on various aspects of each match, including player names, rounds, match outcome, number of games won by each player in each set, and several performance metrics such as first serve percentage, aces, double faults, winners, and unforced errors. The datasets are in CSV format and were imported into Python for data analysis and visualisation. The datasets provide a rich source of information for investigating the performance of tennis players in major tournaments and exploring the relationships between different performance metrics and match outcomes. Each dataset contains the following columns: 'Player1', 'Player2', 'Round', 'Result', 'FNL1', 'FNL2', 'FSP.1', 'FSW.1', 'SSP.1', 'SSW.1', 'ACE.1', 'DBF.1', 'WNR.1', 'UFE.1', 'BPC.1', 'BPW.1', 'NPA.1', 'NPW.1', 'TPW.1', 'ST1.1', 'ST2.1', 'ST3.1', 'ST4.1', 'ST5.1', 'FSP.2', 'FSW.2', 'SSP.2', 'SSW.2', 'ACE.2', 'DBF.2', 'WNR.2', 'UFE.2', 'BPC.2', 'BPW.2', 'NPA.2', 'NPW.2', 'TPW.2', 'ST1.2', 'ST2.2', 'ST3.2', 'ST4.2', 'ST5.2'.

More information about the columns of datasets is available online in the UC Irvine Machine Learning Repository **[1]**.

III. scientific questions/hypothesis

1. **Question:** Do players who win the first set of a match tend to win more matches overall? Is there a relationship between winning the first set and the final outcome of the match? Analyse AusOpen-men's and women's datasets
   Hypothesis: Winning the first set may give players a psychological advantage and increase their confidence, leading to better performance in subsequent sets.

2. **Question:** Do women handle breakpoints more effectively than men? Perform the analysis using the Wimbledon men's and women's tournament datasets.
   Hypothesis: I expect women to be effective handlers of breakpoints as per the analysis of the above question. However, any player who faces more breakpoints may experience tremendous pressure and be more likely to make unforced errors, leading to a lower overall win rate.

3. **Question:** Are there certain types of errors that are more common among players who lose matches? For example, do players who hit more double faults (DBF) or unforced errors (UFE) tend to lose more often? Analyse the French men's tournament.

4. **Question:** How does the distribution of breakpoints won differ between winning and losing players? Perform the analysis for the US men's tournament.
   Hypothesis: The winning players tend to have more breakpoint wins.

5. **Question:** Find which sets are won more by winning players. Perform the analysis using the US men's tournament dataset.

6. **Question:** Does the performance of a player in one aspect, such as first serve percentage, tend to compensate for a poor

performance in another aspect, such as unforced errors? Perform the analysis for the French women's tournament.

7. **Question:** Do players who attempt more net points tend to win more games than those who don't? Does the number of net points attempted by players correlate with the final number of games won? Perform the analysis using the US women's tournament dataset.
   Hypothesis: There should be a positive correlation, as it is one of the best ways to pressure opponents.

8. **Question:** Is there a correlation between a player's first serve percentage and their number of double faults? Perform the analysis using the French men's tournament dataset.
   Hypothesis: Players with a higher first-serve percentage will commit fewer double faults on average as they are good at serving.

IV. Details of libraries and functions

The libraries and modules used are:

[2] Numpy library - A Python library for scientific computing that provides tools for working with arrays and matrices, as well as mathematical functions and linear algebra operations.

[3] Pandas library- A Python library used for data manipulation and analysis. It provides data structures for efficiently storing and accessing large datasets and tools for cleaning, transforming, and analysing data.

[4] Matplotlib library- A Python library used for creating static, animated, and interactive visualisations in Python. It provides various plotting options and customisation tools to create publication-quality figures.

[5] Seaborn library- A Python library based on Matplotlib that provides additional visualisation tools, particularly for statistical data analysis. It provides a high-level interface for creating complex plots with minimal code.

The functions used are:

pd.merge - A function in Pandas that combines two or more data frames based on a common column.

value_counts - A function in Pandas used to count the occurrence of unique values in a column.

df.drop() - method to drop rows or columns from a pandas DataFrame

df.isna() - method to check if any values in a pandas DataFrame are missing (NaN)

df.fillna() - method to fill missing values in a pandas DataFrame with a specified value or method

plt.bar - A function in Matplotlib used to create a bar chart.

sns.histplot(), plt.hist() - A function from the seaborn library used to create a histogram plot of the distribution of data, with options for customising the bins, colours, and other features of the plot.

df.corr() - A method used to calculate the correlation matrix of a pandas data frame, which shows the pairwise correlation coefficients between the columns of the data frame.

plt.show - A function in Matplotlib used to display a plot.

pd.DataFrame - A class in Pandas used to create a data frame from a dictionary or array of data.

pd.read_csv - A function in Pandas used to read a CSV file into a data frame.

pd.groupby - Pandas groupby is used for grouping the data according to the categories and applying a function to the categories. It also helps to aggregate data efficiently. Pandas data frame. group by () function is used to split the data into groups based on some criteria.

sns.scatterplot - The scatterplot function in Seaborn is used to create a scatter plot that displays the relationship between two variables. It is useful for identifying data patterns and trends and can help reveal any correlations between the two variables.

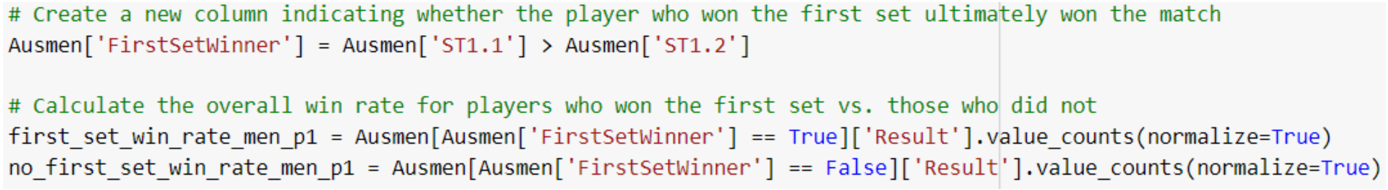
V. Answers to the questions

Question 1: Do players who win the first set of a match tend to win more matches overall? Is there a relationship between winning the first set and the final outcome of the match? Analyse AusOpen-men's and women's datasets
Hypothesis: Winning the first set may give players a psychological advantage and increase their confidence, leading to better performance in subsequent sets.
Answer:
In the analysis, I first loaded the "AusOpen-men-2013.csv" dataset into a pandas DataFrame using the read_csv() function. The dataset contains information about men's singles matches played at the 2013 Australian Open tennis tournament.

Next, a new column is created in the DataFrame called "FirstSetWinner". This column indicates whether the player won the first set or not. This is calculated by comparing the number of games won by each player in the first set (ST1.1 and ST1.2) and assigning a value of True to the "FirstSetWinner" column for the player who won the first set and a value of False for the player who did not.

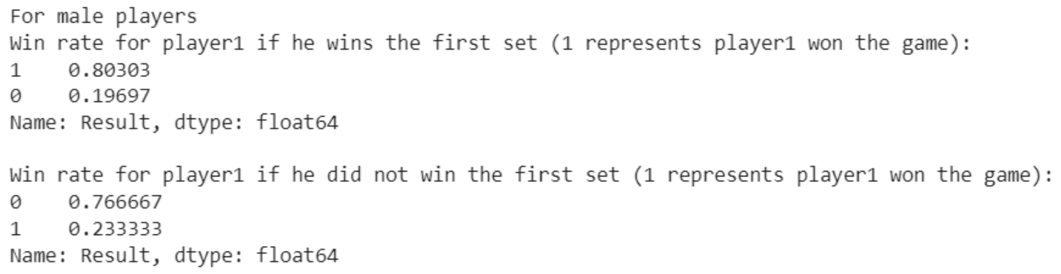
The next step is calculating the overall win rate for players who won the first set versus those who did not. This is done using the value_counts() function, which counts the number of occurrences of each unique value in the "Result" column and returns the result as a pandas Series object. By setting the normalize parameter to True, we can calculate the win rate as a proportion of the total number of matches played.

Finally, the win rates for each group are printed to the console using the print() function. This allows us to compare the win rates for players who won the first set versus those who did not. The same procedure is done for the women dataset also.



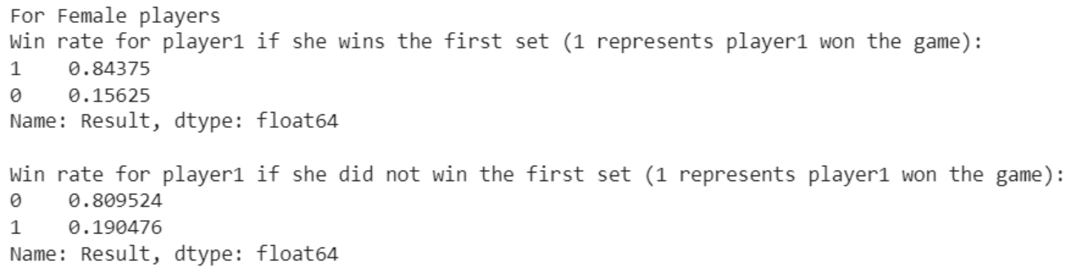
```
# Create a new column indicating whether the player who won the first set ultimately won the match
Ausmen['FirstSetWinner'] = Ausmen['ST1.1'] > Ausmen['ST1.2']

# Calculate the overall win rate for players who won the first set vs. those who did not
first_set_win_rate_men_p1 = Ausmen[Ausmen['FirstSetWinner'] == True]['Result'].value_counts(normalize=True)
no_first_set_win_rate_men_p1 = Ausmen[Ausmen['FirstSetWinner'] == False]['Result'].value_counts(normalize=True)
```



```
For male players
Win rate for player1 if he wins the first set (1 represents player1 won the game):
1    0.80303
0    0.19697
Name: Result, dtype: float64

Win rate for player1 if he did not win the first set (1 represents player1 won the game):
0    0.766667
1    0.233333
Name: Result, dtype: float64
```



```
For Female players
Win rate for player1 if she wins the first set (1 represents player1 won the game):
1    0.84375
0    0.15625
Name: Result, dtype: float64

Win rate for player1 if she did not win the first set (1 represents player1 won the game):
0    0.809524
1    0.190476
Name: Result, dtype: float64
```

For the first player, we can see that for men, the win rate after winning the first set is 80.3 per cent, and the win rate if the first set is lost is only 23.3 per cent. Similarly, for women, the values are 84.3 per cent and 19.04 per cent, respectively.

The analysis results support our hypothesis that winning the first set of a match is strongly correlated with winning the match overall. The much higher overall win rate for players who won the first set suggests that there is indeed a psychological advantage associated with taking an early lead in a match. However, it is essential to note that this analysis does not prove causation; it is possible that players who win the first set are simply better overall and would have won the match anyway, regardless of the first set outcome. Future studies could investigate this question further by controlling for other factors that may influence the outcome of a match, such as player skill level or physical fitness. Also, the win rate after the first set win for women is higher than that of men indicating women have a more psychological advantage over men.

Question 2: Do women handle breakpoints more effectively than men? Perform the analysis using the Wimbledon men's and women's tournaments datasets.
Hypothesis: I expect women to be effective handlers of breakpoints as per the analysis of the above question. However, any player who faces more breakpoints may experience tremendous pressure and be more likely to make unforced errors, leading to a lower overall win rate.
Answer:
To investigate whether women handle breakpoints more effectively than men, I have calculated the break point conversion rate for each player, which is the percentage of break points saved out of the total break points faced. I then plotted the distribution of breakpoint conversion rates for men and women using a histogram. For this, I have defined the bp_conversion_rate function as follows:

The function bp_conversion_rate takes a pandas DataFrame df as input.

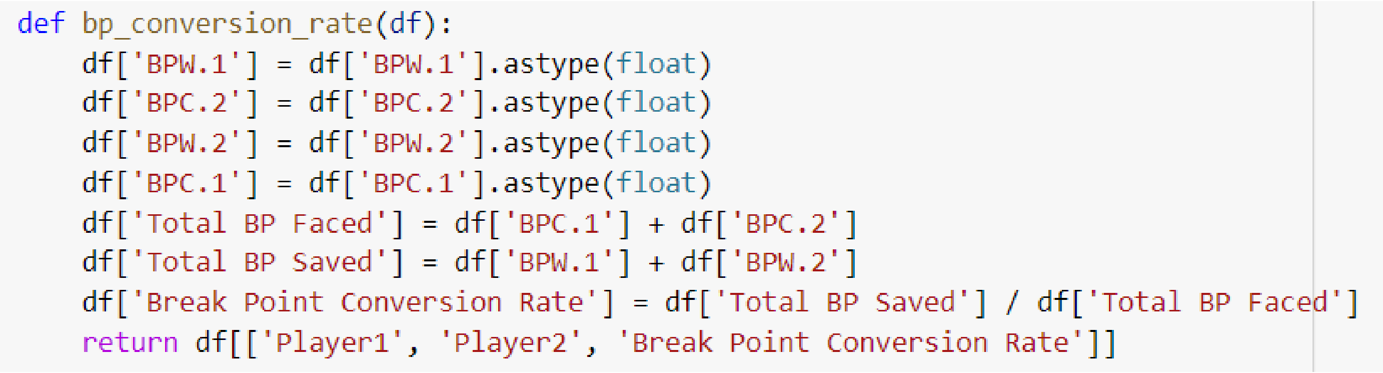
The next four lines of code convert the columns 'BPW.1', 'BPC.2', 'BPW.2', and 'BPC.1' to float data type using the astype() method. These columns represent the number of breakpoints won and faced by each player in the match.

The next two lines of code create two new columns: 'Total BP Faced' and 'Total BP Saved', which represent the total number of breakpoints faced and saved by each player in the match, respectively. The values in these columns are obtained by adding the break points faced and won by each player.

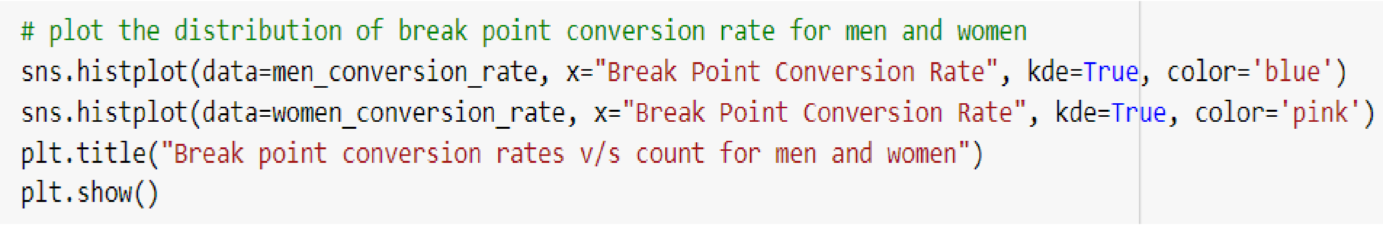
The next line of code calculates the 'Break Point Conversion Rate' for each player by dividing their total number of breakpoints saved by the total number of breakpoints faced. This metric represents the percentage of breakpoints that the player was able to save.

Finally, the function returns a new data frame that includes the player names and their corresponding break point conversion rates. This function is applied to Wimbledon-men-2013.csv and Wimbledon-women-2013.csv datasets.

Overall, this code is used to calculate the break point conversion rate for each player in a tennis match. This metric can be used to assess how well players are able to handle break points and convert them into their own favour.

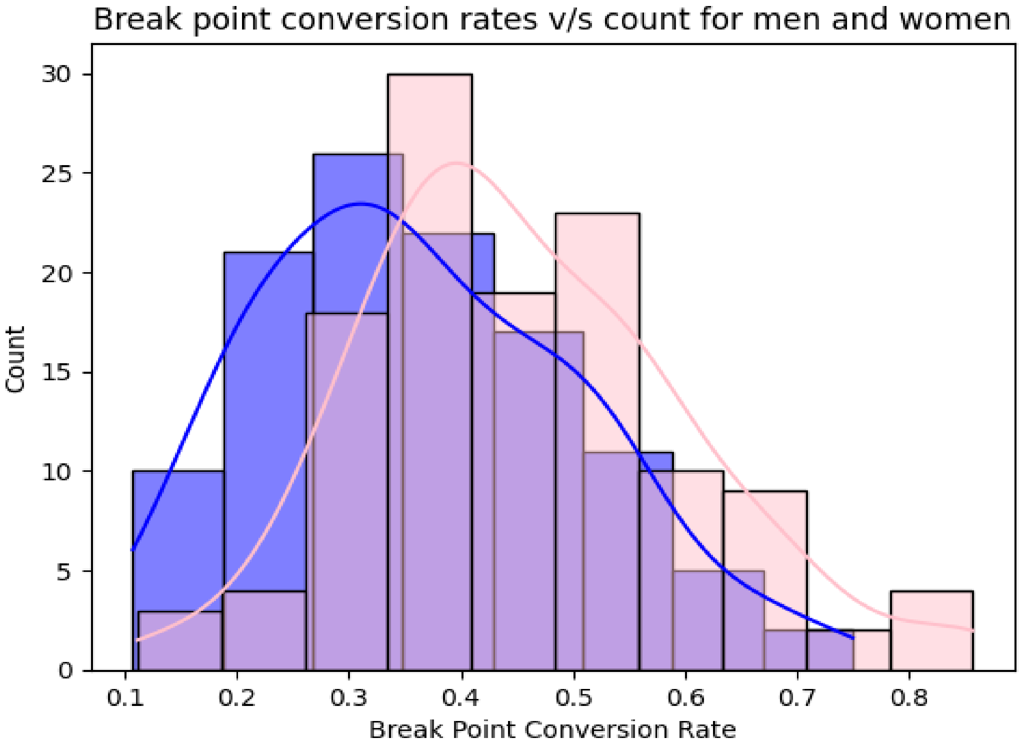


```python
def bp_conversion_rate(df):
    df['BPW.1'] = df['BPW.1'].astype(float)
    df['BPC.2'] = df['BPC.2'].astype(float)
    df['BPW.2'] = df['BPW.2'].astype(float)
    df['BPC.1'] = df['BPC.1'].astype(float)
    df['Total BP Faced'] = df['BPC.1'] + df['BPC.2']
    df['Total BP Saved'] = df['BPW.1'] + df['BPW.2']
    df['Break Point Conversion Rate'] = df['Total BP Saved'] / df['Total BP Faced']
    return df[['Player1', 'Player2', 'Break Point Conversion Rate']]
```



```python
# plot the distribution of break point conversion rate for men and women
sns.histplot(data=men_conversion_rate, x="Break Point Conversion Rate", kde=True, color='blue')
sns.histplot(data=women_conversion_rate, x="Break Point Conversion Rate", kde=True, color='pink')
plt.title("Break point conversion rates v/s count for men and women")
plt.show()
```

Later, I also plotted the distribution of breakpoint conversion rates for men and women using a histogram. Note that the graph is for the Wimbledon tournament dataset.



.

Fig-1: This figure shows the histogram between the breakpoint conversion rates and counts for men and women

The histogram shows that the breakpoint conversion rates for men and women are distributed fairly similarly, with most players having a conversion rate between 10% and 50%. However, there is a slightly higher proportion of women with conversion rates above 30%, while men have a slightly higher proportion of conversion rates below 40%. From the graph, it is clear that there are slightly higher conversion rates for more women, indicating that women have a more psychological advantage over men. This also supports the hypothesis of the previous question of most women having a more psychological advantage over men. It is also clear that only a very small percentage of players (both male and female) have higher conversion rates, indicating that players who face more breakpoints may experience tremendous pressure and be more likely to make unforced errors, leading to a lower overall win rate.

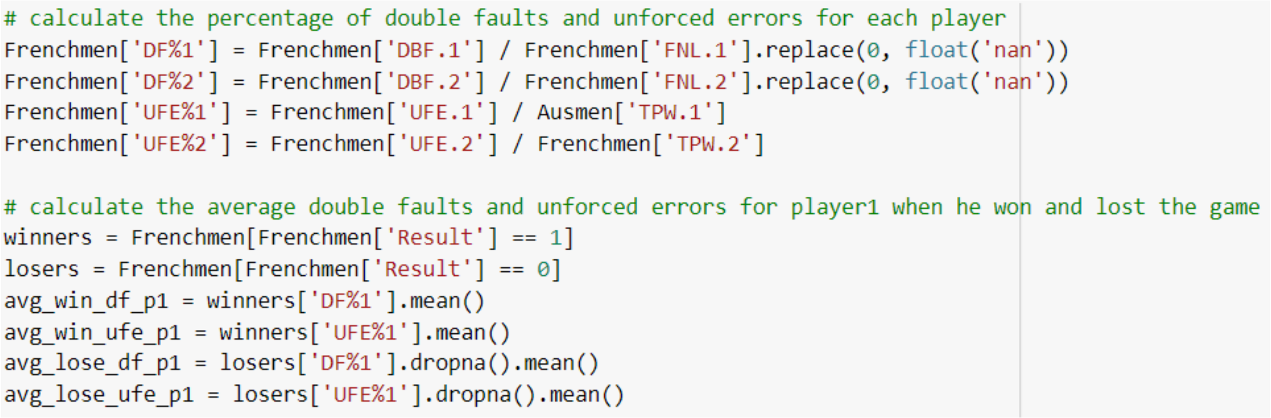
Question 3: Are there certain types of errors that are more common among players who lose matches? For example, do players who hit more double faults (DBF) or unforced errors (UFE) tend to lose more often? Perform the analysis using the French men's tournament.

Answer:

I have first imported the necessary libraries. In this case, I only need pandas. later I read the data from the FrenchOpen-men-2013 CSV file into a pandas DataFrame.

Next, calculate the percentage of double faults and unforced errors for each player. This is done by dividing the total number of double faults or unforced errors by the number of sets won by the player. This gives us a percentage value that we can use to compare players who win and lose matches.

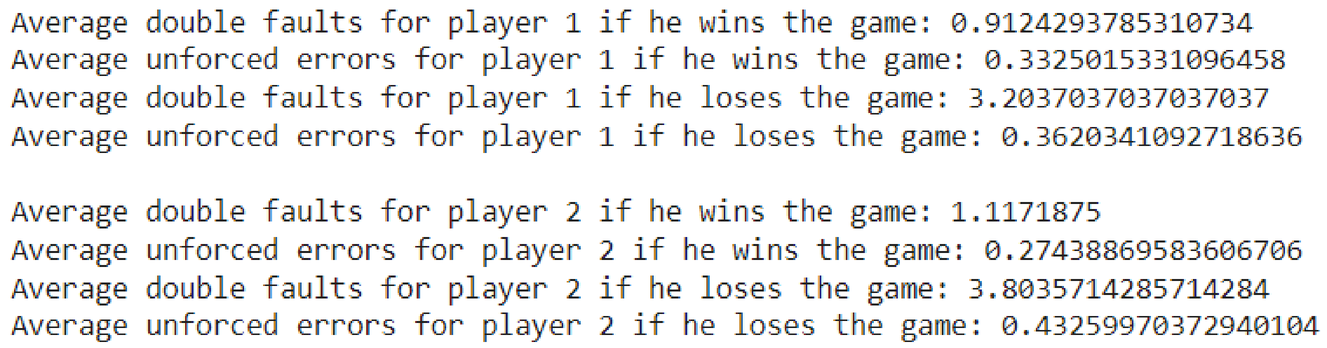
Now, calculate the average double faults and unforced errors for each player when he wins the match and also when he loses the match. This is done by filtering the data frame based on the 'Result' column. I created two new DataFrames, 'winners' and 'losers', that contain only the rows where the 'Result' column is 1 or 0, respectively. I then calculated the mean of the 'DF%1' and 'UFE%1' columns for both DataFrames to get the analysis of player 1, as shown below. Note that result = 1 means player 1 won. A similar format of code is also followed for player 2.



```
# calculate the percentage of double faults and unforced errors for each player
Frenchmen['DF%1'] = Frenchmen['DBF.1'] / Frenchmen['FNL.1'].replace(0, float('nan'))
Frenchmen['DF%2'] = Frenchmen['DBF.2'] / Frenchmen['FNL.2'].replace(0, float('nan'))
Frenchmen['UFE%1'] = Frenchmen['UFE.1'] / Ausmen['TPW.1']
Frenchmen['UFE%2'] = Frenchmen['UFE.2'] / Frenchmen['TPW.2']

# calculate the average double faults and unforced errors for player1 when he won and lost the game
winners = Frenchmen[Frenchmen['Result'] == 1]
losers = Frenchmen[Frenchmen['Result'] == 0]
avg_win_df_p1 = winners['DF%1'].mean()
avg_win_ufe_p1 = winners['UFE%1'].mean()
avg_lose_df_p1 = losers['DF%1'].dropna().mean()
avg_lose_ufe_p1 = losers['UFE%1'].dropna().mean()
```

The results of the analysis show that, on average, players who lose matches tend to hit more double faults and make more unforced errors than players who win matches. This supports the hypothesis that certain types of errors may be more common among players who lose matches.



```
Average double faults for player 1 if he wins the game: 0.9124293785310734
Average unforced errors for player 1 if he wins the game: 0.3325015331096458
Average double faults for player 1 if he loses the game: 3.2037037037037037
Average unforced errors for player 1 if he loses the game: 0.3620341092718636

Average double faults for player 2 if he wins the game: 1.1171875
Average unforced errors for player 2 if he wins the game: 0.27438869583606706
Average double faults for player 2 if he loses the game: 3.8035714285714284
Average unforced errors for player 2 if he loses the game: 0.43259970372940104
```

Based on the analysis, we can conclude that there is a significant relationship between the number of double faults and unforced errors and the likelihood of losing a match. On average, players who hit more double faults and unforced errors tend to lose more often than players who hit fewer errors.
The average number of double faults and unforced errors for winners is significantly lower than that of losers. This suggests that minimising the number of errors is crucial for winning matches.

However, it is important to note that other factors, such as the level of competition and the playing style of opponents, may also influence the outcome of a match. Therefore, further research is needed to determine the specific types of errors that have the greatest impact on match outcomes and to develop strategies to minimise these errors.

Question 4: How does the distribution of breakpoints won differ between winning and losing players? Perform the analysis for the US men's tournament.
Hypothesis: The winning players tend to have more breakpoint wins.
Answer:
Break points are crucial in tennis as they provide an opportunity for a player to win a game on their opponent's serve. In this analysis, I have explored the difference in the distribution of breakpoints won by winning and losing players.

For the analysis, I have used histograms to analyse and visualise the distribution of breakpoints won and created in tennis matches.

Firstly, I have created a histogram of the break points won by Player 1 for both winning and losing matches. This is done as follows:

The plt.hist() function creates a histogram. The first argument is the data to be plotted, which is df[df["Result"] == 1]["BPW.1"] for the winning matches and df[df["Result"] == 0]["BPW.1"] for the losing matches. The alpha parameter is set to 0.5 to make the bars semi-transparent. The label parameter is set to "Winning" and "Losing" to create a legend. The xlabel and ylabel functions are used to label the x and y axes. The legend() function displays the legend, and plt.show() displays the plot.

The same code is also used for player 2. Overall, the code creates two histograms to compare the distributions of break points won by Player 1 and Player 2 for winning and losing matches. The alpha parameter is set to 0.5 to make the bars semi-transparent, and the label parameter is used to create a legend. The xlabel and ylabel functions label the x and y axes, and the legend() function displays the legend. The plt.show() function displays the plot.

The graphs are shown below. Clearly, we can see that the distribution of breakpoints won by winning players is generally higher than the distribution of breakpoints won by losing players. This observation supports my hypothesis that the winning players tend to have more breakpoint wins.

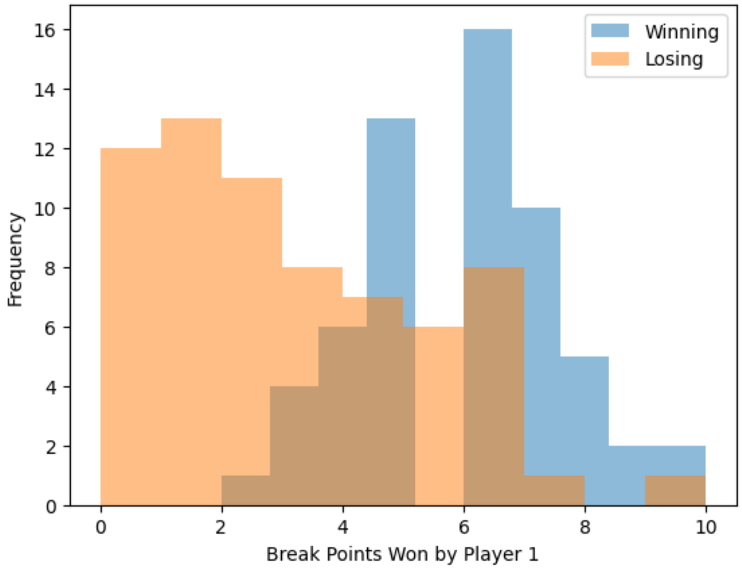


Fig-2: This figure shows the distribution of breakpoints won by player 1 in matches where he won or lost.

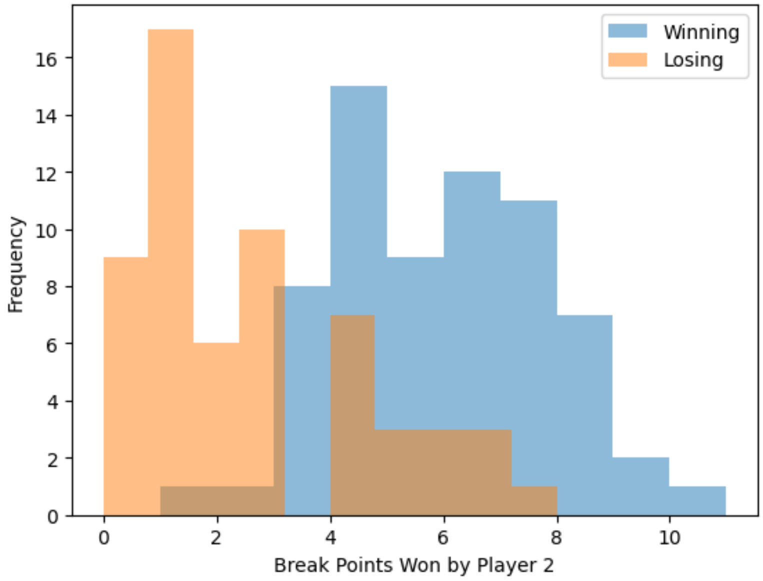


Fig-3: This figure shows the distribution of breakpoints won by player 2 in matches where he won or lost.

Question 5: Find which sets are won more by winning players. Perform the analysis using the US men's tournament dataset.
Answer:
To achieve the above task, I first wrote a code that does the following:
Firstly, the dataset is filtered to include only matches won by Player 1 by selecting all rows where the "Result" column is equal to 1. Next, dictionary set_wins is initialised to count the number of times each set is won by Player 1.

Later, A for loop iterates over each row of the filtered dataset df_wins and checks which player won each set by comparing the scores in columns ST1.1, ST2.1, ST3.1, ST4.1, and ST5.1 to the corresponding scores in columns ST1.2, ST2.2, ST3.2, ST4.2, and ST5.2. If Player 1 wins a set, the count for that set is incremented in the set_wins dictionary. A bar chart is plotted using matplotlib to display the number of set wins for Player 1. The same process is repeated for Player 2 by filtering the dataset for matches won by Player 2 and checking which player won each set.

I have obtained the following results:

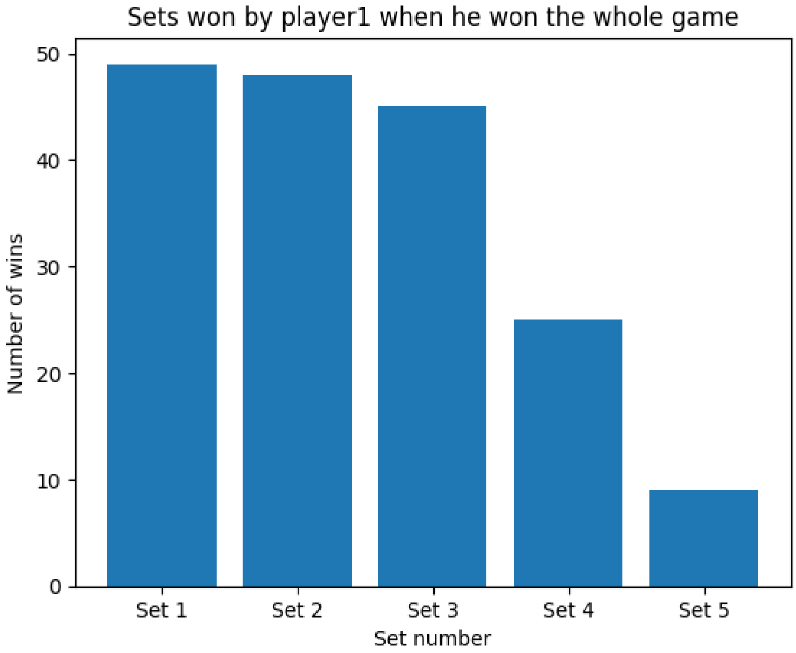


Fig-4: This figure shows the sets won by player1 when he won the whole game

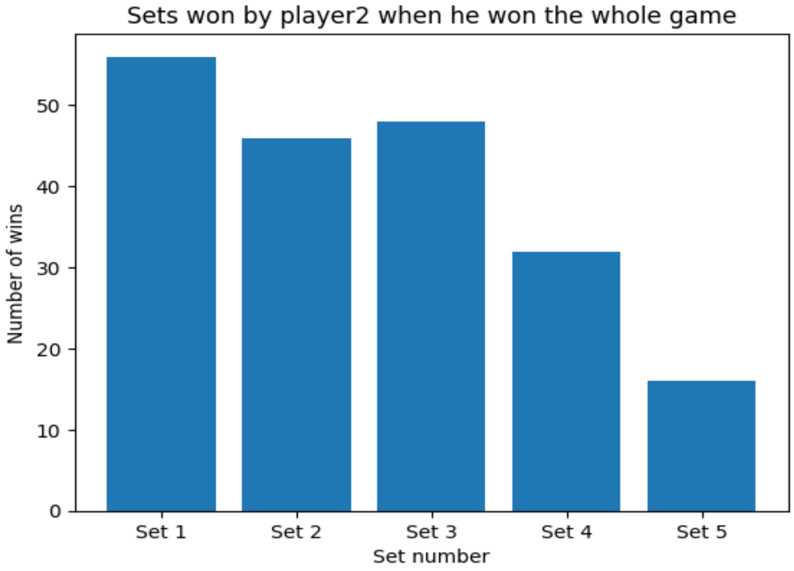
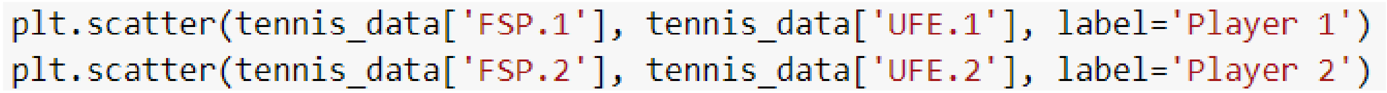


Fig-5: This figure shows the sets won by player2 when he won the whole game

From the graphs, it can be seen that the winning players mostly score in the first three sets, i.e. they play more aggressively in the first three sets. It is self-evident that winning in the initial games itself helps in building self-confidence and also suppresses the confidence of the opponent. However, it's also possible that playing more aggressively can lead to more errors and more fatigue, which can lead to losing the match. Thus, the relationship between playing style and winning is complex and can depend on multiple factors.

Question 6: Does the performance of a player in one aspect, such as first serve percentage, tend to compensate for a poor performance in another aspect, such as unforced errors? Perform the analysis for the French women's tournament.
Answer:
To analyse the question of whether a player's performance in one aspect of the game compensates for a poor performance in another aspect, I will look at the relationship between a player's first serve percentage (FSP) and their number of unforced errors (UFE).

I started by importing the dataset into Python using the pandas library and performing some data cleaning to remove missing values and unnecessary columns. Then, we will create a scatter plot to visualise the relationship between FSP and UFE for each player. I also calculated the correlation coefficient between these two variables to quantify the strength of the relationship.



```
plt.scatter(tennis_data['FSP.1'], tennis_data['UFE.1'], label='Player 1')
plt.scatter(tennis_data['FSP.2'], tennis_data['UFE.2'], label='Player 2')
```

The resulting scatter plot shows the relationship between a player's FSP and UFE for each player. Each point represents one player's performance in a single match. We can see that there is a general trend of players with higher FSPs having fewer UFEs, indicating that a strong performance in one aspect of the game may compensate for a poor performance in another. That means that professional players avoid committing silly mistakes.

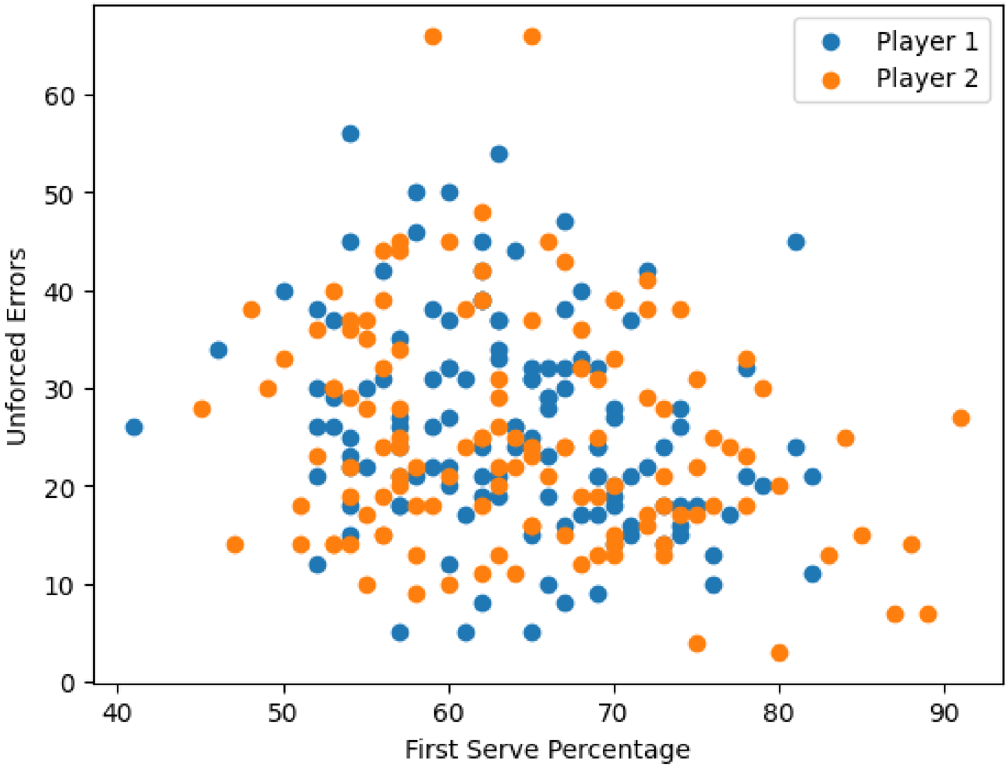
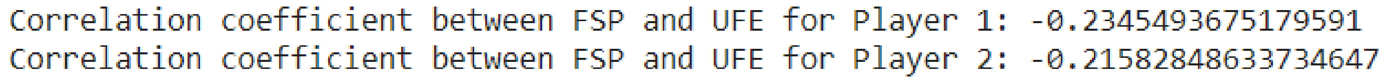


Fig-6: This figure shows the relationship between UFEs and FSPs for each player.

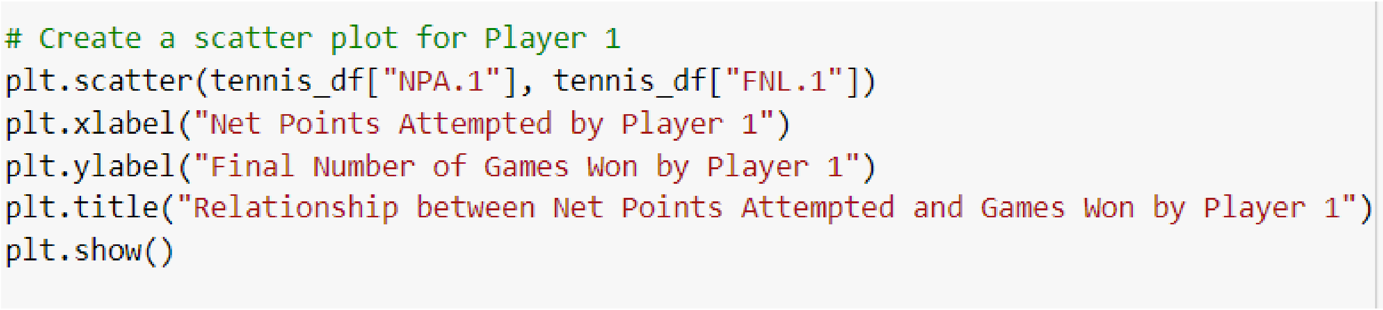
The correlation coefficient between FSP and UFE for Player 1 is -0.2340, indicating a moderate negative correlation between these two variables. This suggests that as a player's FSP increases, their UFEs tend to decrease.



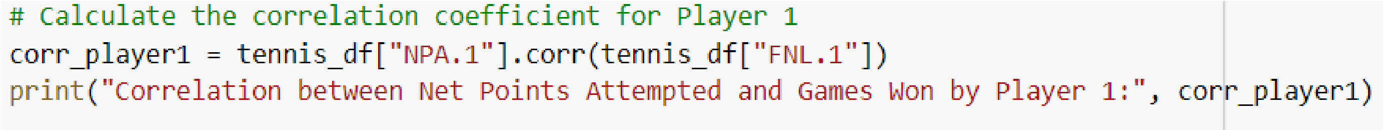
```
Correlation coefficient between FSP and UFE for Player 1: -0.2345493675179591
Correlation coefficient between FSP and UFE for Player 2: -0.21582848633734647
```

Question 7: Do players who attempt more net points tend to win more games than those who don't? Does the number of net points attempted by players correlate with the final number of games won? Perform the analysis using the US women's tournament dataset.
Hypothesis: There should be a positive correlation, as it is one of the best ways to pressure opponents.
Answer:
In tennis, a net point is a point that is played when a player comes to the net (the area of the court close to the net) during a point and hits the ball before it bounces. The aim of coming to the net is to put pressure on the opponent and to be in a better position to hit a winner. If the player wins the point at the net, it is called a net point won.

I want to investigate if there is a correlation between the number of net points attempted by a player and their chances of winning the game. I will separately analyse the data for both players, using scatter plots to visualise the relationship between the number of net points attempted and the final number of games won.



```
# Create a scatter plot for Player 1
plt.scatter(tennis_df["NPA.1"], tennis_df["FNL.1"])
plt.xlabel("Net Points Attempted by Player 1")
plt.ylabel("Final Number of Games Won by Player 1")
plt.title("Relationship between Net Points Attempted and Games Won by Player 1")
plt.show()
```



```
# Calculate the correlation coefficient for Player 1
corr_player1 = tennis_df["NPA.1"].corr(tennis_df["FNL.1"])
print("Correlation between Net Points Attempted and Games Won by Player 1:", corr_player1)
```

I expect that players who attempt more net points will have a higher chance of winning the game because they will be able to put more pressure on their opponents and win more points.

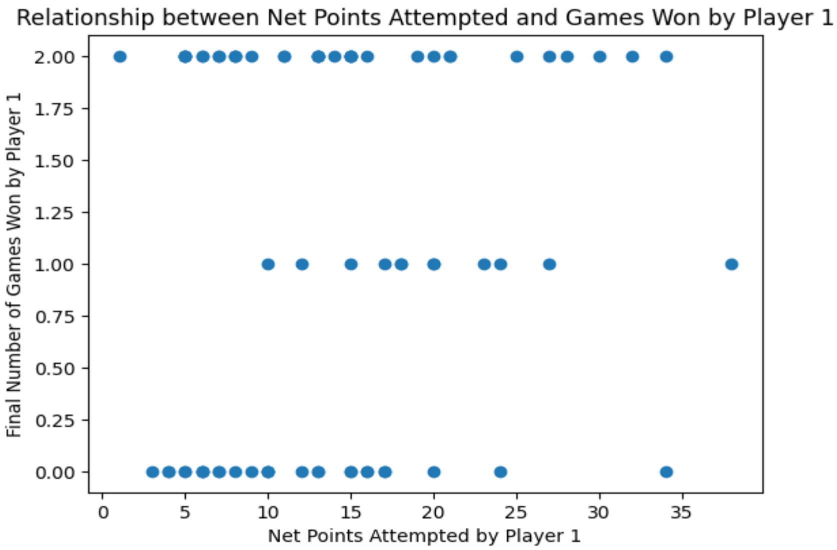


Fig-7: This figure shows the relationship between NPAs and FNLs for player 1.

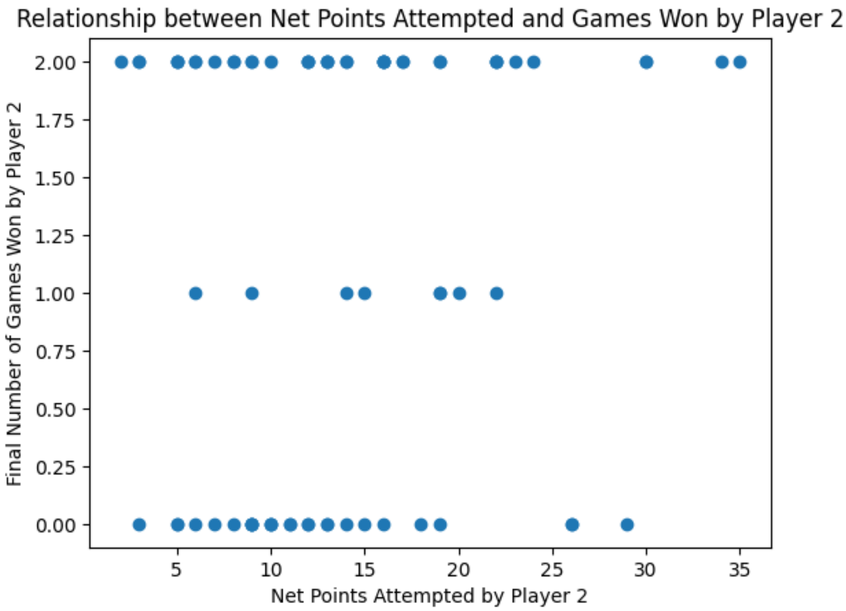
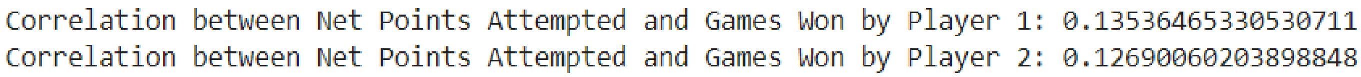


Fig-8: This figure shows the relationship between NPAs and FNLs for player 2



```
Correlation between Net Points Attempted and Games Won by Player 1: 0.13536465330530711
Correlation between Net Points Attempted and Games Won by Player 2: 0.12690060203898848
```

Surprisingly, I have found out that there is a very weak correlation between the number of net points attempted and the number of games won by both players; this result overrules my hypothesis of having more wins for players who attempt more net points.

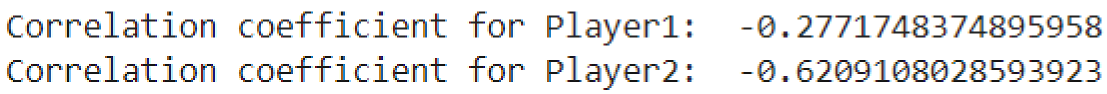
Question 8: Is there a correlation between a player's first serve percentage and their number of double faults? Perform the analysis using the French men’s tournament dataset.
Hypothesis: Players with a higher first-serve percentage will commit fewer double faults on average as they are good at serving.
Answer:
To analyse the correlation between a player's first serve percentage and their number of double faults, I have used the pandas library in Python. I have used the French men’s tournament dataset as mentioned in the question.

First, I imported the necessary libraries and read the dataset. Then, I created a scatter plot to visualise the relationship between first serve percentage and the number of double faults for both players.

Finally, we can calculate the correlation coefficient between the first serve percentage and the number of double faults for both players. The correlation coefficient for each player will give us an idea of the strength of the relationship between first serve percentage and the number of double faults. A negative correlation coefficient would support the hypothesis that players with a higher first-serve percentage will commit fewer double faults on average.

I have got the correlation coefficient to be negative, indicating that my hypothesis is right. Players with a higher first-serve percentage tend to commit fewer double faults because they are likely to be more accurate and consistent in their serves. A higher first serve percentage means that the player is able to successfully land more of their first serves in the opponent's court, which puts them in a more advantageous position in the point. When players miss their first serve, they have to rely on their second serve, which is usually slower and less accurate than their first serve. This makes it more difficult to win points and increases the likelihood of committing double faults.

In contrast, players with a higher first-serve percentage are able to keep the pressure on their opponent, making it more difficult for them to return the ball successfully. This gives the server an advantage in the point and reduces the likelihood of committing double faults.



```
Correlation coefficient for Player1:  -0.2771748374895958
Correlation coefficient for Player2:  -0.6209108028593923
```

The above result can also be seen from the scatter plots that I have plotted, which are given below.

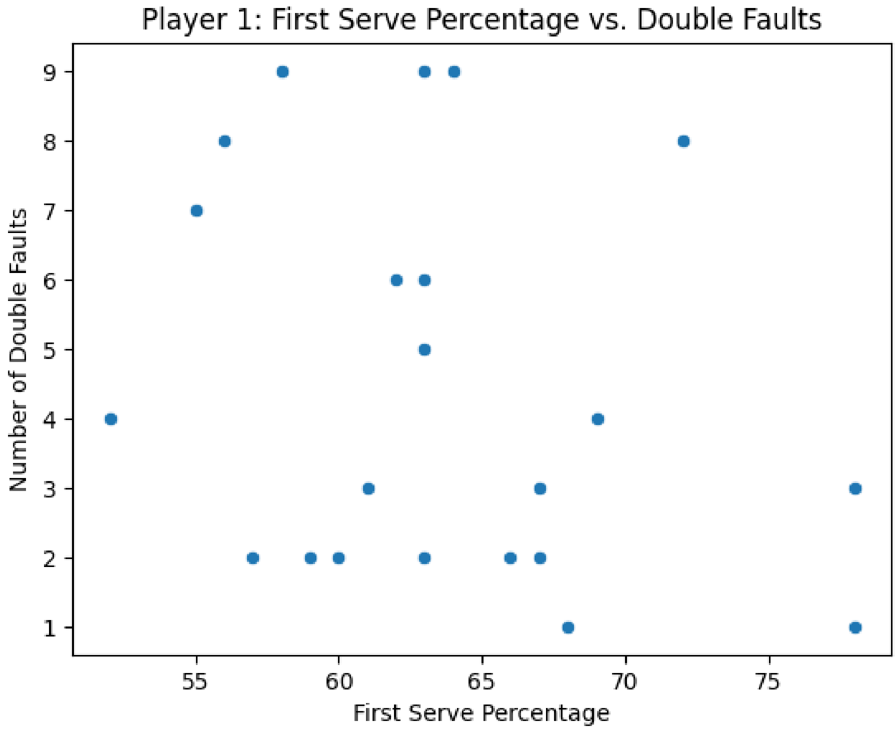


Fig-9: This figure shows the relationship between FSPs and DBFs for player 1

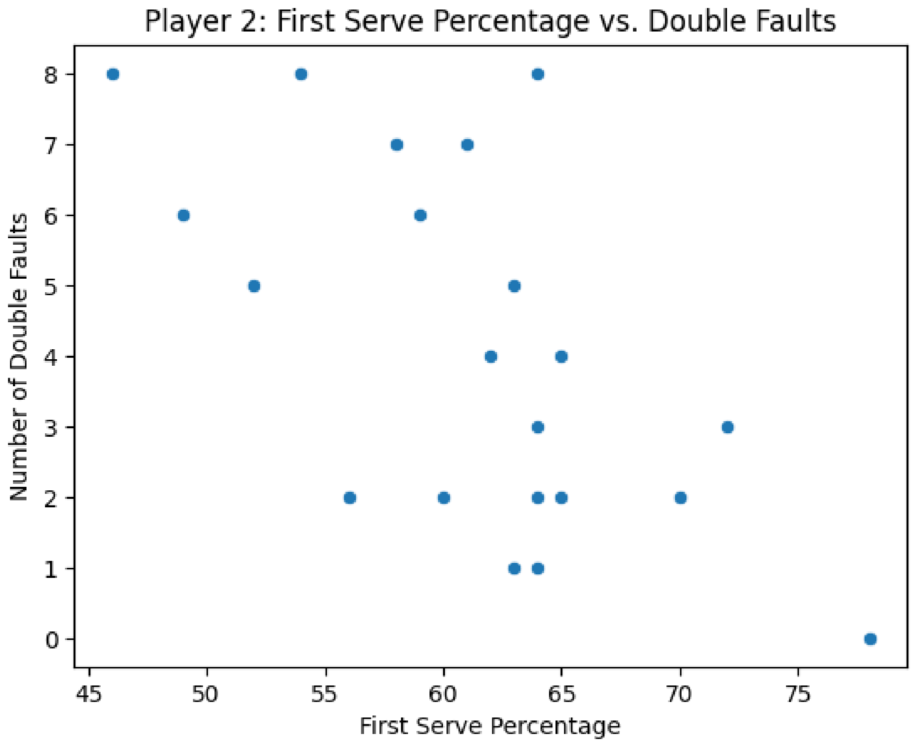


Fig-10: This figure shows the relationship between FSPs and DBFs for player 2

VI. SUMMARY AND OBSERVATIONS

1. Question 1:

The results of the analysis show that players who win the first set of a match have a significantly higher overall win rate compared to those who do not. For example, male players who win the first set have a win rate of 80.3%, while those who do not have a win rate of only 23.3%. This suggests that there is a strong relationship between winning the first set and winning the match overall, supporting our hypothesis that winning the first set gives players a psychological advantage and increases their confidence, leading to better performance in subsequent sets. Also, for women, the win rate after winning the first set is a bit higher, which is 84.3 per cent, which indicates women have a more psychological advantage over men.

However, it's important to note that this analysis does not prove causation. It is possible that other factors, such as player skill level or physical fitness, could also influence the outcome of the match.

2. Question 2:

The histogram shows that the breakpoint conversion rates for men and women are distributed fairly similarly, with most players having a conversion rate between 10% and 50%. However, there is a slightly higher proportion of women with conversion rates above 30%, while men have a slightly higher proportion of conversion rates below 40%. From the graph, it is clear that there are slightly higher conversion rates for more women, indicating that women have a more psychological advantage over men. This also supports the hypothesis of the previous question of most women having a more psychological advantage over men. It is also clear that only a very small percentage of players (both male and female) have higher conversion rates, indicating that players who face more breakpoints may experience tremendous pressure and be more likely to make unforced errors, leading to a lower overall win rate.

3. Question 3:

Based on the analysis, we can conclude that there is a significant relationship between the number of double faults and unforced errors and the likelihood of losing a match. On average, players who hit more double faults and unforced errors tend to lose more often than players who hit fewer errors.

The average number of double faults and unforced errors for winners is significantly lower than that of losers. This suggests that minimising the number of errors is crucial for winning matches.

However, it is essential to note that other factors, such as the level of competition and the playing style of opponents, may also influence the outcome of a match. Therefore, further research is needed to determine the specific types of errors that have the greatest impact on match outcomes and to develop strategies to minimise these errors.

4. Question 4:

From the graphs, we can clearly see that the distribution of breakpoints won by winning players is generally higher than the distribution of breakpoints won by losing players. This observation supports my hypothesis that the winning players tend to have more breakpoint wins.

Break points can certainly play a key role in winning a tennis match. A break point occurs when the serving player is one point away from losing the game, which gives the returning player a significant advantage. If the returning player can win the breakpoint, they will break serve and gain a lead in the set. Breaking serve is often critical to winning a set, and winning a set is critical to winning a match. So, winning more break points than one's opponent can certainly increase the likelihood of winning a match. However, it is not the only factor in winning a match, as other factors such as serving, returning, and overall consistency also play a role.

5. Question 5:

From the graphs, it can be seen that the winning players mostly score in the first three sets, i.e. they play more aggressively in the first three sets. It is self-evident that winning in the initial games itself helps in building self-confidence and also suppresses the confidence of the opponent. However, it's also possible that playing more aggressively can lead to more errors and more fatigue, which can lead to losing the match. Thus, the relationship between playing style and winning is complex and can depend on multiple factors.

6. Question 6:

The resulting scatter plot shows the relationship between a player's FSP and UFE for each player. Each point represents one player's performance in a single match. We can see that there is a general trend of players with higher FSPs having fewer UFEs, indicating that a strong performance in one aspect of the game may compensate for a poor performance in another. That means that professional players avoid committing silly mistakes.

The correlation coefficient between FSP and UFE for Player 1 is -0.2340, indicating a moderate negative correlation between these two variables. This suggests that as a player's FSP increases, their UFEs tend to decrease.

7. **Question 7:**

I have obtained the following results:
Correlation between Net Points Attempted and Games Won by Player 1: 0.13536465330530711
Correlation between Net Points Attempted and Games Won by Player 2: 0.12690060203898848
Clearly, there is a very weak correlation between the number of net points attempted and the number of games won by both players; this result overrules my hypothesis of having more wins for players who attempt more net points.

8. **Question 8:**

I got the correlation coefficient to be negative, indicating that my hypothesis is correct. Players with a higher first-serve percentage tend to commit fewer double faults because they will likely be more accurate and consistent in their serves. A higher first-serve percentage means that the player can successfully land more of their first serves in the opponent's court, which puts them in a more advantageous position in the point. When players miss their first serve, they have to rely on their second serve, which is usually slower and less accurate than their first serve. This makes it more difficult to win points and increases the likelihood of committing double faults.

In contrast, players with a higher first-serve percentage can keep the pressure on their opponent, making it more difficult for them to return the ball successfully. This gives the server an advantage in the point and reduces the likelihood of committing double faults.

VII. REFERENCES

1. Tennis Major Tournament Match Statistics, 2019. [Online]. Available: https://archive-beta.ics.uci.edu/dataset/300/tennis+major+tournament+match+statistics. [Accessed: April 22, 2023].

2. NumPy: Travis E, Oliphant. "A guide to NumPy", Proceedings of the 7th Python in Science Conference, vol. 1, 2008, pp. 1-7.

3. Pandas: Wes McKinney. "Data structures for statistical computing in Python", Proceedings of the 9th Python in Science Conference, vol. 445, 2010, pp. 56-61.

4. Matplotlib: J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

5. J. Waskom, M. Botvinnik, O. Hobson et al., “seaborn: v0.11.2 (November 2021),” Zenodo, 08-Dec-2021. [Online]. Available: https://doi.org/10.5281/zenodo.5737980. [Accessed: 23-Feb-2023].

VIII. ACKNOWLEDGEMENT

I want to sincerely thank Professor Shanmuga R for his important advice and assistance in the completion of the project. My knowledge and skills have benefited greatly from his experience and commitment. I am grateful for his never-ending inspiration and enthusiasm, which have helped me to flourish in my academic endeavours. I will always be beholden to him for his mentorship, and I sincerely appreciate all the time and effort he has put into me.