Data Narrative-3

Jinil Patel, 22110184
Computer Science department,
IIT Gandhinagar
Gandhinagar, India
jinilkumar.patel@iitgn.ac.in

I. OVERVIEW OF DATASET

The dataset consists of eight different CSV files, each corresponding to a major tennis tournament in 2013, including the Australian Open, French Open, US Open, and Wimbledon, for both men's and women's singles matches. Each file contains data on the individual matches played during the tournament, with a total of 2546 rows and 42 columns for each tournament.

The columns include various statistics for each player, such as the number of sets won, percentage of first serves in, percentage of first serve points won, number of aces served, number of double faults committed, number of winners hit, number of unforced errors committed, and number of net points attempted and won. Additionally, there are columns indicating the round of the tournament in which the match was played and the result of the match, with 'W' indicating a win by Player1 and 'L' indicating a loss by Player1.

There are also columns indicating the number of points won by each player in each set, as well as the total points, won column for each player. Overall, the dataset provides a comprehensive view of the performance of players in major tennis tournaments in 2013, allowing for an in-depth analysis of player performance and comparisons across tournaments and genders.

II. SCIENTIFIC QUESTIONS

1. The player with a higher percentage of first serves in a match is typically considered the favorite to win.
2. The Average number of aces served is greater in the Wimbledon league compared to France Open.
3. Describe the causes of the variations in Net points a player generates due to changes in the court.
4. Players having higher first serve percentage makes less double fault errors.
5. It is quite difficult for a player with a high first serve to win many first serve points.
6. Describe the variation of Breakpoints over the variation of leagues that played in tennis.
7. The top four players in a tournament make less double faults and unforced errors than the average value of each variable, respectively.
8. Can the amount of winner's hits made, unforced errors committed by the player, and double faults committed by the player be used to forecast the outcome of the game?

III. DETAILS OF LIBRARY AND FUNCTIONS

1. **Pandas:** Pandas is a popular open-source library for data analysis in Python. It provides easy-to-use data structures and data analysis tools for handling and manipulating large datasets, including data input/output, data cleaning, filtering, aggregation, and visualization. Pandas is widely used in various industries, including finance, healthcare, social sciences, and engineering, and it has become an essential tool for data analysts, data scientists, and machine learning practitioners. Pandas also integrates well with other libraries such as NumPy, Matplotlib, and Scikit-learn, making it a powerful tool for data science workflows.
2. **Matplotlib.pyplot:** Matplotlib.pyplot is a Python library that provides a variety of functions to create different types of plots and visualizations. It is built on top of the Matplotlib library and provides a simple interface for creating basic visualizations such as line charts, scatter plots, bar charts, histograms, and more. Matplotlib.pyplot is a popular tool for data visualization in the scientific community, and it is widely used in fields such as biology, physics, finance, and engineering. It is highly customizable, allowing users to fine-tune every aspect of their plots to create high-quality visualizations that effectively communicate insights from the data.

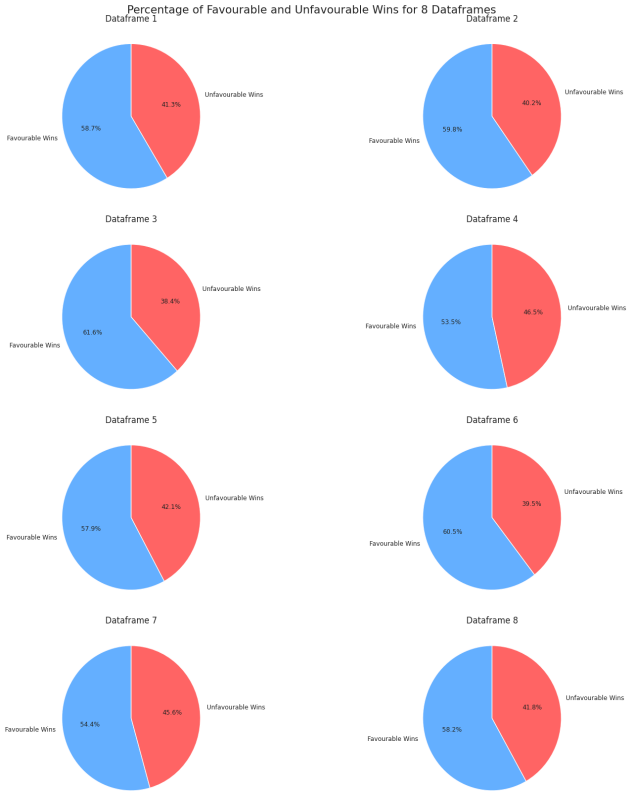
3. **Seaborn:**Seaborn is a data visualization library built on top of Matplotlib that provides a higher-level interface for creating informative and attractive statistical graphics. It provides a variety of visualization types such as scatterplots, line charts, heatmaps, violin plots, and more, and includes features for color palettes, themes, and aesthetics that make it easy to create visually appealing and informative plots. Seaborn is widely used in data science and statistical analysis to explore data and communicate insights to others. It is particularly useful for visualizing complex relationships between multiple variables and for creating publication-ready figures.
4. **pd.read_csv():** to read in CSV files
5. **pd.merge():** to merge DataFrames based on a common column
6. **groupby():** to group the data by tag name
7. **mean():** to calculate the average rating for each tag
8. **idxmax():** to find the index of the maximum value in a Series
9. **value_counts() function:** used to count the number of times each rating appears in the DataFrame
10. **sum() function:** used to calculate the total number of ratings
11. **bar() function:** used to create the bar graph of the rating probabilities
12. **set_xlabel() and set_ylabel() functions:** used to label the x and y axes of the plot
13. **set_title() function:** used to set the title of the plot
14. **show() function:** used to display the plot on the screen.

IV. ANSWERS TO THE QUESTIONS

1. In tennis, the serve is one of the most important aspects of the game. It is the shot that starts every point, and a player who serves well has a significant advantage over their opponent. The serve system in tennis involves a player hitting the ball over the net into the opponent's service box. There are two serves allowed, the first serve and the second serve. The first serve is typically hit with more power and spin and is considered more important because it is the player's best opportunity to win the point outright. If the first serve is missed, the player gets a second serve, which is usually hit with less power and more accuracy.

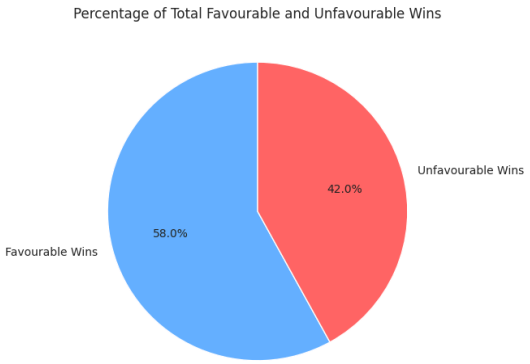
The code provided calculates the percentage of favorable and unfavorable wins based on which player had a higher percentage of first serves in a match. The function ***get_win_count*** takes a dataframe as input and calculates the number of favorable and unfavorable wins based on the first-serve percentage of the players. The ***pie_plot*** function creates a pie chart showing the percentage of favorable and unfavorable wins for a given dataframe.

The code then creates a list ***win_count_list*** containing the favorable and unfavorable win counts for each of the eight dataframes. The code then loops over this list and creates a pie chart for each dataframe showing the percentage of favorable and unfavorable wins. Finally, the code calculates the total number of favorable and unfavorable wins across all eight dataframes and creates a pie chart showing the percentage of total favorable and unfavorable wins.



The percentage of favorable and unfavorable wins for each of the 8 dataframes may be seen in the pie charts. Each pie chart depicts the percentage of wins for the player with the higher percentage of first serves (favorable wins) and the player with the lower percentage of first serves

(unfavorable wins) in various tennis matches represented by the dataframes.



The assumption that the player with a higher percentage of first serves is often thought to be the favorite to win is further supported by the pie chart displaying the percentage of overall favorable and unfavorable wins across all dataframes. The graph demonstrates that, out of all the matches in the dataset, the player with a greater percentage of first serves won 58% of them, while the player with a lower percentage of first serves only won 42% of them.

The data as a whole reveals that the first serve is, in fact, a crucial element in tennis matches and can significantly affect the result of a game.

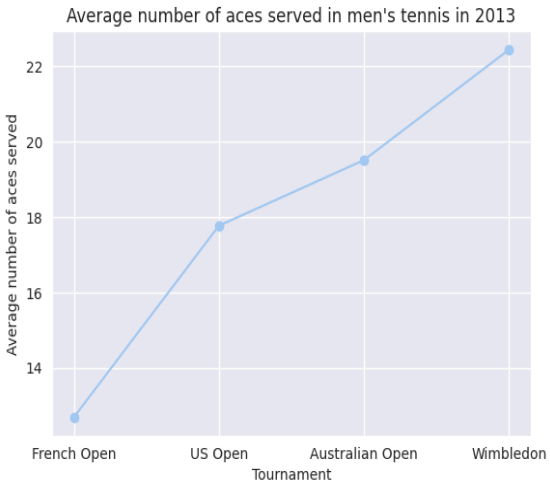
2. Tennis is a game that can be played on grass, clay, or hard courts, among other surfaces. Each surface has distinct qualities that can have an impact on how the ball behaves when it bounces off of it. The manner of play and the tactics that players use during the game can be significantly influenced by these traits.

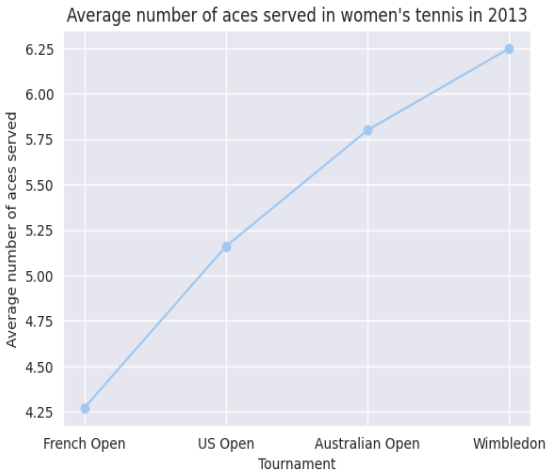
   The kind of surface used during serving can have a big impact on how many aces a player can hit. A serve that lands inside the limits of the opponent's court without being touched by the opposition is known as an ace. Aces are a crucial component of a player's toolkit since they enable them to score rapidly and exert pressure on their rivals.

   Wimbledon's grass courts are renowned for being quick and low-bouncing. The ball bounces lower and faster on grass because there is less friction there than on other surfaces. This makes it more challenging for the opposition to predict the serve and prepare to return it. Players at Wimbledon, therefore frequently serve more aces than in other competitions.

   The French Open's clay courts, on the other hand, play substantially slower than grass courts. Clay has greater friction than other surfaces; therefore the ball bounces higher and slower there. This allows the opposing team more time to set up and position themselves for the serve and return. As a result, compared to Wimbledon, players in the French Open frequently hit less aces.

   Major tennis events like the US Open and Australian Open are held on hard courts.The medium-paced hard courts let players to execute a wide range of shots, including serves that may result in aces. The bounce on hard courts is slower than that of clay courts, while it is more predictable than that of grass courts, giving rivals an opportunity to return serves. A multitude of variables, including the atmosphere, altitude, and player prowess, have an impact on the number of aces served at these competitions. Overall, these competitions tend to have more aces served than the French Open but fewer than Wimbledon. As opposed to the hard courts used at the US Open, the Plexicushion surface for the Australian Open is a little bit slower, which can make it simpler for rivals to return serves.





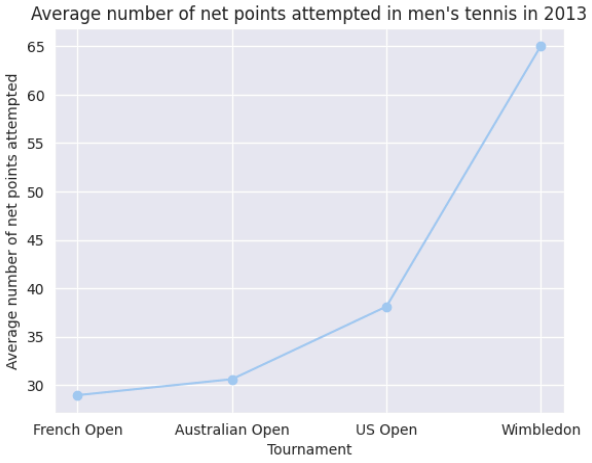
On average, men's tennis players served more aces at Wimbledon than at the US Open and Australian Open, while doing the opposite in the French Open, according to the first graph. The second graph similarly demonstrates that Wimbledon, the US Open, and the Australian Open had the highest average number of aces served by women's tennis players, while the French Open had the lowest average number of aces served. The graph's findings confirm the theory describing how the variation in court surfaces influences the quantity of aces served.

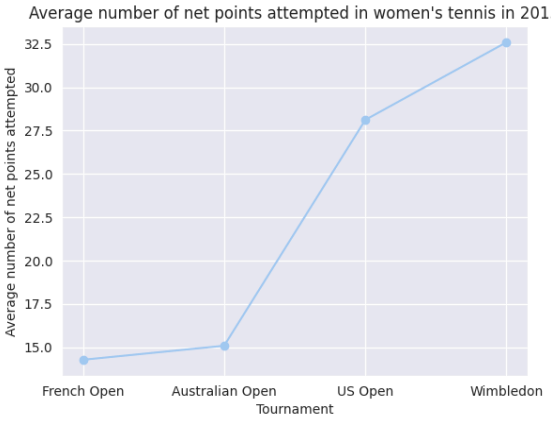
The amount of aces served can be influenced by a variety of additional elements in addition to the court surface, including the players' skill levels, serve speed, and placement. The gap in the amount of aces served between Wimbledon and the French Open, however, is largely due to the differences in playing characteristics between grass and clay courts.

3. Tennis players who hit the ball into their opponent's court but allow it to touch the net before it bounces over are said to have earned a "net point." The player who hit the ball is given the point if this occurs. The point goes to the opponent's player if the ball contacts the net but doesn't cross over. Players can score net points during serves as well as rallies, and they can be very important in determining the result of a match. Gaining net points might make a player feel more in control of the game and have a psychological advantage.

The court's surface can have an impact on net points for various tennis competitions. For instance, the grass court at Wimbledon is typically the quickest and has the lowest bounce. Players may have opportunities to approach the net and attempt volleys or drop shots as a result of this.

Similar to Wimbledon, the US Open and Australian Open are played on hard courts, where players may have more opportunities to smash flatter shots that may be challenging to return. As a result, there may be more net play at these tournaments than at the French Open. Conversely, the slower surface and higher bounce of the clay courts used for the French Open may make it harder to strike winners at the net, resulting in fewer attempts at net play.





The results of the two graphs are consistent with the provided justification. Men's tennis' average number of net points attempted is highest at Wimbledon and lowest at the French Open, according to the line graph, which is consistent with the idea that Wimbledon's grass court fosters more net play. The line graph for women's tennis similarly demonstrates that the average number of net points attempted is highest at Wimbledon

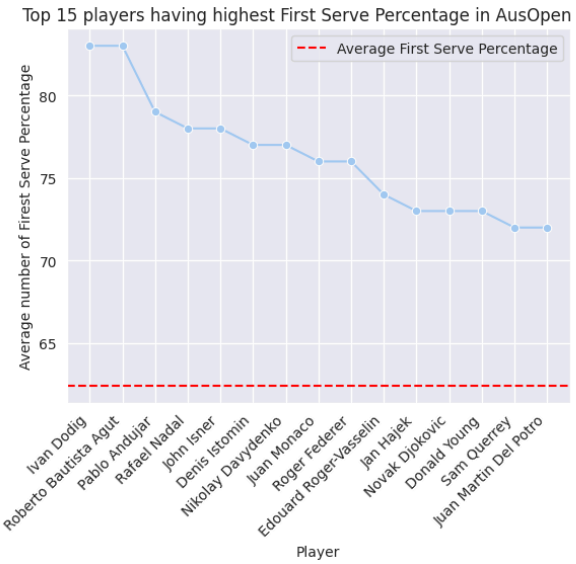
and lowest at the French Open, which is also consistent with the notion that the clay court at the French Open makes it more difficult to hit winners at the net.

4. In tennis, a double fault error occurs when a player makes two consecutive faults while serving, resulting in the loss of the point. A fault occurs when a player's serve fails to land within the boundaries of the opponent's service box, or when the server commits a foot fault by stepping over the baseline or onto the court before making contact with the ball.

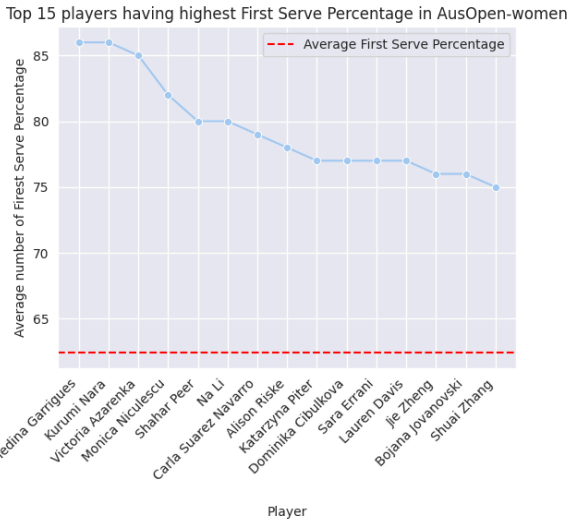
we believe that players with greater first serve percentages commit less double fault errors. Because In general, a player who serves more frequently has better control over their serve, allowing them to accurately place the ball inside the service box. This precision reduces the risk of making a foot fault or serving outside of the court, the two most common causes of double fault errors.

A player with a higher first serve % may also decide to play more cautiously and hit a slower, safer first serve to ensure that the ball enters play rather than a more forceful, riskier serve that may increase the likelihood of a double fault.

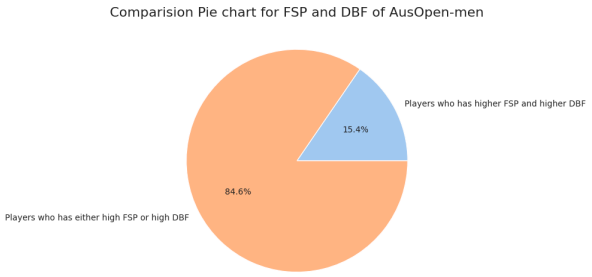
To support the aforementioned hypothesis, the given data from the Australian open men's and women's tournament The top 15 players with the highest first serve % were extracted first, followed by the top 15 players with the lowest double fault errors.

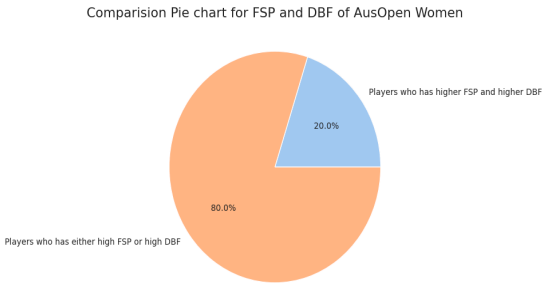


The top 15 players in the Australian Men's Open are represented by the above graph, which shows their first serve percentage.



The top 15 players in the Australian Women's Open are represented by the above graph, which shows their first serve percentage.





From the two pie charts above, we can observe that, for men and women, respectively, 15.4% and 20% of players had high first serve percentages and low double fault errors.

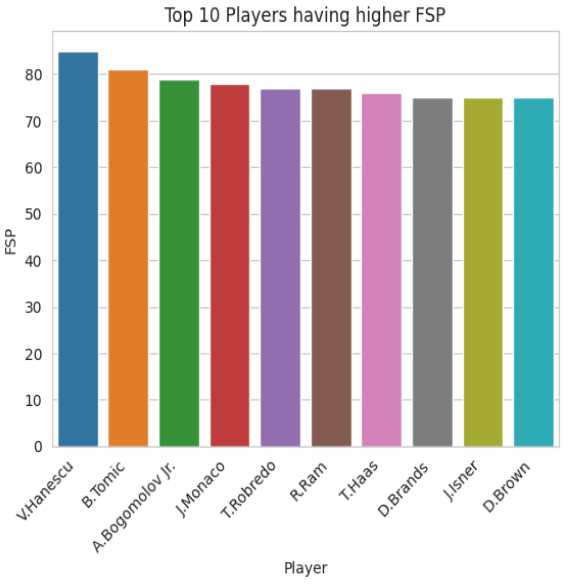
Since there were a very small number of players who possessed both abilities, our first hypothesis was not entirely accurate. However, the players from the Australian men's Open who have both abilities were primarily ranked beneath 10 internationally. It follows that players with high

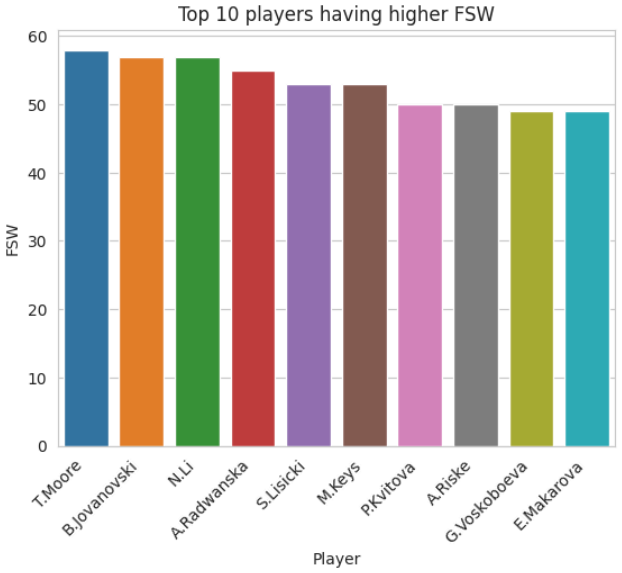
first serve percentages and less double fault errors are those who have high worldwide rankings.

5. We have used the Wimbledon men's data frame to verify it.

```
new_df_1 = df8[['Player1', 'FSP.1']]
new_df_1 = new_df_1.sort_values('FSP.1', ascending=False)
top_10_players_1 = new_df_1.head(10)
new_df_2 = df8[['Player2', 'FSP.2']]
new_df_2 = new_df_2.sort_values('FSP.2', ascending=False)
top_10_players_2 = new_df_2.head(10)
merged_df = pd.concat([top_10_players_1.rename(columns={'Player1': 'Player', 'FSP.1': 'FS
                       top_10_players_2.rename(columns={'Player2': 'Player', 'FSP.2': 'FS
sorted_df = merged_df.sort_values('FSP', ascending=False)
sorted_df.drop_duplicates(subset=['Player'], keep='first', inplace=True)
top_10_players_FSP = sorted_df.head(10)
```

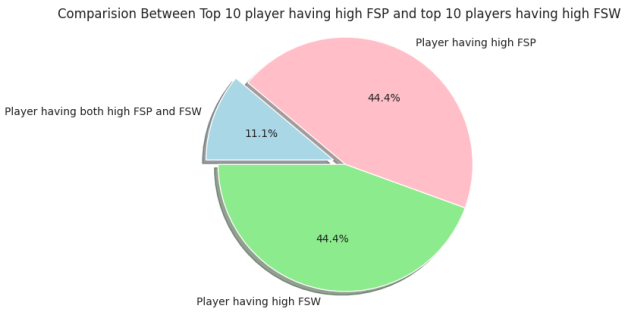
The method chooses from a given dataset the top 10 tennis players with the highest FSP (First Serve Percentage). In order to accomplish this, the top 10 players are chosen after choosing columns for Player 1 and their FSP and sorting the results by FSP in descending order. Then it repeats the process for Player2 and their FSP, joins the dataframes, sorts by FSP in decreasing order, and deletes duplicate player data for FSP values below a certain threshold. The top 10 tennis players with the highest FSP are displayed in the resulting dataframe and The top 10 players with high First Serve Won points were determined using the same procedure.






```
players_FSW = set(top_10_players_FSW['Player'])
players_FSP = set(top_10_players_FSP['Player'])
players_both = players_FSW.intersection(players_FSP)
players_either = players_FSW.symmetric_difference(players_FSP)
print("Players present in both top 10 lists:")
print(players_both)
print("Players present in either top 10 list:")
print(players_either)
```

The top 10 tennis players with the highest FSW (First Serve Won) and FSP (First Serve Percentage) scores are compared in two sets by code. By utilizing the intersection method, it finds the players who are present in both sets and prints them. Additionally, it uses the symmetric_difference technique to find and report the players who are present in one set but not both. The output demonstrates the shared and distinct players between the two sets of the top 10 tennis players with the highest FSW and FSP values.



There are very few players who have high FSP and FSW, as can be seen. this is due to the fact that while in tennis having a high first serve percentage (FSP) is crucial, it is not always the deciding factor in first serve points. First serve points are determined by the quality of the serve, where it is placed, and the opponent's skill on the return. It may not always be the case that a player

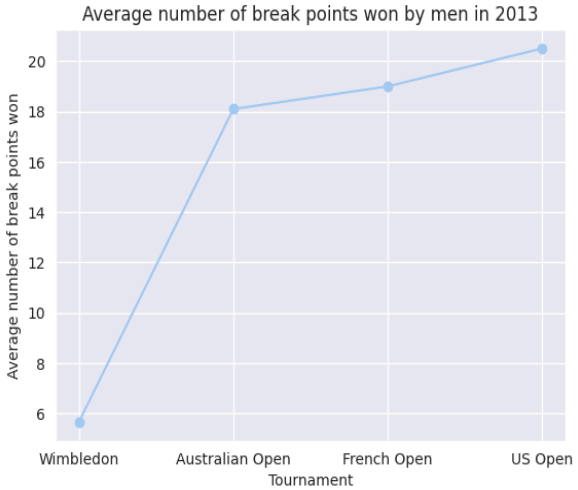
with a high FSP has a strong or accurate serve, which makes it simpler for their opponent to return the ball and win the point. In contrast, a player with a lower FSP might have a serve that is more potent and challenging to return, leading to a higher rate of first serve points won.

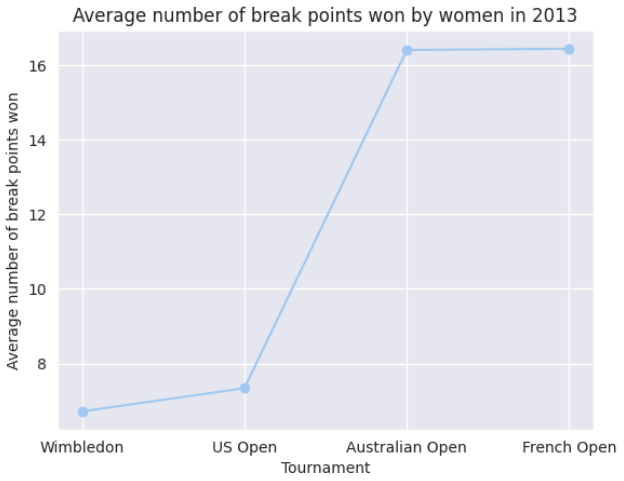
Overall, while having a high FSP is crucial for success in first serve points in tennis, it is not the sole determinant. A player's ability to win first serve points depends on a number of variables, including the quality and location of the serve as well as the opponent's talent in returning the ball.

6. A breakpoint in tennis happens when the player taking the serve has a chance to win the match by taking the following point on the opponent's serve. Because it allows the returning player a chance to take control of the set or match, a breakpoint is a crucial juncture in a match.

Break points typically happen when the player serving makes an error, such as hitting a double fault or an unforced error, giving the returning player the upper hand. If the returning player wins the point, it is called a "broken serve," and they will have won the match. If the serving player wins, they are said to have "saved the breaking point," and the game will proceed.

It's important to note that due to the tennis scoring system, break points are very important. Tennis, in contrast to many other sports, has a special scoring system that favours momentum and consistency. If the score is tied at 40-40 (known as "deuce"), a player must win two straight points to win the game. Otherwise, a player must win four points to win the game. This implies that a single break point can make all the difference in a game, set, or even a match.





The graphs show that there are differences in the average number of break points gained between men and women and between tournaments. The US Open and Australian Open often have higher break point averages for both men and women, and men tend to win more break points than women.

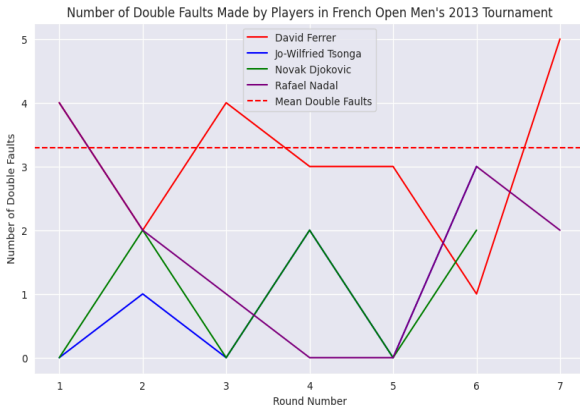
These variations can be attributable to the various court surfaces used in each competition. Hard court surfaces, like those at the US Open and Australian Open, tend to result in speedier games with more reliable bounces, as was already established. Players may be able to be more aggressive with their returns, as a result, creating more break point chances.

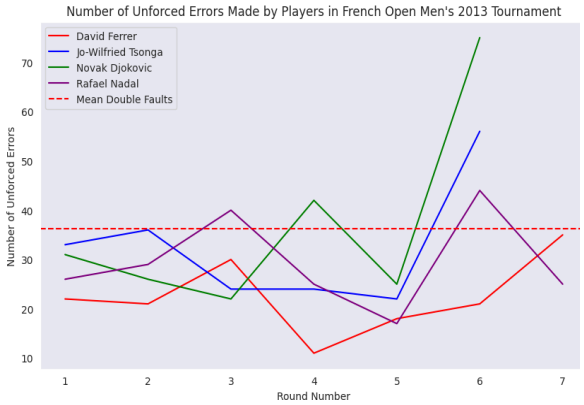
The French Open's clay courts, on the other hand, tend to result in slower matches with higher bounces. There may be less opportunities for breakpoints as a result of players finding it harder to strike wins and induce errors from their opponents.

Grass court surfaces, like those at Wimbledon, can result in low bounces and quicker matches, which might favor serve-and-volley play and aggressive returns. However, because grass court surfaces are variable, it can be difficult for players to remain consistent, which lowers the average number of breakpoints gained.

In conclusion, the type of court surface used at each Grand Slam competition can have a big impact on how many breakpoints players are able to win. This may result in different player tactics and games, which could change the typical number of breakpoints scored among tournaments and between men and women.

7. The top four players at the French Men's Open were Rafael Nadal, Novak Djokovic, Jo-Wilfried Tsonga, and David Ferrer.





The two graphs below, which display double fault and unforced errors, display the Top 4 players in the French Men's Open. As can be observed, the values of the following variable are frequently lower than their mean values in the majority of rounds.

This is due to the fact that the top four tennis players are often the most talented and seasoned players, who have a greater understanding of the game and have spent years perfecting their technique. Less double faults and unforced errors are produced as a result of their improved consistency and accuracy in their serve and groundstrokes. These players also possess a higher degree of mental toughness, which enables them to maintain their composure and commit fewer errors even while under duress. Because of their increased talent and mental toughness over time, it is not surprising that they commit less double faults and unforced errors than the ordinary player.

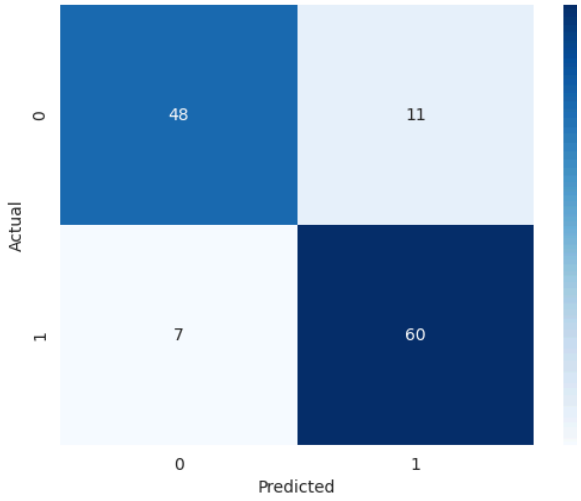
8. Winners, unforced errors, double faults, and breakpoints are all crucial indicators that, taken together, offer insightful information about a tennis player's performance and can be used to forecast a match's outcome. A player's ability to execute successful shots, seize control of the game, and score points is demonstrated through wins. Conversely, unforced errors highlight a player's shortcomings and may serve as a gauge of their physical and mental health throughout the game. Double faults are a significant part of the game since they show a player's dependability and consistency in serving.

   Another crucial indicator for assessing a player's success is breakpoints. When the player's opponent serves back and is just one point away from winning the game, a breakpoint occurs. The player can gain momentum and possibly win the game if they are successful in saving the breakpoint. They risk losing the game and their momentum, which could have an effect on the rest of the match, if they are unable to save the breakpoint.

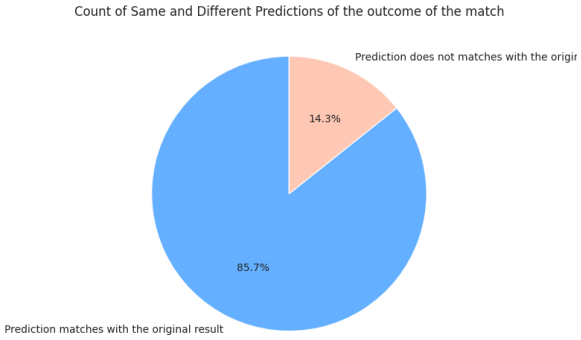
   By combining all of these measures, analysts are better able to identify a player's strengths and weaknesses, gauge their physical and mental condition throughout a match, and anticipate a match's outcome.

```
predictions = []

for index, row in df1.iterrows():
    diff1 = row['WNR.1'] + row['BPW.1']- row['UFE.1'] - row['DBF.1']
    diff2 = row['WNR.2'] + row['BPW.2']- row['UFE.2'] - row['DBF.2']

    # predict the outcome based on the player with the higher difference
    if diff1 > diff2:
        predictions.append(1)
    else:
        predictions.append(0)
```

The code estimates the difference between the winner's hits and breakpoint wins and the unforced errors and double faults for both players in a tennis match by iterating through each row of the DataFrame 'df1'. A diff1 bigger than a diff2 predicts the first player to win (represented by 1), whereas a diff1 less than or equal to a diff2 predicts the second player to win (represented by 0). After that, the prediction is made based on the player with the higher difference. The list of 'predictions' contains these predictions.



A confusion matrix produced by the list prediction is shown in the above graphic. It tends to be a diagonal matrix.



We can see from the pie chart above that 85.7% of the predictions were accurate. Therefore, we can draw the conclusion that by examining the aforementioned player skills, we can somewhat predict the outcome of the game.

V. SUMMARY OF OBSERVATION

1. Wimbledon and the French Open have the highest and lowest average number of aces served per match, respectively.
2. The player with a higher percentage of first serves in a match is typically considered the favorite to win.
3. In men's and women's tennis, the average number of net points attempted is highest at Wimbledon and lowest at the French Open, which is in line with the influence of court surfaces on net points.
4. Tennis players that have a high first serve percentage (FSP) are not more likely to win first serve points (FSW), but there are other criteria that are very important as well, such as the quality and placement of the serve as well as the opponent's skill in returning the ball.
5. In tennis, the average number of break points gained varies depending on the court surface, with harder courts offering more opportunities for break points and clay courts offering fewer.
6. Even though a high first serve percentage may not always translate into fewer double faults, players who possess both skills are likely to be highly ranked. According to data from the Australian Open men's and women's competitions, there aren't many players that succeed in both, proving that it's challenging to master both at once. On the other hand, for players who do have both skills, it might be a sign of their general ability and success on the court.
7. The top four players in a tournament make less double faults and unforced errors than the average value of each variable, respectively.
8. With an accuracy of 85.4%, we can predict the result of the game based on the number of winners hits made, unforced errors made by the player, break points earned, and double faults made by the player.

VI.REFERENCES

1. “Pandas Documentation — Pandas 1.5.3 Documentation,” n.d. https://pandas.pydata.org/docs/
2. “Matplotlib — Visualization with Python,” n.d. https://matplotlib.org/.
3. “NumPy Documentation — NumPy v1.24 Manual,” n.d. https://numpy.org/doc/stable/
4. “Seaborn: Statistical Data Visualization — Seaborn 0.12.2 Documentation,” n.d. https://seaborn.pydata.org
5. GreekForGreeks. "Python Tutorials." GreekForGreeks, https://www.greekforgreeks.com/python-tutorial.

VI. ACKNOWLEDGMENT

I would like to express my deepest gratitude to our supervisor, Prof. Shanmuga, for his invaluable guidance and constructive feedback throughout this project. The data set is accessed from http://lib.stat.cmu.edu/datasets/colleges/ which helped with the data collection and analysis.