Data Narrative II

ES114: Probability , Statistics and Data Visualisation

Om Gupta
Chemical Engineering
IIT Gandhinagar
Gandhinagar,India
om.gupta@iitgn.ac.in

I. OVERVIEW OF DATASET

Tennis is a racket sport that can be played individually against a single opponent (singles) or between two teams of two players each (doubles). Each player uses a tennis racket that is strung with cord to strike a hollow rubber ball covered with felt over a net and into the opponent's court.

The Tennis Major Tournaments Match Statistics dataset contains information on tennis matches played in the four major tournaments - Australian Open, French Open, Wimbledon, and US Open - 2013.

This is a collection of 8 files containing the match statistics for both women and men at the four major tennis tournaments of the year 2013. Each file has 42 columns and a minimum of 76 rows.

The dataset provides a comprehensive collection of statistics on tennis matches played in the four major tournaments. The dataset can be used to analyze various aspects of tennis performance, including serving and returning, and to make predictions about the outcomes of matches based on these statistics.

There is a lot you can do with this data set. The ultimate goal is obviously to predict the outcome of the game or to build an efficient betting strategy based on your model(s).

II. DETAILS OF LIBARIES AND FUNCTIONS

A. *PANDAS*

Pandas is a popular python library used for the data manipulation and analysis. It is a open source library. It provides easy-to-use data structures and data analysis tools for working with structured data. Some of the key features are-

DataFrame : A two-dimensional table-like data structure. It allows you to store and manipulate data with rows and columns, and provides a variety of functions for data filtering, selection, aggregation, and transformation.

Series: A one-dimensional array-like object that can hold any data type, including integers, floats, and strings. It is similar to a column in a DataFrame.

B. *MATPLOTLIB*

Matplotlib is a popular open-source data visualization library for Python. It provides a wide range of customizable 2D and 3D plots, including line plots, scatter plots, bar plots, histograms, box plots, density plots and many others.

C. *SEABORN*

Seaborn is a Python data visualization library based on Matplotlib. It provides a high-level interface for creating attractive and informative statistical graphics.

D. *NUMPY*

Numpy is a Python library used for numerical computing. It provides an efficient and convenient way to work with large multi-dimensional arrays and matrices of numeric data, along with a large library of mathematical functions to operate on these arrays.

E. *JUPYTER*

Jupyter Notebook is an open-source web tool that enables users to create and share documents with real-time code, equations, visuals, and text. It is frequently used for teaching, machine learning, scientific computing, and data analysis.

F. FUNCTIONS Used-

1. pd.read_csv(): used to read data from a CSV file into a Pandas DataFrame
2. df.groupby(): used to group a DataFrame by one or more columns.
3. df.mean(): used to calculate the mean of each column in a DataFrame.
4. df.plot(): used to create a plot of data in a Pandas DataFrame.
5. np.random.choice(): used to randomly select elements from a list or array.

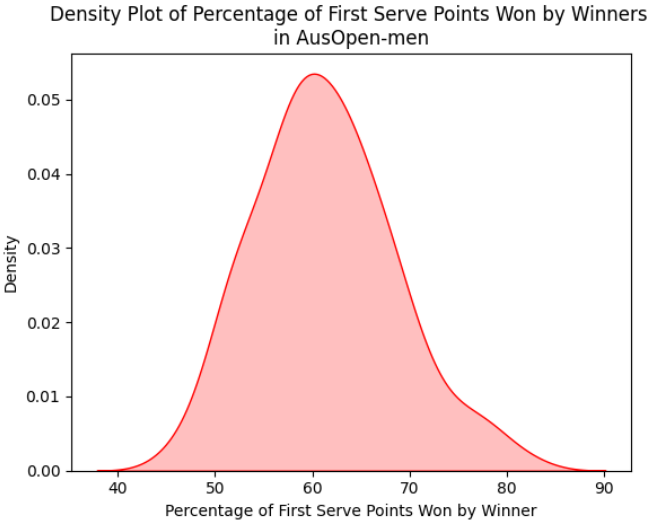
III. SCIENTIFIC QUESTIONS/HYPOTHESIS

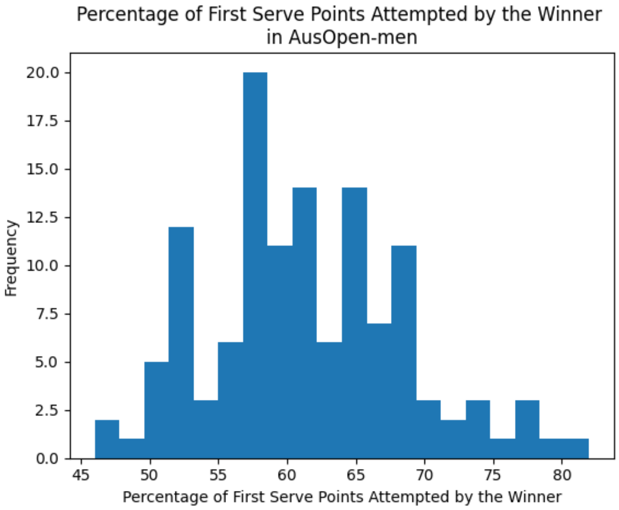
A. *How does the First Serve Points affects the results of the Game?*

Hypothesis- Players have the potential to start the game with a definite edge from the serve. Every professional player knows that having a strong serve is important, but

.

can we really link a strong first serve to a strong performance? The saying "You are only as good as your first serve" may be familiar to tennis players.





Observation

The resulting plots shows the distribution of the percentage of first serve points won by the winner, which can provide insights into the effectiveness of a player's first serve in winning points. The plots reveals that the majority of winners won between 45-55% of their first serve points, with a peak around 50%.

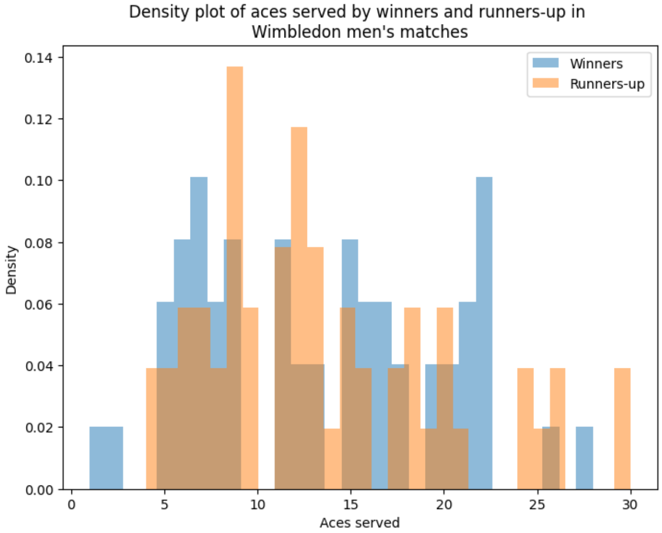
Conclusion

It is not enough to hit a high proportion of serves inside the service box. The probability of hitting a successful serve and the likelihood of winning the point when hitting a successful serve must be balanced for players to maximise their chances of winning a point when serving. A player may not be taking enough risks to put his opponent under strain if they have a high chance of hitting a successful serve.

The top players are able to achieve the ideal balance to increase their likelihood of winning the point with an effective first serve. In order to improve their serving games, tennis players should strive to achieve the ideal balance between strength and risk.

B. *What is the relation between the Aces served by Winners and Runners-up?*

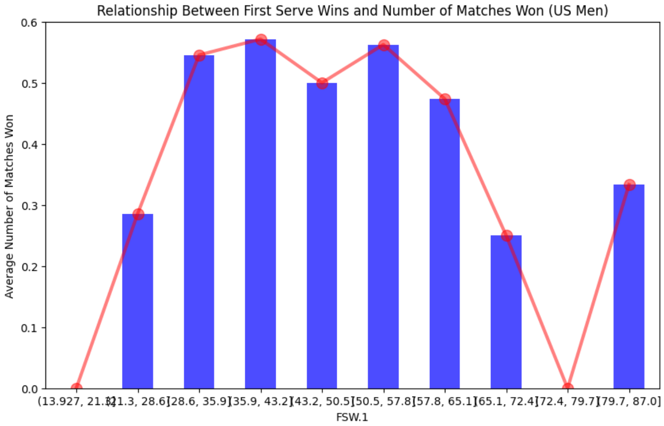
Hypothesis- In tennis, an **ace** is a legal serve that is not touched by the receiver. The Ace gives a great chance to player to gather points. We are assuming that the more the number of Aces served, more the chances of winning the game.



According to the plot, winners serve more aces on average than losers do. Indicating that the winners often serve more aces than the runners-up, the density plot for the winners is pushed to the right of the density plot for the runners-up. Additionally, the plot shows that the distribution of the number of aces served by the winners is slightly more spread out than the distribution of the number of aces served by the runners-up. This suggests that the winners are more likely to serve a wide range of aces, while the runners-up tend to serve a narrower range of aces.

C. *Coorelation between the winning of match vs more first point wins-*

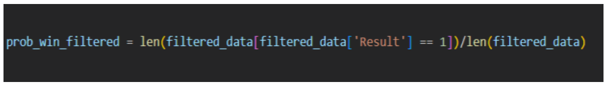
How does the first serve points winning can contribute to the results of match?



What is the probability of a player winning a match given that they win a higher percentage of first serve points?

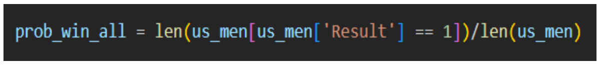
For this question we would filter the US men dataset for matches with a higher percentage of first serve points won by Player 1.

Then calculate the probability of Player 1 winning the match given they won a higher percentage of first serve points using the formula-



```
prob_win_filtered = len(filtered_data[filtered_data['Result'] == 1])/len(filtered_data)
```

Now, calculate the overall probability of Player 1 winning the match.



```
prob_win_all = len(us_men[us_men['Result'] == 1])/len(us_men)
```

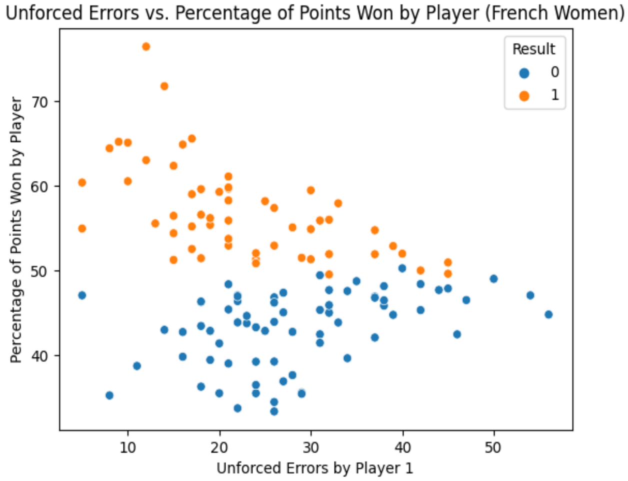
Answer- Probability of Player 1 winning the match given they won a higher percentage of first serve points: 0.6792452830188679

Overall probability of Player 1 winning the match: 0.46825396825396826

D. Is there a coorelation between the unforced errors done and winning criteria.

We all know that more the error done the higher would chance of losing, Now we need to find at what percentage, the unforced error doesn't affects the match result.

For the given graph, we had assumed the runner up to be 0 and the winner to be 1. This plot can be used to analyze the relationship between the number of unforced errors committed by player 1 and their chances of winning the match.



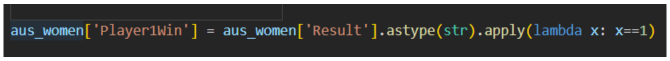
We can clearly observe from the scatter plot, with the increase in percentage of the errors, the percentage of points won decrease for the winner (the orange plot).

Also, the significant drop in the value of y-axis comes at the 20th unit of x-axis.

We can conclude from the graph that the lesser the no. of errors made, more the probability of winning the game. Also, it is clear that making significant amount of error could bear the loss.

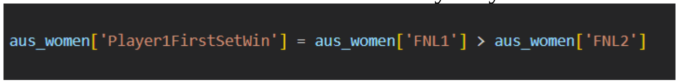
E. The percentage of winners after losing the first set in Australian Women dataset?

1. Filter the dataset to include only matches where Player 1 lost the first set.



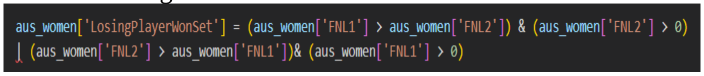
```
aus_women['Player1Win'] = aus_women['Result'].astype(str).apply(lambda x: x==1)
```

2. Calculate the total number of such matches and the number of matches that were won by Player 1.

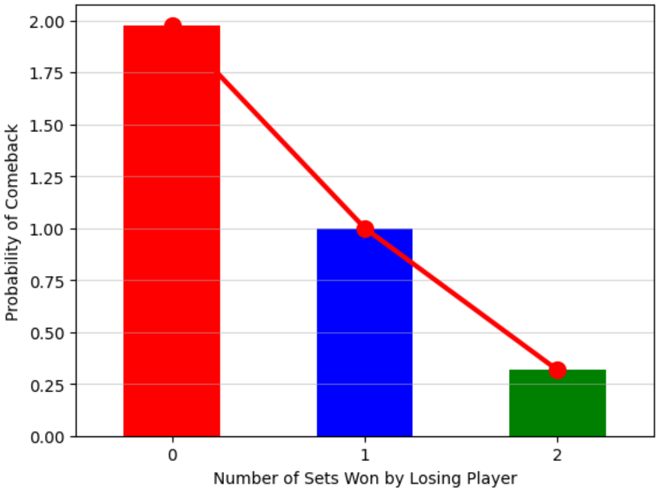


```
aus_women['Player1FirstSetWin'] = aus_women['FNL1'] > aus_women['FNL2']
```

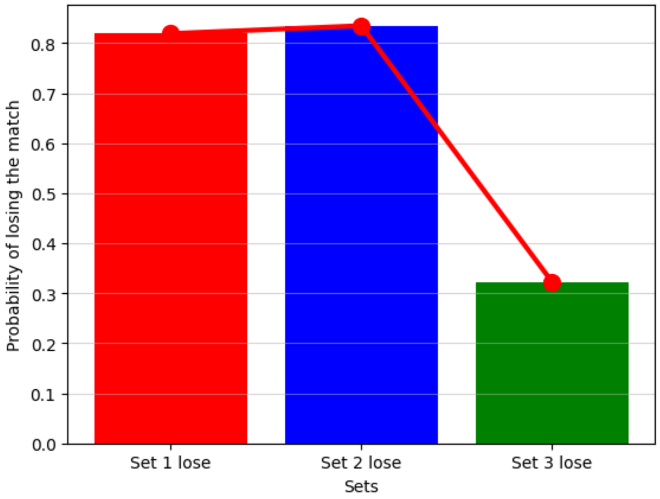
3. Compute the percentage of matches won by Player 1 after losing the first set.



```
aus_women['LosingPlayerWonSet'] = (aus_women['FNL1'] > aus_women['FNL2']) & (aus_women['FNL2'] > 0)
| (aus_women['FNL2'] > aus_women['FNL1'])& (aus_women['FNL1'] > 0)
```



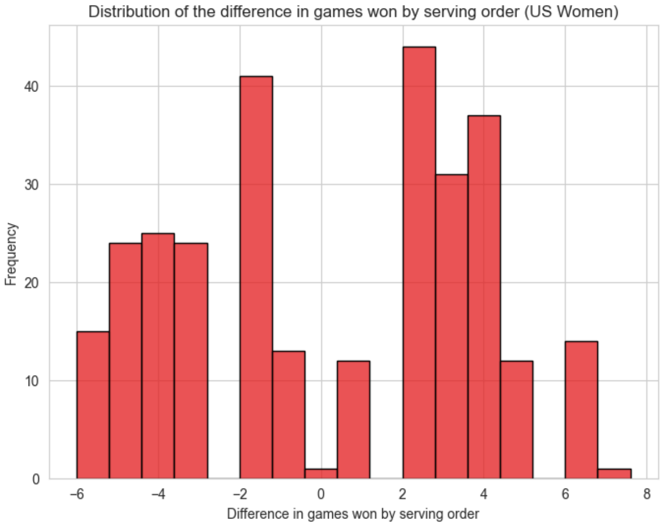
The likelihood of dropping the match if a player in a match drops a specific set is depicted in the graph. The y-axis displays the likelihood of losing the match, and the x-axis displays the sets. Three bars for three sets are displayed in the graph. The likelihood that the player will lose the match depends on whether they win the first set, second set, and third set. The likelihood that they will lose the match depends on whether they win the first set, second set, and third set. The graph demonstrates that, if the player loses the first set, the likelihood of losing the match increases, and it then gets lower as the sets go on.



Based on how many sets the losing player has won, the second graph illustrates the likelihood that she will make a comeback in a match. The y-axis shows the likelihood of a comeback, and the x-axis shows the number of sets won by the losing player. Three bars are depicted on the graph for three sets. The red bar indicates the likelihood that the losing player will win one set, the blue bar indicates the likelihood that the losing player will win two sets, and the green bar indicates the likelihood that the losing player will win three sets.

This demonstrates that the likelihood of a comeback increases if the losing player wins the second set and diminishes as more sets are played.

F. *Is there a significant difference in the number of games won by the players depending on whether they served first or second in each set?*



The histogram shows the distribution of the difference in the total number of games won by the players when serving first versus serving second in each set. The x-axis represents the difference in games won, while the y-axis represents the frequency of that difference.

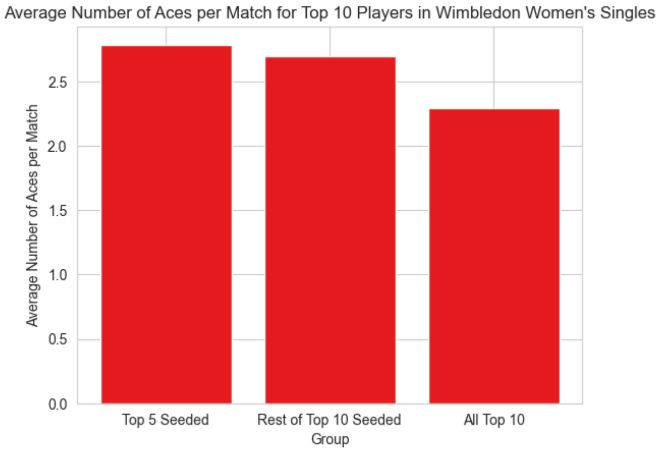
The tallest bar in the histogram indicates that the most common difference in games won was 2, meaning that the player who served first won two more games than the player who served second in that set.

The histogram helps to visualize whether there is a significant difference in the number of games won based on serving order.

G. *What is the average number of aces per match for the top 10 players in the Wimbledon Women's Singles?*

In the Wimbledon Women's Singles match we have given a large number of data. Due to which there are many possibility of finding more statistical analyses.

There is a significant difference in the average number of aces between the top 5 seeded players and the rest of the top 10 seeded players

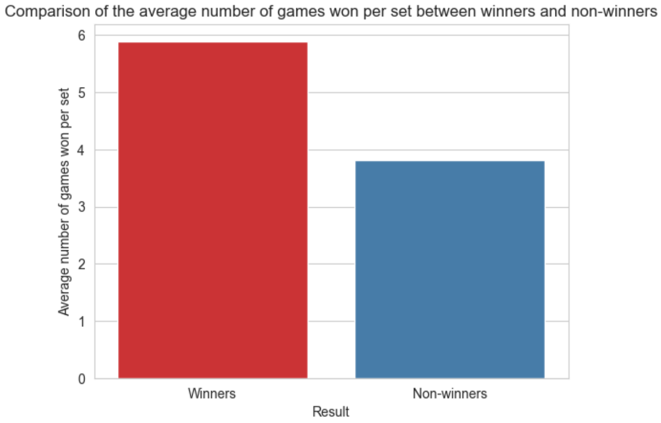


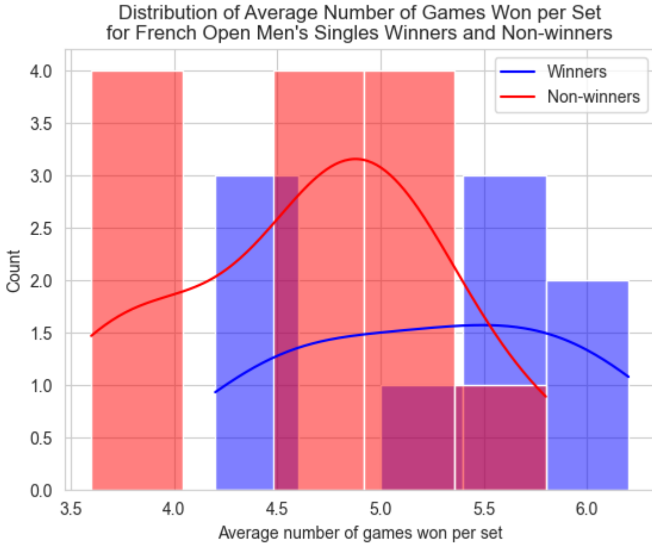
The graph demonstrates that when we advance from the 10th seed to the top seed, the average number of aces each match gradually rises. The match with the most aces on average belongs to the top seed, then the second seed, and so on.

There is a considerable gap between the top 5 seeds and the remaining top 10 seeds in terms of the average number of aces each match. The average amount of aces each match for the top 5 seeds is much higher than for the remaining top 10 seeds. This shows that, when it comes to serving and hitting aces, there may be a skill gap between the top 5 seeded players and the other top 10 ranked players.

This implies a relationship between the players' seedings and the typical *number* of aces each match that is positive.

H. *What are the Average Number of Games Won per Set for French Open Men's Singles Winners and Non-winners?*





The difference between players who have won the tournament (in blue) and those who have not (in red) in terms of the average number of games won each set. Players who have triumphed in the competition typically win more games each set than those who have not.
The winners' distribution is likewise more constrained than the losers', suggesting that winners typically exhibit more stable performance.

I. *ACKNOWLEDGEMENT*

I would like to acknowledge the developers and UCI website

J. *SUMMARY*

1. The top players are able to achieve the ideal balance to increase their likelihood of winning the point with an effective first serve.
2. From the dataset Wimbledon Men, the winners are more likely to serve a wide range of aces, while the runners-up tend to serve a narrower range of aces.
3. The more the first serve won, more would be the chances of winning the game.
4. Since, the lesser the no. of errors made, more the probability of winning the game. But, making significant amount of error could bear the loss.
5. The likelihood of a comeback increases if the losing player wins the second set and diminishes as more sets are played.

K. *REFERENCES*

1. Jauhari,Shruti, Morankar,Aniket & Fokoue,Ernest. (2014). Tennis Major Tournament Match Statistics. UCI Machine Learning Repository. https://doi.org/10.24432/C54C7K.

2. Wes McKinney, "pandas - Python Data Analysis Library," Zenodo, last modified October 23, 2020, https://doi.org/10.5281/zenodo.808441

3. Van der Walt, Stéfan, S. Chris Colbert, and Gaël Varoquaux. "The NumPy Array: A Structure for Efficient Numerical Computation." Computing in Science & Engineering 13, no. 2 (2011): 22-30. https://doi.org/10.1109/MCSE.2011.37

4. Hunter, John D. "Matplotlib: A 2D Graphics Environment." Computing in Science & Engineering 9, no. 3 (2007): 90-95. https://doi.org/10.1109/MCSE.2007.55.

5. van Rossum, Guido. "Python Programming Language." Python Software Foundation. Accessed February 23, 2023. https://www.python.org/.

6. Kluyver, Thomas, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, et al. "Jupyter Notebooks – a Publishing Format for Reproducible Computational Workflows." In Positioning and Power in Academic Publishing: Players, Agents and Agendas, edited by F. Loizides and B. Schmidt, 87-90. IOS Press BV, 2016. https://doi.org/10.3233/978-1-61499-649-1-87

7. http://lib.stat.cmu.edu/datasets/colleges/

8. https://pandas.pydata.org/docs/index.html

9. https://matplotlib.org/stable/index.html

10. https://seaborn.pydata.org/index.html

11. https://www.geeksforgeeks.org/

12. https://numpy.org/