# Data Narrative 3 Report

Name: Mohit Maurya
*Discipline:B-tech M-tech Electrical Engineering*

Roll number: 22110145

## I. Overview of the dataset

The data set contain statistics for both man and women at four major tennis tournament namely Australian-open, French-open Us-open, and Wimbledon for the year 2013. Files for men's and women's statistics are different so all in all, we have 8 files with each file containing a minimum of 76 rows and 42 columns.

In each file there is information like the name of the player and their opponent, the result of the match, rounds, first serve percentage for both players, first serve won for both players, net point attempted and net point won by each player, breakpoint created and won by each, point gain by each players in each set, Unforced error of each player, etc.

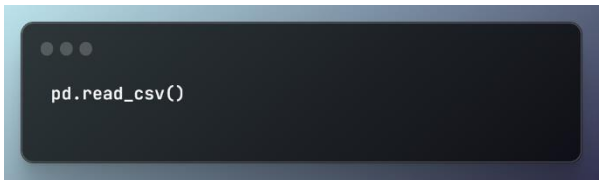
## II. Scientific Question / Hypotheses

1) Do players with a high number of Aces tend to win the match in all tournaments?
2) Is there any relation between winning 1st set and winning the match?
3) Which among the four tournaments for men is tougher when it comes to winning a breakpoint?
4) What is the probability that the player will win the 2nd set, given that he won the 1st set?
5) Is there any change in Unforced Error (UFE) committed by players in Australian Open and French Open?
6) What is the relationship between Net Point Attempted and Net point Won for different tournament?
7) What is the Distribution of Unforced Error in all the tournaments of women?
8) What are the average First serve percentage and average First serve won for women in the Australian Open, French Open, and Wimbledon?

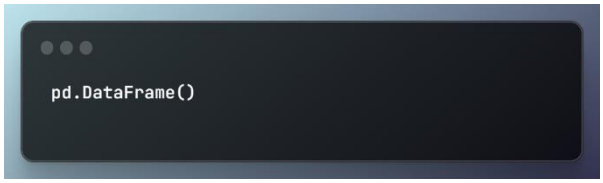
## III. Details of Library and Functions

To answer the above scientific question, I have used three libraries: Pandas, Matplotlib, Seaborn and Numpy.

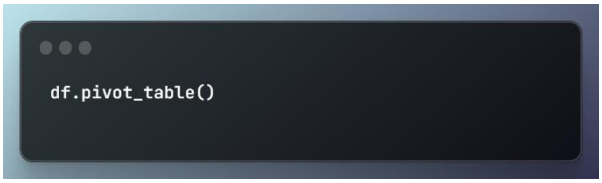
*1) The function used in library Pandas are*

- 


The use of this function is to read the excel file, which is in .csv format; and return the dataframe of the information in that excel file

- 


This function is used to convert dictionaries, lists, 2-D numpy, or series into dataframes

- 


This function takes input like an index and a mathematical function like sum,count, mean, etc., and returns a dataframe whose index is the same as the specified index. That mathematical function is applied to the rest of the columns.

- 
```
.head()
```
This function when used without specifying input, give first five rows of the dataframe; when information which is a number(n), is specified, it provides the first n row in the dataframe.

- 
```
.sort_values()
```
This function, when provided with the columns of the dataframe it sorts the columns in ascending order along with the dataframe. If ascending =True is passed in the input, and if False is passed instead of True dataframe is sorted in descending order

- 
```
.value_count()
```
This function returns the series of frequency/count of the unique values.

- 
```
.index
```
It gives the index of the dataframe as an array

- 

*2) The function in matplotlib are*

- 
```
fig,ax=plt.subplots()
```
It plots the graph of the two values given as input. Also, it takes other inputs as color to change the color of the bar. Also it can create multiple subplots.

- 
```
plt.xlabel()
```
Use to label the x axis

- 
```
plt.title()
```
Use to label y axis

- 
```
ax.text()
```
Use to add text on the graph.

*3) The Function in Seaborn are*

- 
```
sns.set()
```
This help to set color and change the size of all the plot in the code

- 
```
sns.barplot()
```
This function plot the bar graph

- 
```
sns.regplot()
```
This a scatter which comes with liner regression model

- 
```
sns.stripplot()
```
This is function which help to plot scatter plot but it help to plot different categories

*4) The Function in Numpy are*

- 
```
np.where()
```
It provide the array of the values which satisfy the condition sepecified in as an input.

*5) Some general function used are*

- 
```
.sum()
```
Provide the sum of the input.

## IV. ANSWER TO THE QUESTIONS FOR USNEWS DATASET.

*1) Do players with a high number of Aces tend to win the match in all tournaments?*

To respond to this question I first find whether player 1 is the winner or player 2 is the winner then I calculated the winner's aces then I counted the number of winners whose aces are greater than their opponents and stored them in the list. I did this for all the 4 tournaments then I plotted a joint bar graph of total winner, players who are the winner and have more Aces and players who are the winner but have fewer Aces.

From the graph, it was clear that there are more number of player who won the match with a higher number of Aces than those who have won the match but with fewer Aces.

It also shows that winning Aces plays essential role for winning the match.

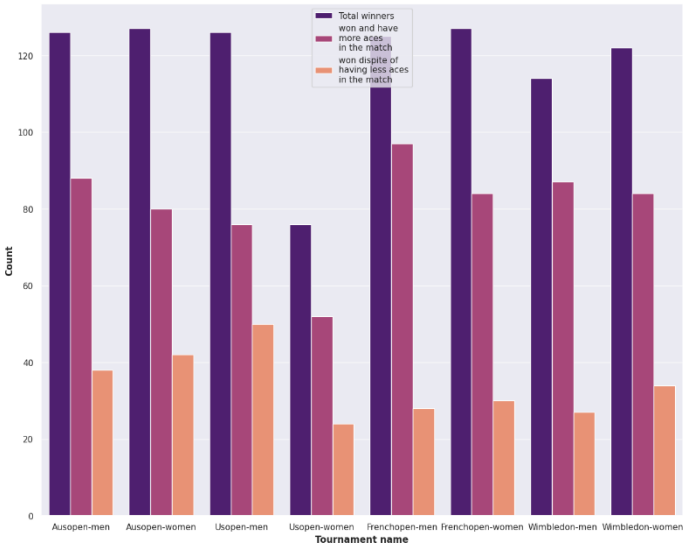


Fig. for question 1

*2) Is there any relation between winning 1st set and winning the match?*

I produced the dataframe in order to provide an answer. In the dataset, I counted the number of player who won the 1st set and the match for each file for both men and women I also counted the number of the player who won the match but loose the 1st set.

The graph showed that for both men and women in all four tournament there were more number of player who

won the 1st set and the match and there were very fewer number of player who lost the 1st set but won the match.

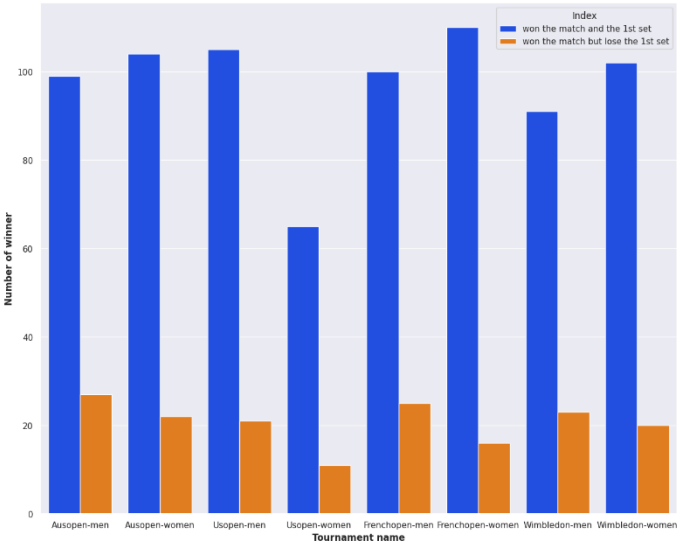



Fig. for question 2

*3) Which among the four tournaments for men is tougher when it comes to winning a breakpoint?*

To answer this question I created the separate data frame for all 4 tournaments (only for mens) then I added all the match which was held in round 1 for all tournament. After this, merges all 4 dataframe and sum up all the breakpoints created by both player and did the same for the breakpoint won.

I found that for Australian Open, French Open, and Us Open, Break point created by the player were less as compared to Breakpoint created in Wimbledon but this totally changes when it came to break point won as for above 3 tournament breakpoint won was way more than breakpoint won in Wimbledon.

This shows that its difficult to create breakpoints in the Australian Open, French Open and Us Open than Wimbledon but it easy to win breakpoints on all three compared to Wimbledon

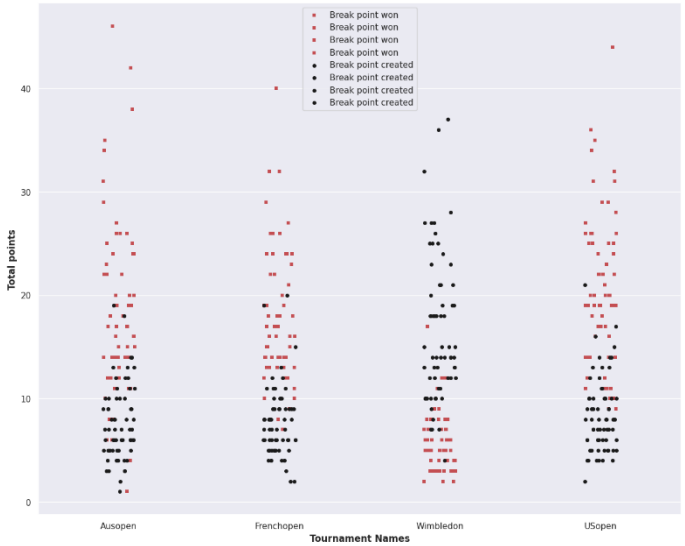



Fig. for question 3

*4) What is the probability that the player will win the 2nd set, given that he won the 1st set?*

To answer this question I separated out the player who won the 1st set irrespective of weather they won the match or not then in that data frame I counted the number of player who won the 2nd set and caculated the probability and plotted the stem graph for both men and women and did this for all the 4 tennis Tournament.

By looking at the graph it was founded that probability of winning the 2nd set is approximately 0.7 for both men and women in all 4 tournament. Also a strange thing was appeared in graph that for all tournament except Us open mens have high probability of winning the 2nd set than women but in Us open this was not the case the reason for this is unknown.

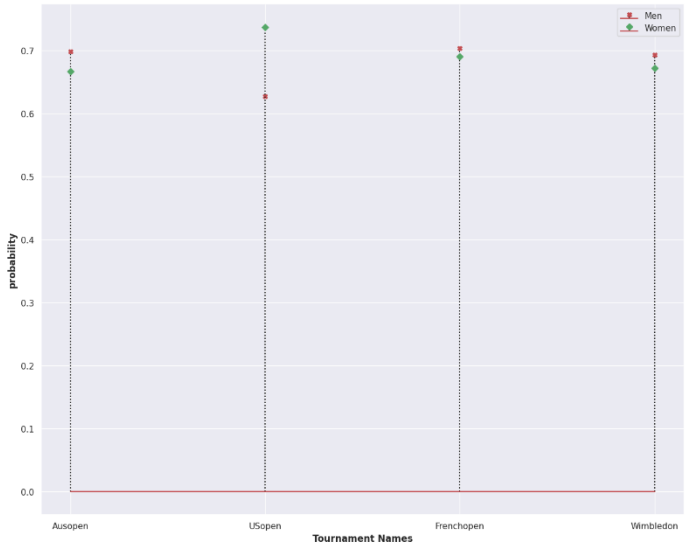



Fig. for question 4

*5) Is there any change in Unforced Error (UFE) committed by players in Australian Open and French Open?*

For this, I different functions and plotted a stem plot of 2 out of 4 tournaments namely the Australian open and French Open for men. Then I created a dataframe for common player who played in both tournaments and found their Unforced error.

The final plot was a little bit inclusive so I found out of 92 player 56 players have committed more unforced error in the French open than the Australian open.

So there is a negative change in the performance of player as they committe more silly errors.

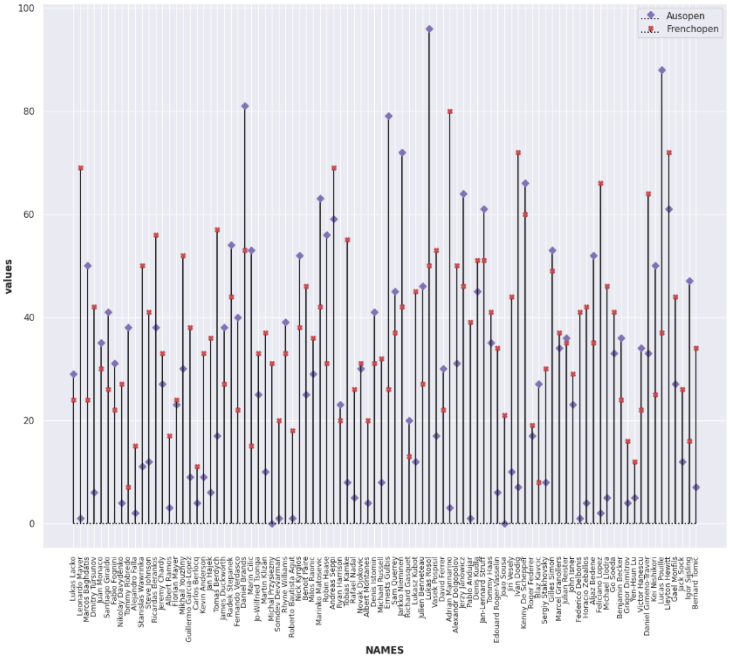



Fig for question 5

*6) What is the relationship between Net Point Attempted and Net point Won for different tournament?*

In order to answer this I ploted the regplot between Net point attempted and Net point won and found that it was linearly proportional for all tournament as which is intentive but strange thing was slop of line for mens Australian open, Us open, French open was more than slop for Wimbledon which says that in all the above 3 tournament Net points won is easier to get even when Net point attempted is low but for Wimbledon Its nearly a 45 degree line which mean a player get more net point won by attempting more Net points

But for women this behaviour is seen for different tournaments as slope of line for French open and Australian open is more than slope for line Us open and Wimbledon open mean for women in French open and Australian open Net points won is easier to get even when Net point attempted is low but for Wimbledon Its nearly a 45 degree line which mean a player won more net point by attempting more Net points

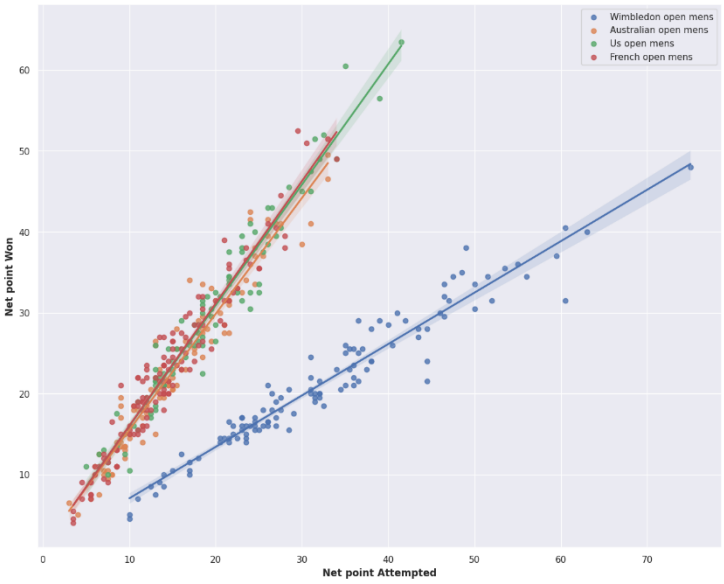



Fig. graph for mens tournament (question 8)

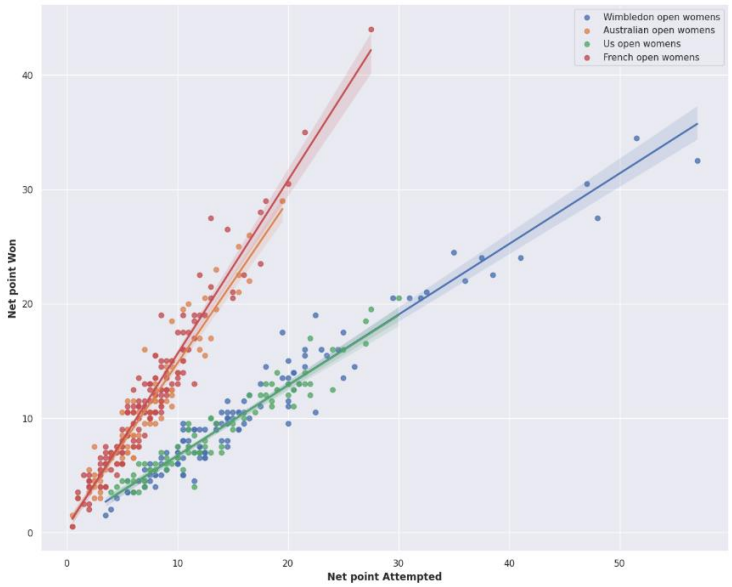



Fig. graph for women tournament (question 8 )

*7) What is the Distribution of Unforced Error in all the tournaments of women?*

In order to answer this question I created the pandas series for all 4 tournament of women and plotted a violin plot of it and made some conclusions.

Some conclusions are that the median of the Australian open is the highest and median of the Usopen is the least mean player make more error Australian open than in Us open another reason for this can be as the data quantity in Usopen is less as compared to the Australian open which resulted in decrease in overall mean. Another observation is for

Australian open the we have the highest and lowest UFE as compared to other also for Australian open and Wimbledon have a wide spread of data.

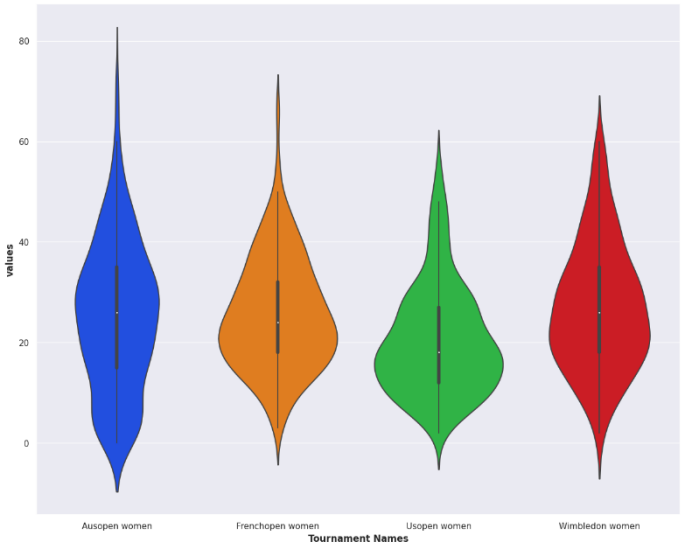



Fig. for question 7

*8) What are the average First serve percentage and average First serve won for women in the Australian Open, French Open, and Wimbledon?*

For this question I created the different dataframe for all 3 women tournament which contain name of all player and their first serve percentage and first serve won. Then took out the common player from all 3 tournament and founded the average of first serve percentage and first serve won.

I plotted the bar graph of this data and found that there is a large variation the both the averages also there are two player with very high first serve won average as compared to others.

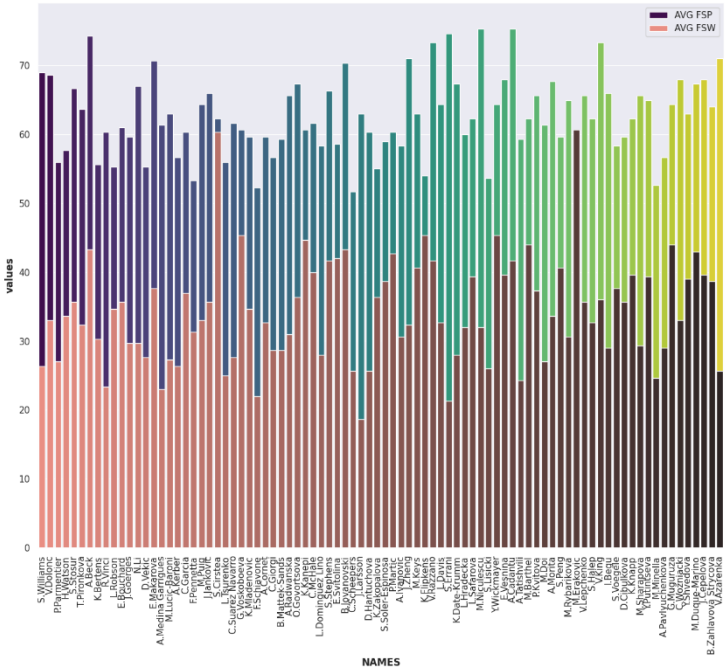



Fig. for question 8

## V. Summary of the Observations

In conclusion the daaset offer the stastical data for the 4 major tennis tournament happened in 2013. From the analysis of data we can draw a certain conclusion

Some conclusion are player with the higher number of Aces in the match tend to win the match mean Aces contribute a major role in gaining a point against the opponent. There is a high correlation between winning the 1st set and winning the match as majority of winner have won their 1st set. Also there is a very high probability that a player who win their 1st set will win their 2nd set also. In Wimbledon its easier to create a break point and harder to win a breakpoint as compared to others tournament. Player commites more silly error in French open then in the Australian open. Also there is a liner relation between Net point attempted and Net point won. There is much variation in UFE for women in 4 tournament with the Australian open have the highest median. Average first serve is around 60 and average first serve won is around 30 for womens who played all 4 tournaments.

These were some of the conclusion drawn from the dataset,

## VI. References

*1) "Matplotlib 3.7.1 Documentation#." Matplotlib documentation - Matplotlib 3.7.1 documentation. Accessed March 27, 2023.* https://matplotlib.org/stable/index.html

*2) NumPy documentation. Accessed March 27, 2023.* https://numpy.org/doc/

*3) "Pandas Documentation#." pandas documentation - pandas 1.5.3 documentation. Accessed March 27, 2023. https://pandas.pydata.org/docs/*

*4) "Pandas Tutorial." GeeksforGeeks. GeeksforGeeks, March 17, 2023. https://www.geeksforgeeks.org/pandas-tutorial/*

*5) "Scipy Documentation#." SciPy documentation - SciPy v1.10.1 Manual. Accessed March 27, 2023. https://docs.scipy.org/doc/scipy/*

*6) "W3Schools Free Online Web Tutorials." W3Schools Online Web Tutorials. Accessed March 27, 2023. https://www.w3schools.com/*

*7) "An Introduction to Seaborn¶." An introduction to seaborn - seaborn 0.9.0 documentation. Accessed April 20, 2023. http://man.hubwiz.com/docset/Seaborn.docset/Contents/Resources/Documents/introduction.html*

8) *Create beautiful images of your code. Accessed April 20, 2023. https://ray.so/*

## VII. Acknowledgements

1) *Thanks to all TA's for hepling me in DATA NARRATIVE.*
2) *Thanks to Piyush singh.*
3) *Thanks to Hari balaji.*
4) *Special thanks to Mithlesh tandon.*
5) *Special thanls to Famida sayyed.*