# ES114 Probability & Statistic
# Data Narrative-3

Archit Dhakar 22110031

***OVERVIEW* —The Tennis Major Tournament Match Statistics dataset contains information on Tennis matches held in 2013 in 4 different countries. The dataset comprised of Men and Women's matches with detailed information of statistical data such as Number of rounds, result, first serve percentage and number of Aces won etc. The result of first player wins the match is shown by 1 and second player wins the match is shown by 0. Each dataset consist of 7 rounds. This dataset can be used to analyze the performance of players and other factors such as cognitive motor skills of players across various countries.**

***Keywords— FSP.1: first serve percentage for player 1 , ACE.1: Number of aces won by player 1, UFE.1 Unforced errors committed by player 1, NPA.1 : Net points attempted, WNR.1 : Number of winners, DBF: Double fault errors.***

## I. SCIENTIFIC QUESTIONS/ HYPOTHESIS

1. Create a graph with the correlation coefficient between a player's first serve percentage and their first serve win percentage. Explain what this plot represents w.r.t Australia Men's match 2013 dataset.

2. Player's playing mindset and mental balance plays an important role. Create a covariance matrix with the following random variables: break points created by player1,break points won by player 1, break points created by player 2 and break points won by player 2. Explain what does covariance matrix signifies.

3. Good player's generally plays well in both defensive and offensive side with aggressive shots which have higher chance of errors. Show that unforced error plays a important role in winning the match in Tennis. Explain the above statement using some valid plots and facts.

4. From the Wimbledon Women's Tennis match show that which factors or features from the dataset which helped the player who played in round 7. Compare her success factors with other player who lost in first round using appropriate plots.

5. Is there any correlation between number of winners and unforced error done by a player. Justify the statement using Linear regression for US open Women's dataset.

6. Show the division of Average double faults committed by players in different rounds. Use pie chart to signify this data. Also discuss the fact that Average double faults committed decreases with increase in rounds.

7. Find the probability that for a selected player wins a match if they have higher number of winner point that his/her opponent. Show the above probabilities for different number of winner points using appropriate plot.

8. Player's motivation plays an important role in sports tournaments. Using appropriate plots show that less first serve won percentage directly affects the second serve won percentage. State the reason for above fact.

## II. DETAILS OF LIBRARIES & FUNCTIONS

Libraries and functions used to visualize the data are:

1. NumPy library: It is used to analyze the tabular and csv data more easily.
2. Seaborn library: To plot probability density function.
3. dropna function to drop '*' values from the dataset.
4. Matplotlib library: It is used to visualize the dataset in visual form by creating different plots such as pie chart, bar graph, histogram etc.
5. Pandas Library: It is used to create the DataFrame from the given csv files and read the bigger dataset in easy way. It is also used to visualize dataset with the help of Scatter plot, line graph etc.
6. Scipy library is used to use linear algebra and other math functions in pyhton.
7. read_csv function to read the csv files in Python.
8. .head() and .tail() to read some important values of data
9. sorted function to sort the columns of dataframe.
10. Sklearn library : It is used to analyze data for the machine learning based algorithms.

## III. ANSWER OF QUESTIONS

1. We know that in Tennis match first serve and points won in first serve plays a important role to win the match.

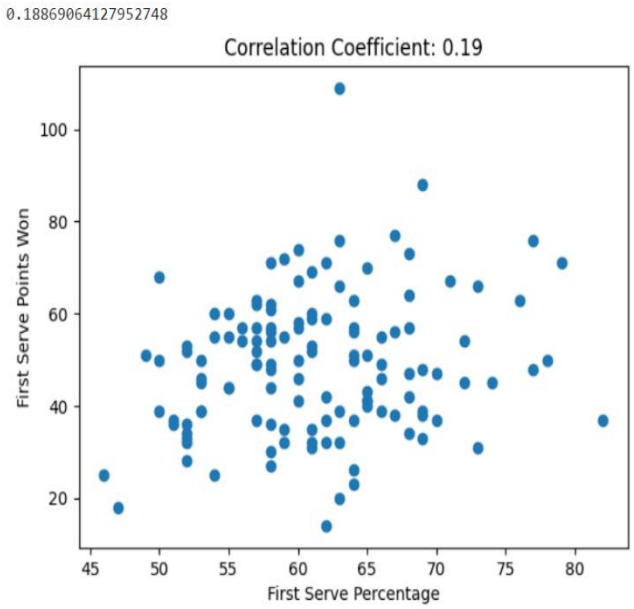


Fig. 1. Scatter plot representing the relation between First serve percentage and First serve won

From the above Scatter plot we can observe that correlation Coefficient is 0.19 which shows that both factors are positively related to each other. But as the number is very small so we can say that even if a player got the chance to serve first instead of that the chance of winning that serve was very less.

Also from the scatter plot we can see that points are densed in center which show that for approx. 50 percentage of serves players are able to won. This show that playing mindset and endurance of player is the deciding factor in winning the first serve.

2. We know that Covariance matrix shows the covariance between two random variables. In our data set we can use the concept of covariance between two features of French Tennis match. Here is the Covariance matrix for Four features of French Men's 2013. These features vectors are Break points created by player 1, Break points won by player 1, Break points created by player 2 and break points won by player 2.

```
Covariance matrix
[[ 6.27380645  9.28116129 -0.97103226 -1.51503226]
 [ 9.28116129 26.34232258 -3.094      -3.44716129]
 [-0.97103226 -3.094       6.468       9.60206452]
 [-1.51503226 -3.44716129  9.60206452 25.61651613]]
```

Fig. 1. Covariance Matrix where each element in the matrix shows the Covariance of any of two above features.

Let M be the Covariance matrix. If we observe the $M_{21}$ element of matrix, it shows the covariance between Break point won by player 1 and Break point created by player 1. The value is 9.28 approx. which shows the both the features are highly correlated and if the number of break points created increases, break point winning percentage increases rapidly. This factor also shows the Human Psychology of player that if we get more number of chances in any field we tend to do our work more accurately.

Now Lets observe the $M_{23}$ element of matrix. We can see that covariance for break points won by player 1 and break points created by player 2 have -3.09 covariance which is negative. This value shows that if player 1 won more number of break points it create the pressure on other player and as a result player 2 is not able to create good number of break points. The factor here is motivation, player 1 has less pressure compare to 2 so break points winning probability increases for him.

Accordingly we can observe other elements from the matrix and show the dependency that how performance of one player directly affects the performance of other player in French Men's match 2013.

3. It is the fact that if a player commit more number of errors which are just due to his less attentive playing or other managing factors known as Unforced errors. Here we will analyze the Australia Open Women's match 2013 that more the number of unforced errors committed by players less is the chance of winning the match. We will observe using two plots where one is comparison bar graph and other is scatter plot showing the Unforced errors of players.

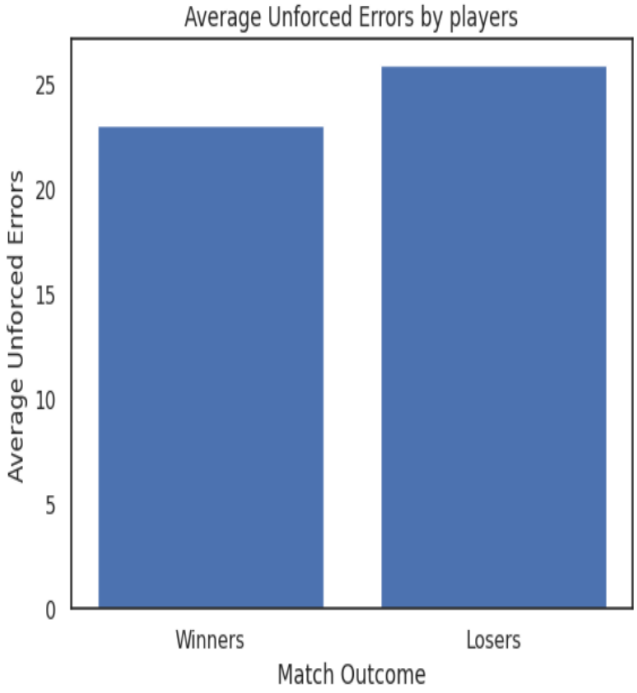


Fig. 3. A comparison bar graph representing the average unforced errors by winning player and loser player.

From this graph we can easily observe that player who won the match commit less number of Unforced errors which increase their motivation to play accurately. Meanwhile players who commit more number of Unforced errors are in pressure and commit more mistakes and hence the probability of losing the match increases.

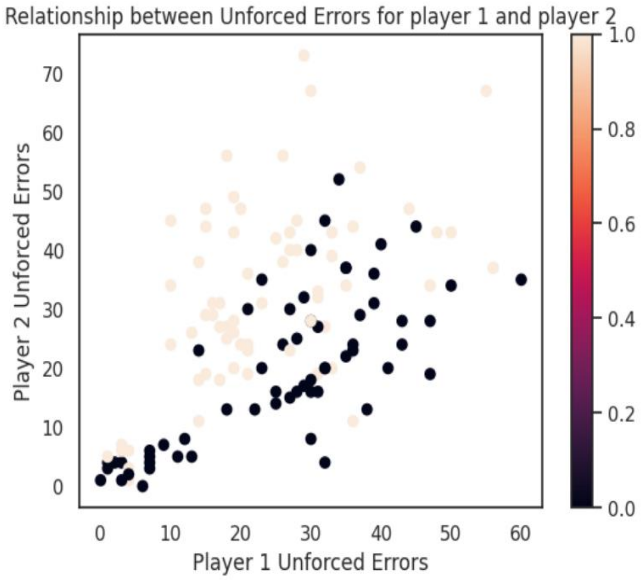


Fig. 4. This is the scatter plot representing the unforced errors committed by Australia Women's Tennis players during 2013 match.

From this Scatter plot we can observe that on an average player who lose the match commits more errors as the density of points is more in upper side of plot.

Meanwhile players who won the match have more point density on center of plot. These players also commits decent number of unforced errors but less then above one.

4. From the Wimbledon Women's match dataset we can observe that first serve , winning point, net points and other factors are common features of player who played in round 7. So first using appropriate code we find the players who played in round one and seven. Now let's visualize these features using a comparison bar graph plot.

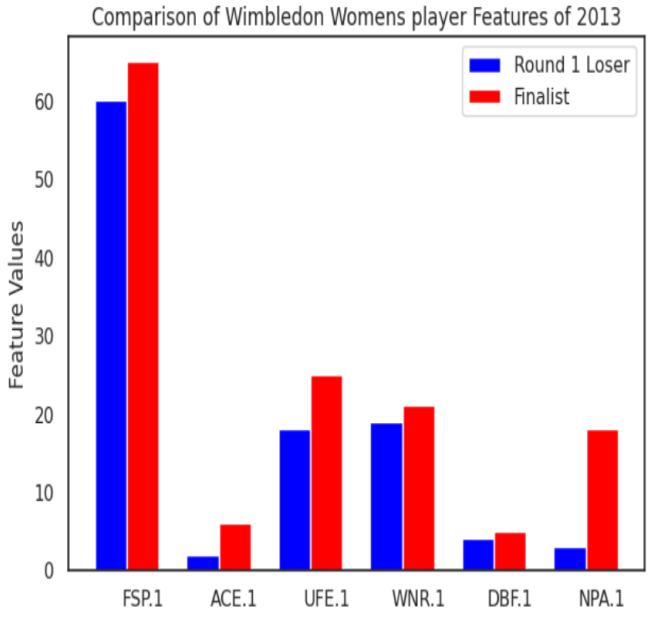


Fig. 5. A comparison bar graph plot showing the important features of both round one and seventh player.

From this graph one can observe that finalist player exceeds almost in all aspects. Net point plays very important role in overall score of player, here we can observe that player who lost in round 1 has less ability to win net points. Also Ace points are also the one of the deciding factors in the Tennis match because it creates the pressure on the opponent player that she is not even able to receive the serve of player.

Hence these all were the deciding factors in the match played in Wimbledon. Also we can observe that Unforced error were more in number of finalist player because when they play aggressive shots in tennis then there are fair chances that they can commit a Unforced error. So these all were worth noting observations on the dataset.

5. In Tennis match, a winner occurs when a player unable to touch the ball with their racquet before it bounce twice in the court. And Unforced errors are the errors committed with players own careless mistake. Using the linear Regression on the US Men's data set we can observe the visual image.

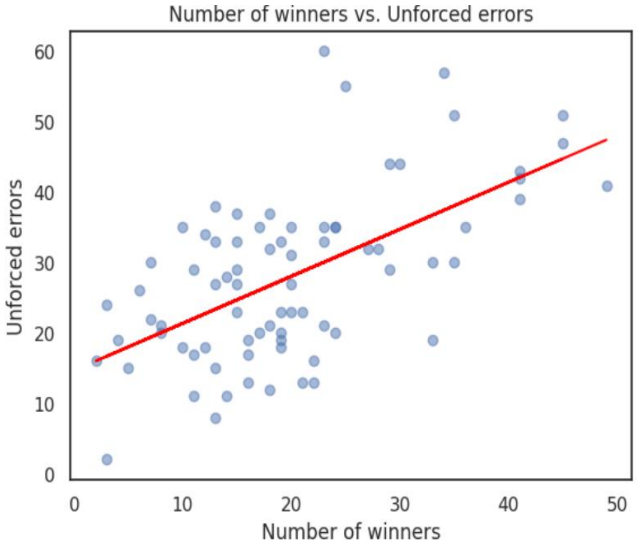


Fig. 6.Linear regression analysis for unforced error and Number of winners in US Women's Tennis match 2013.

Coefficient of determination: this shows that how both parameters depends on each other.

Slope of regression line: If slope that line Is positive that it shows that if value of one increases other's value increases.

Coefficient of determination: 0.35

Slope of Regression line: 0.67

From the image and these values we can say that both of these quantities fairly depends on each other. The reason for this fact is if a player wins more number of winners means he has played exceptionally well shots both in defense and offense. When a player plays aggressive and fast shots then it's naturally that he will commit some mistake in those risky shots. So in conclusion we can say that number of winners really depends on the unforced errors. linear regression line is better way to visualize because line clearly helps to map the data points.

6. We said that a player have committed double fault if he is not able to receive the serve of opponent for continuous for two times. In US Women's data set have seen that there is unusual trend in the average double faults committed by players in different rounds. First using graph let's visualize the dataset.

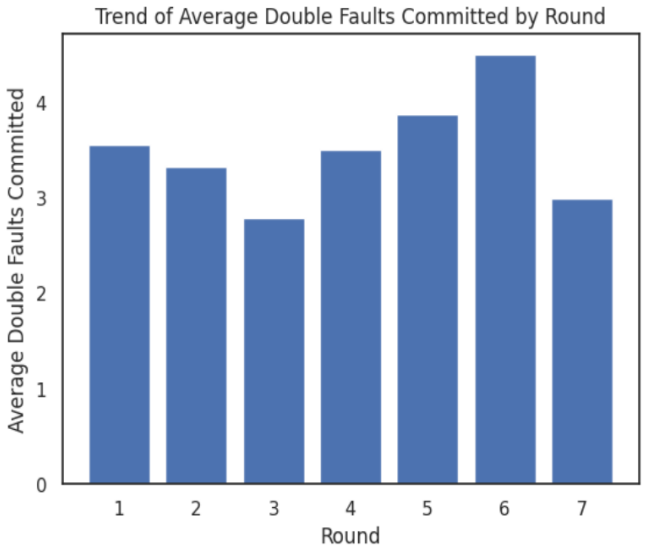


Fig. 7.Bar graph plot showing Double faults in each round.

From this data set we can observe that when player are in their staring or end of matches commit these errors. One of the reason is anxiety because initially in the beginning of tournament due to anxiety some players commit more mistakes. Also when they have played fair amount of matches then in last second round due to mind changing thoughts and nervousness their performance decreases and hence they commit more no of Average Double faults in last rounds.

7. Number of winners are always advantage to the player. SO the player who have more number of advantage will have more chance of winning the match then his/ her opponent. Using code for probability we have found from French Women's match dataset the probability that a given player have the probability of 0.67 of winning the match if he/she has more number of winning points. We can observe this using the below plot.

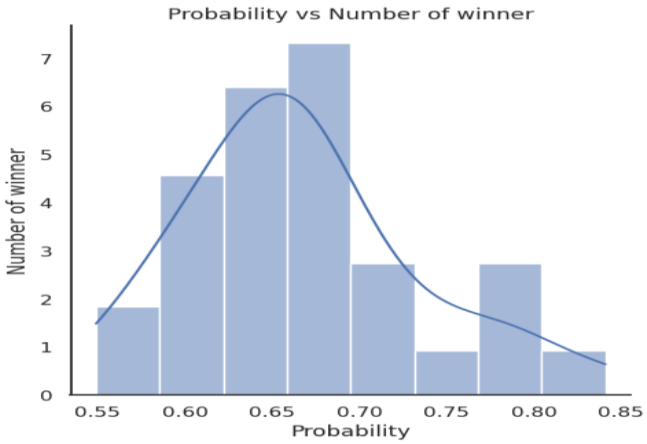


Fig. 8.Plot showing the relation between winning probability and number of winners in French Women's Tennis dataset of 2013 matches.

We can easily observe from this probability density kind of function that peak of graph is high where number of winners are more. This shows the above said fact. From this plot we can find the probability when number of winner were 3 and probability was approx. 0.60. Similarly we can observe for others too.

8. Motivation and consistency plays an important role in player's overall improvement. If some player win more points in first serve then in second round he has confidence to won the serve with more probability. Meanwhile if a player's first serve is itself is bad then there is higher chances of wasting the second serve. With the help of scatter plot let's visualize these facts.

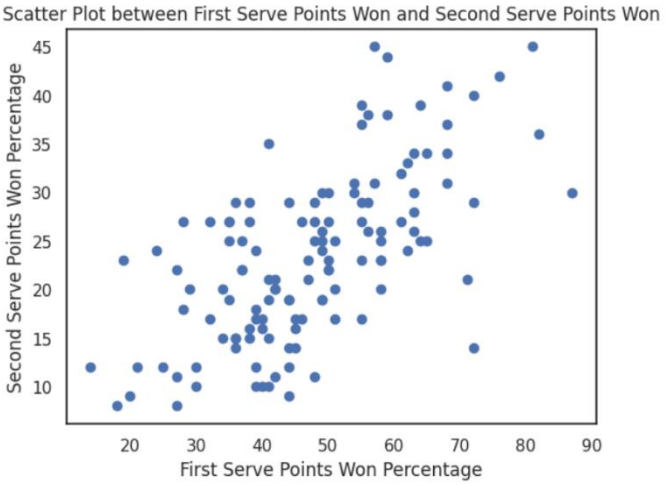


Fig. 9. Scatter plot representing the variation of first serve won and second serve won in US Women's Tennis match 2013.

From this plot we can easily observe that data points are highly correlated to each other. For less number of First serve won, Second serve won percentage is less. While for greater number of first won percentage second won percentage is more. Hence Scatter plot justify the above statement of confidence and motivation.

## IV. SUMMARY OF OBSERVATIONS

In the Tennis dataset provided for different countries there were many factors which can directly help in improvement of players if they pay attention correctly. There were many factors such as first serve win percentage which help to increase the confidence of player while playing if this percentage is high. Number of winners, number of aces and number of net points won plays a main role in the whole journey of matches. These factors acts as advantages to players.

We also observe in the data set that in starting of tournament player commits more number of Double faults because of factors like anxiety. Unforced error was also the factor which change the trajectory of player. If player is good in both defensive as well as aggressive shots then he/ her might won more number of winners point. These players usually commits more unforced errors. We have seen this face using Linear regression. We also find the probability that a player wins because of his first serve won percentage. Using the feature analysis of Finalist player and loser played we found some plots which can help player to improve on those points and do better in next tournament. We

also have discussed that how control of pressure and other mental factors plays an important role in Tennis match.

## V. Refrences


[1]. To plot the bar graphs in python
[2]. To open the CSV file in python.
[3]. To remove * values from dataset.
[4]. To plot line graph using matplotlib.
[5]. Scatter plotting in python using pandas dataframe.
[6]. Seaborn library to plot coloured graphs.


## VI. Acknowledgment


I would like to express my special thanks of gratitude to Professor Shanmuga for his guidance and support to evoking critical thinking in data analysis.


Date: 22-04-2023

Name: Archit Dhakar