Data Narrative

Analysing the data of different tournaments of tennis

Ruchit Jagodara
Student of CSE department
Roll no. 22110102
IIT Gandhinagar
Gandhinagar,Gujarat
ruchit.jagodara@iitgn.ac.in

***Abstract*— This data narrative aims to provide the data analysis done by Ruchit Jagodara on various tournaments of tennis. This data will give you the full scientific analysis of data on the basis of different types of scores which are generally considered in tennis. Also, this data analysis is done in such a way that every person can understand that.**

A. Overview of the Dataset

The dataset, ‘Tennis Major Tournament Match Statistics’ provides the data of different tennis tournaments. In that data, they have divided scores in different categories like net points, Ace points, winner points and with these they have provided the total points of both players. They have also provided the data regarding points in all sets of the match. Player’s skill-based data (i.e., unforced error and double faults) is also provided in this dataset. Moreover, they have given data of eight different tournaments in which four are men tournaments and four are women tournaments.

B. Scientific Questions/Hypotheses

I. What is the probability that a person who is finishing the game in a smaller number of rounds is doing less unforced error as compared to the average unforced error done by a person who plays more rounds to win?

II. Is a good first serve is enough to build a good score?

III. Does the performance of the player increase as round increases?

IV. Is there any relation between the total net points attended and total net points won by a player? Is it affecting the total score of a player?

V. Is there any effect of round on the ratio of break points won and break points created? Analyse the same thing but using the ratio of break points won and unforced error done by that player.

VI. What is the probability of winning a match after losing the first and second set in that match?

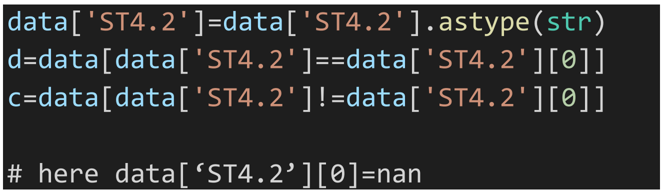
VII. Winner points of a player should be more if opposite player does more unforced errors.

VIII. Is there any major effect of ace on the first serve win?

C. Answers to the Questions

I. What is the probability that a person who is finishing the game in a smaller number of rounds is doing less unforced error as compared to the average unforced error done by a person who plays more rounds to win?

It is obvious that if you are not doing mistake than you will win in few sets, only. To prove that vice versa is also true, first of all we have to take the data in which 4th and 5th set had not played. We can carry out this process using the following sequence of code,(We are using “AusOpen-men-2013.csv” data for this analysis.)

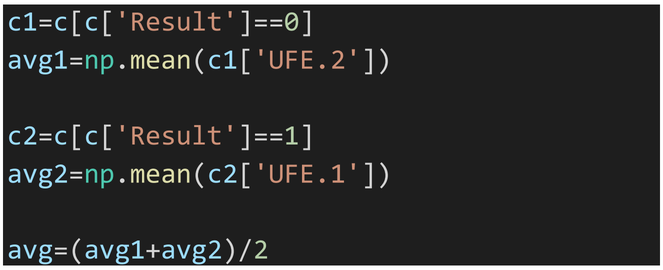


```
data['ST4.2']=data['ST4.2'].astype(str)
d=data[data['ST4.2']==data['ST4.2'][0]]
c=data[data['ST4.2']!=data['ST4.2'][0]]

# here data[‘ST4.2’][0]=nan
```

Here, we have taken the data of “AusOpen-men-2013.csv” dataset and we have stored it in the variable named data using pandas.read_csv() function. Here, after applying above code we will get two variables d and c, where d contains the data in which more than 4 set had been played and c contains data of those matches in which only 1st, 2nd and 3rd set were played.

Now, we need an average value of unforced error made by the player who is winning after playing 4th or 5th set. So, for that again we wrote nearly same sequence of code as above.

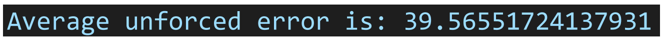


```
c1=c[c['Result']==0]
avg1=np.mean(c1['UFE.2'])

c2=c[c['Result']==1]
avg2=np.mean(c2['UFE.1'])

avg=(avg1+avg2)/2
```

Here, first of all we found the winning person and then we got a mean of unforced errors of all that person. So, now we have average of unforced error which is ,



```
Average unforced error is: 39.56551724137931
```

Now, we will see that how many data is there in d with unforced error less than the avg (we will see the unforced error of that player who has won that match). We will count total number of players that satisfies these criteria. Now we will apply probability equation,

$$probability = \frac{Total\ number\ of\ desired\ players}{Total\ number\ of\ players}$$

So, with the help of this equation we will get the value of probability as 0.9701492537313433. So, as probability is so high, we can say that it is better to perform well in first sets.

II. Is a good first serve is enough to build a good score?

If a player does a good serve than the probability of getting an ace on it increases so overall first serve win increases. So, first of all, we should check if there is an effect of first serve percentage on the first serve win. So, for that, we will plot a graph between first serve percentage and first serve win for both the players. (We are using "AusOpen-women-2013.csv" data for this analysis.)

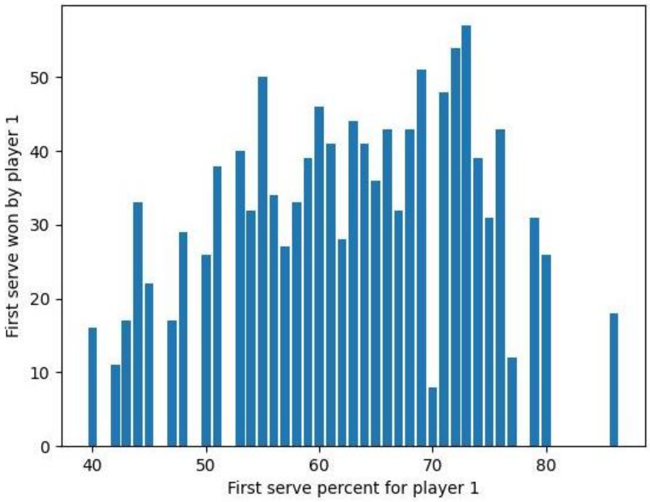


Fig. 1 plot between first serve percentage and first serve win for player 1

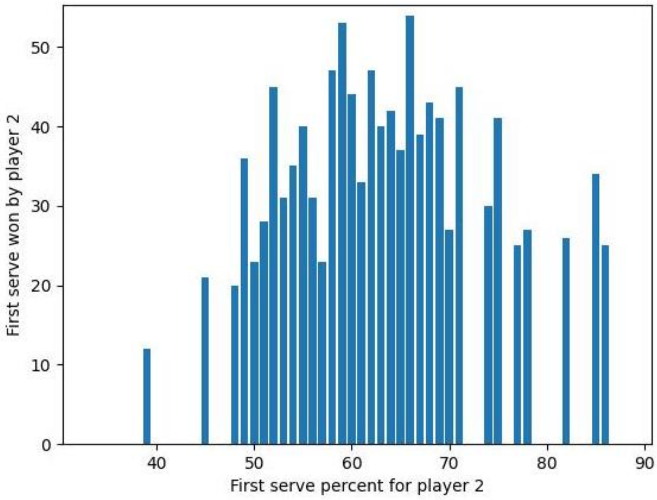


Fig. 2 Plot between first serve percentage and first serve win for player 2

Here, from both Fig. 1 and Fig. 2 you can see that first serve percent does not follow a linear trend with respect to first serve win so from here we can say that it only depends on the skill of the player that he is able to do the first serve win or not.

Now, we have to see that if one does more points by scoring first serve win, then how much that affect on total number of points. So, to observe this, we made a scatter graph between the first serve win and total points made by both players. So, with the help of following code we will get the desired graph,

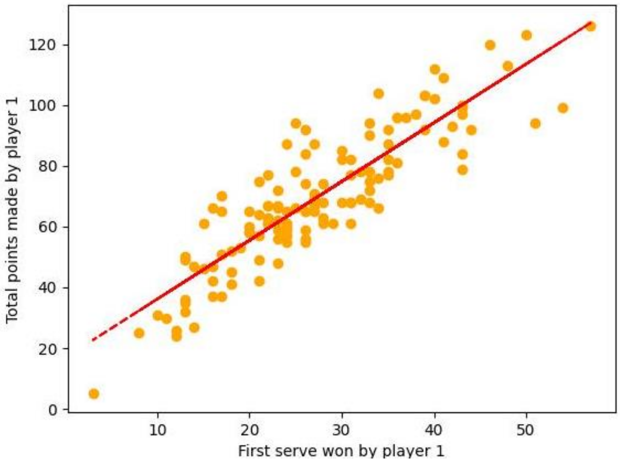


Fig. 3 scatter plot between first serve won and total points earned by player 1

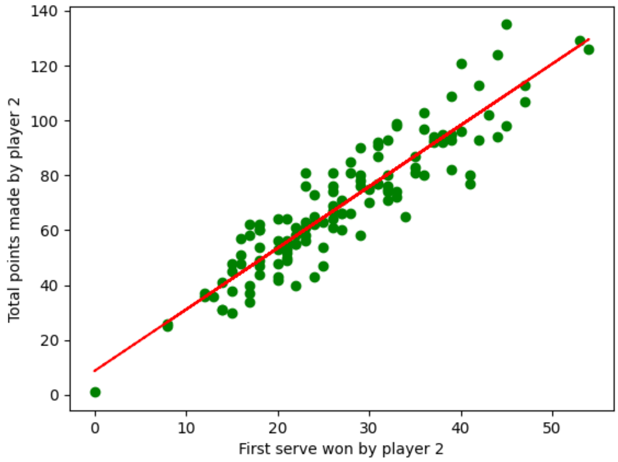


Fig. 4 scatter plot between first serve win and total points earned by player 2

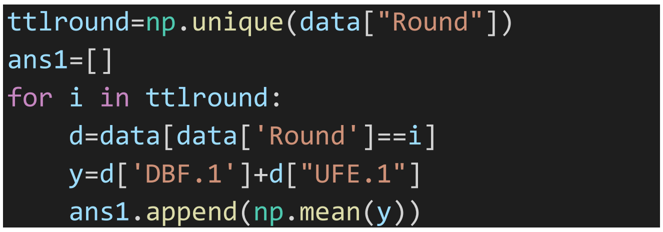
Here, in Fig. 3 and Fig. 4, the red line shown is representing the regression line. We can see that total number of points is also increasing with almost same rate as first serve win points, so first serve win points may be contributing most of points in total points won by a player. So, to obtain that we will divide the first serve win and total points won by a player and then we will do an average of that. So, the result, we are getting is,

Average percentage score of first serve win in total point made by player 1 is: 39.35490575138401 %

Average percentage score of first serve win in total point made by player 2 is: 38.96030990206541 %

III. Does the performance of the player increase as round increases?

It is generally seen that when a person reaches to a higher round like finals, semi-finals, his/her performance will be increased as compared to the other rounds. So, to find this, we have considered the unforced error and double fault errors done by a player in each round. To find out this, we have to differentiate the whole data into small ones which have data of a particular round and after that we can calculate our desired values one by one for each round, so for that we have used below code, (We are using “FrenchOpen-men-2013.csv” data for this analysis.)



```
ttlround=np.unique(data["Round"])
ans1=[]
for i in ttlround:
    d=data[data['Round']==i]
    y=d['DBF.1']+d["UFE.1"]
    ans1.append(np.mean(y))
```

With the use of above code, we will get the average value of sum ofdouble fault and unforced error for player 1. Similarly, we will get the value for player 2, also. And we will plot a pie chart for these data.

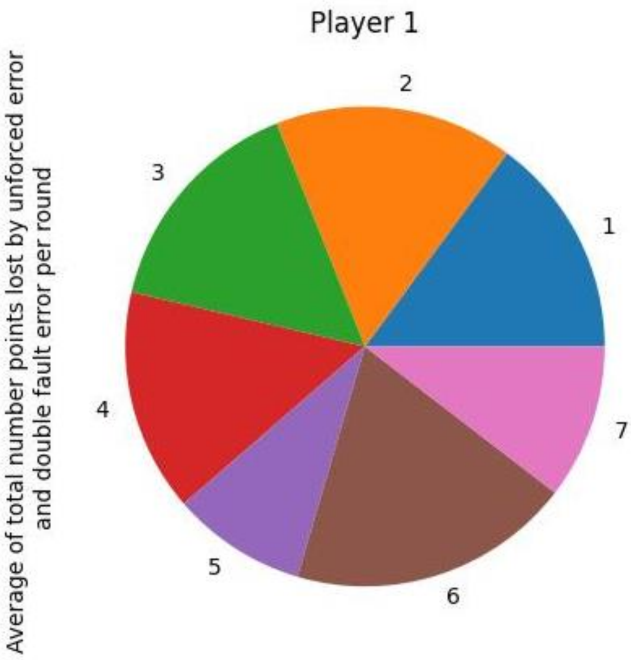


Fig. 5 Pie chart for the roundwise average points which are lost by either unforced error or double fault (for player 1)

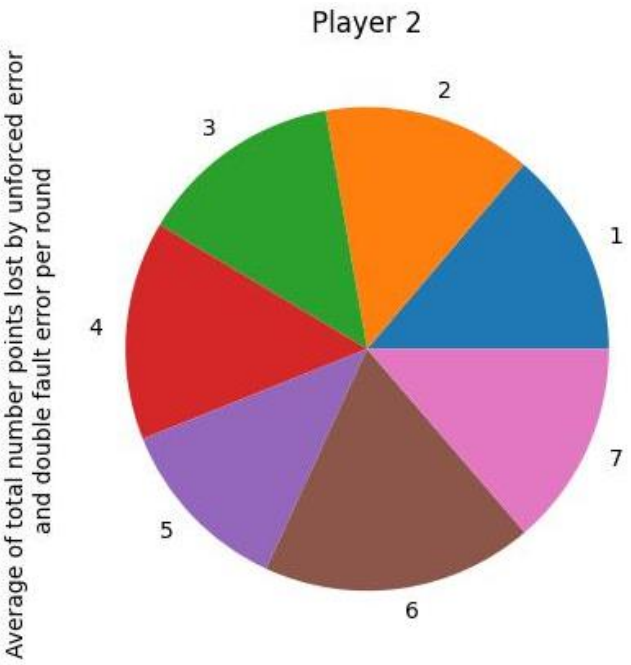


Fig. 6 Pie chart for the roundwise average points which are lost by either unforced error or double fault (for player 2)

In Fig. 5 and Fig. 6, we can easily see that there is no specific trend which is beeing followed by the points of unforced error and double faults, so from this data, we can say that mistakes done by a player in a match reamins nearly the same for all the matches irrespective of the round in which it is beeing played.

IV. Is there any relation between the total net points attended and total net points won by a player? Is it affecting the total score of a player?

Net points are the most important factor for a match because it is fully based on the skill of the player. First of all, we need to see that if there is any relation between net points attended and net points won by a player. So, for that, we have plotted a graph between the total net points won by a player and total net points created by a player, (We are using “FrenchOpen-women-2013.csv” data for this analysis.)

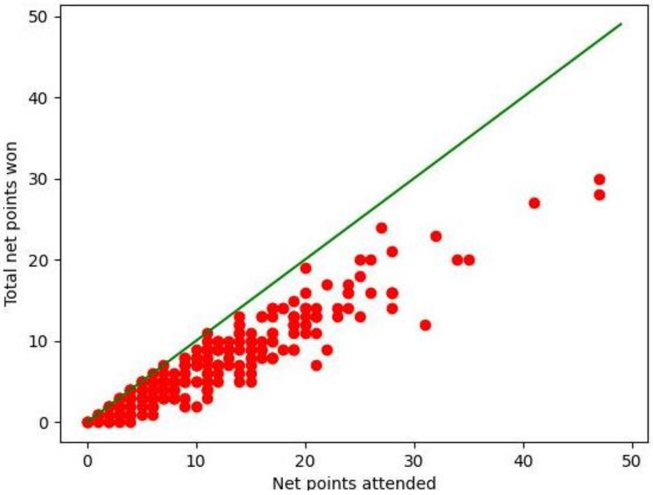


Fig. 7 Graph between net points attended and net points won

Green line in the Fig. 7 shows that that is the maximum value of net points won for a particular value of net points attended because net points won can not be greater than net points attended. In Fig. 7, we can see that, when value of net points attended is very low, scattered points are very near to the line and as the value of net points attended increases, the distance between the line and points increases. So, we can say that when net points attended are low, probability of losing a net point is very less. So, from this analysis, we can say that a player should attend a smaller number of net points to win the match.

We will find the correlation factor to see that at which extent net points win and total net points attended are related. So, with the help of numpy.corrcoef() function, the correlation coefficient between these two vectors comes out to be 0.95030957, which is very high, so we can say that net points win and net points attended are highly related to each other.

So, to check the above analysis, we will plot a graph between net points attended and total points earned by a player.

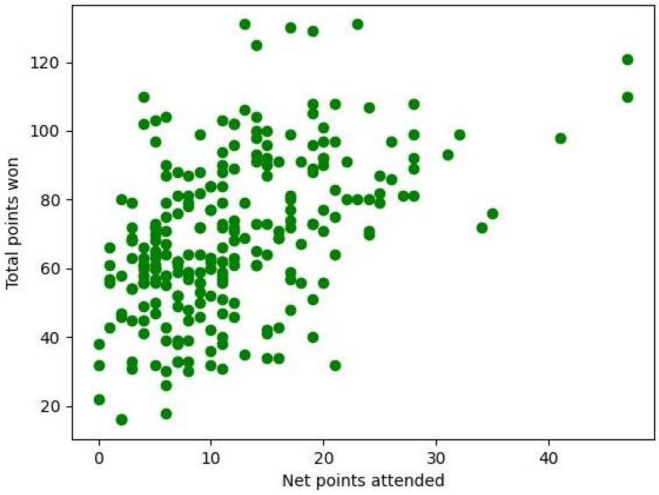


Fig. 8 Scatter plot between net points attended and total points

In Fig. 8, we can easily see that there is no particular trend that this graph is following. So, from this we can say that net points contributing very less in total points earned by a player.

V. Is there any effect of round on the ratio of break points won and break points created? Analyse the same thing but using the ratio of break points won and unforced error done by that player.

Here, break point plays a very important role in the match because if you win the break point it means that you won that set. So, to find if round affects the ratio of break points created and break points won or not, we will plot a pie chart which contains average of ratios of break points won and break points created. (We are using "USOpen-women-2013.csv" data for this analysis.)

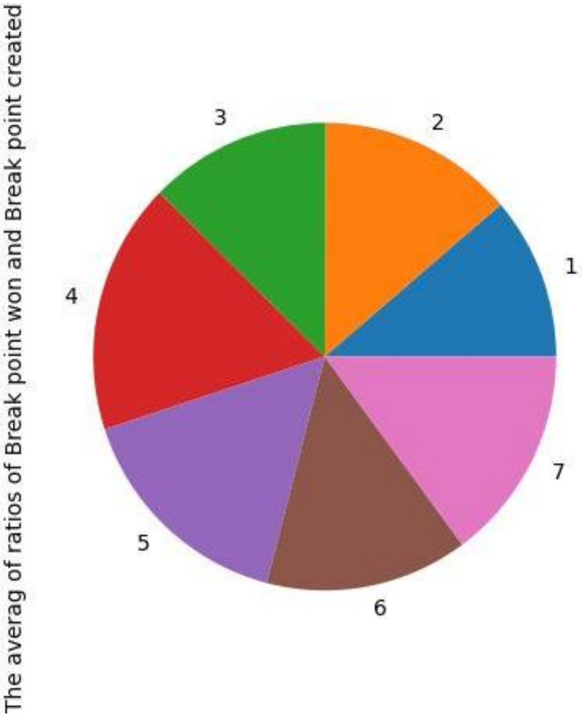


Fig. 9 Pie chart of average of ratios of break points won and break points created per round

In Fig. 9, we can see that in each round the average ratio remains almost the same, so we can say that there is not a particular trend for the ratio of break points won and break points created with respect to the round in which that match is being played. Here, it is also possible that if a player is doing too much mistakes, then he will miss the break point created. So, we will see the ratio of break points won and unforced error with respect to round.

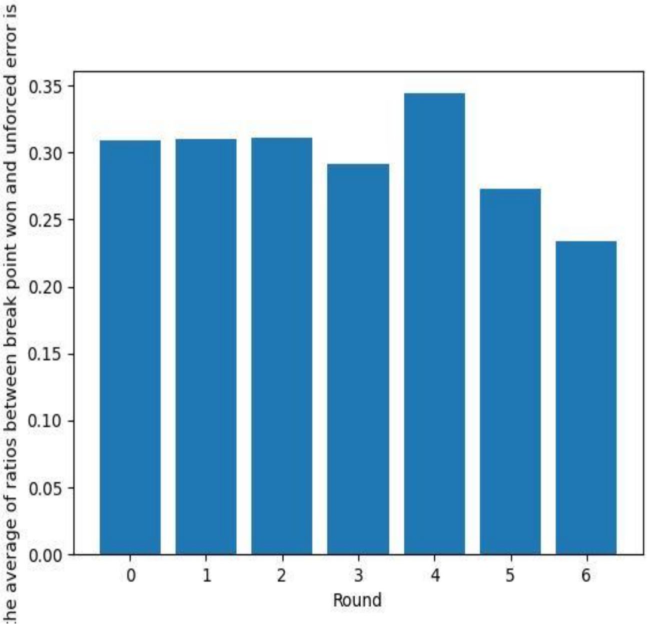
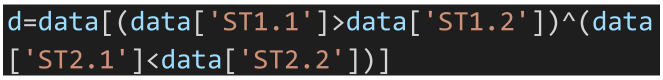


Fig. 10 graph of the average of ratios of break points won and unforced error vs round

In Fig. 10, we can see that all the bar plots have nearly the same height, so we can observe that round does not matter anyhow on the break point.

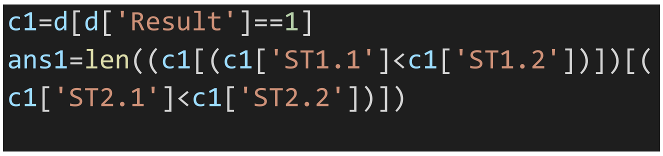
VI. What is the probability of winning a match after losing the first and second set in that match?

To solve this question, we need to split the main data into a smaller one in which first player either losses or wins first two sets in a particular match. So to find that, we have used below code, (We are using “USOpen-men-2013.csv” data for this analysis.)



```
d=data[(data['ST1.1']>data['ST1.2'])^(data['ST2.1']<data['ST2.2'])]
```

Here, it will see that, in which data, winner of 1st set and 2nd set are same. With the help of above code, we will get our desired data in the form of variable d. After that, we will see in which matches player 1 has won the match but lost in 1st and 2nd set and we will note down the length of that data to find how many players are there with this situation and this can be done by below code,



```
c1=d[d['Result']==1]
ans1=len((c1[(c1['ST1.1']<c1['ST1.2'])])[(c1['ST2.1']<c1['ST2.2'])])
```

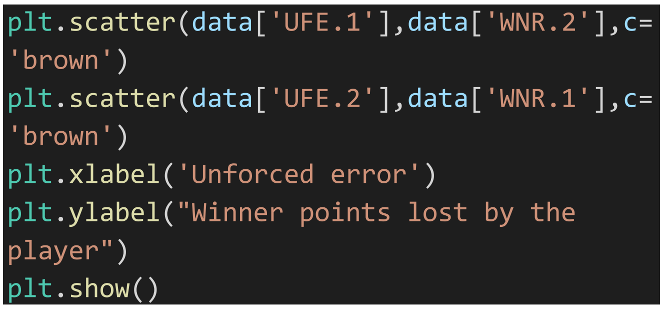
Similarly, we will find the same thing for player 2, also. And after getting number of total players who satisfies the condition, we will divide it with total number of players who lost 1st and 2nd set.

So, after doing this thing, probability comes out to be 0.0379746835443038, so we can say that after losing 1st and 2nd set, it is very hard to come back and win the match.

VII. Winner points of a player should be more if opposite player does more unforced errors.

Winner points are those points which are very common but very important in a match. Winner point is that point which simply don’t satisfy any criteria. So, it is obvious that if your opponent is doing very much mistake means his/her unforced error is very high then your winners point count will be so high.

So, to analyse this we need to plot a graph between winners point of a player and unforced error of opponent player. And we can do this with the help of following sequence of code, (We are using “Wimbledon-men-2013.csv” data for this analysis.)



```
plt.scatter(data['UFE.1'],data['WNR.2'],c='brown')
plt.scatter(data['UFE.2'],data['WNR.1'],c='brown')
plt.xlabel('Unforced error')
plt.ylabel("Winner points lost by the player")
plt.show()
```

And the result we are getting from this is,

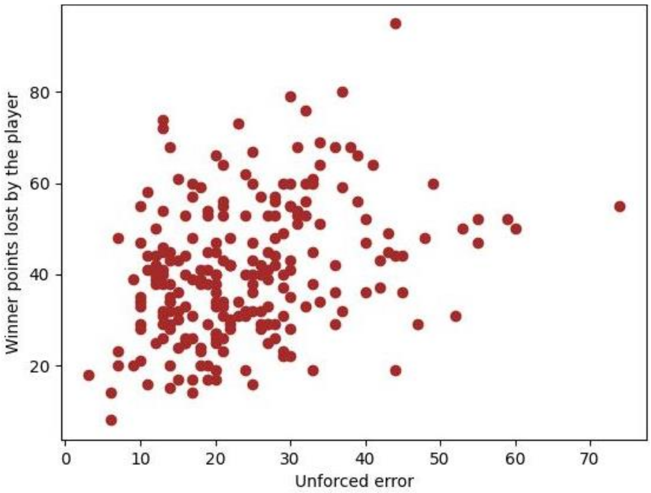


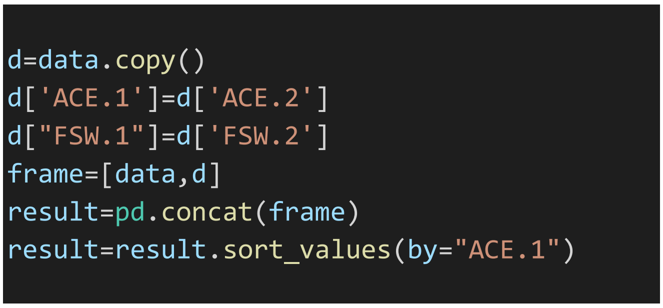
Fig. 11 Scatter plot between Unforced error of one player and winner point of opposite player

In Fig. 11, we can see that there is not a particular trend for winner points lost by a player and unforced error done by that player. So our hypotheses that there should be a trend between these two factors is false.

VIII. Is there any major effect of ace on the first serve win?

An ace point is earned when the serve is in but not hit at all by the receiving player. So, it is obvious that ace point should depend on the first serve of a player. If a player is doing a good first serve then it is possible that his/her chance of getting an ace increase. But it is generally shown that first serve win is directly related to the first serve percentage. So, to see the variation of ace points with first serve win we will plot a scatter graph between these two quantities.

To plot a graph, we need the data of both the players who are playing a match on the same column so that it becomes easy to handle. So, we made a copy of the data and added the columns of player 2 in the columns of player 1 using following sequence of code, (We are using “Wimbledon-women-2013.csv” data for this analysis.)



```
d=data.copy()
d['ACE.1']=d['ACE.2']
d["FSW.1"]=d['FSW.2']
frame=[data,d]
result=pd.concat(frame)
result=result.sort_values(by="ACE.1")
```

Now, we will plot the scatter graph between first serve win and ace points won by every player.

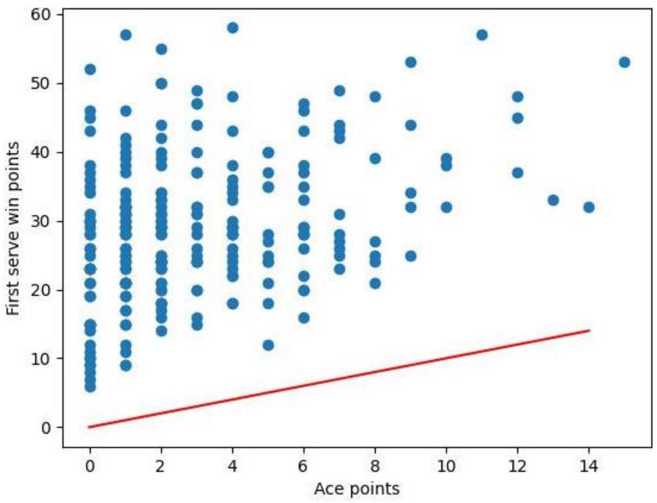


Fig. 12 scatter plot between ace points and first serve win points

In Fig. 12, the red line represents the minimum value that first serve win can take. In Fig. 12, we can see that minimum distance of points from the line increasing as the value of ace points increases, but this does not matter a lot because most of points are following a random trend in this.

So, we can say that there is no major effect of ace points in first serve win.

D. Details of Libraries and Functions

1. Pandas library is very useful while handling some kind of data especially when we are dealing with csv files.

2. pandas.read_csv() function reads the data of the csv file which we have given as input to the function and we can store the value of that with the help of assigning operator '='.

3. Matplotlib is a very useful library as it helps to handle the graphs which we are going to plot and it also helps to show the graphs on the screen.

4. NumPy is a library in python that helps us to do a wide range of operations on multidimensional arrays.

5. np.corrcoef() is a function provided by NumPy library in python that calculates correlation coefficient matrix for a given set of input arrays.

6. copy() function is very useful function while working with a dataframe or lists because if we don't use this function to copy a dataframe or a list and we directly assign the value then some times when we make changes to new variable old one also got changed.

E. Unanswerable Questions

1. "Winner points of a player should be more if opposite player does more unforced errors." - The reason why this hypothesis is not true is unanswerable because there are no specific reasons for this.

F. Summary of the Observations

1. From observations, we can say that if you are doing less mistakes than you can win the match in 3 sets only and your chances to win the match also increases.

2. First serve win plays a major role on the total points, although there is no relation between first serve win and first serve percentage. But first serve win totally depends on the skill of a player.

3. There is no effect of round on the performance of the player.

4. Player should attend smaller number of net points to win the match because when net points attended is small then the probability of getting a net point win is very high.

5. Net points contribute very less in total points earned by a player.

6. There is no effect of round on the ratio of break points created and break points won.

7. It is nearly insane for a player to win a match after loosing 1st and 2nd set of the match.

G. References

1. Oliphant, Travis E. A guide to NumPy. Vol. 1. USA: Trelgol Publishing, 2006.
2. Tosi, Sandro. Matplotlib for Python developers. Packt Publishing Ltd, 2009.
3. McKinney, Wes. Python for data analysis: Data wrangling with Pandas, Numpy, and IPython. "O'Reilly Media, Inc.", 2012.

H. Acknowledgements

I am very thankful to my professor Shanmuga and all teaching assistant for the course ES114 who shared this data to me and who thought that I am eligible to analyse this data.