Tennis Major Tournament Data Analysis

Parag Sarvoday Sahu
Electrical Engineering
(Roll No. - 22110179)
IIT Gandhinagar
Gandhinagar, Gujarat
parag.sahu@iitgn.ac.in

***Abstract*—To mine the given datasets containing information on the four major championships in tennis and report the observations.**

Keywords—python, pandas, data-analysis, tennis

I. Overview of the Dataset

The description of the datasets is as follows:

The given dataset has information on the following tennis tournaments held in the year 2013:

1. Australian open
2. French open
3. Wimbledon
4. US Open

The dataset has two CSV files for each tournament. One file contains information about the men's tournament and another contains information about the women's tournament. Each file has information on each match played in the tournament with numerical data on number of aces hit, double faults committed, net points scored, first serve percentage etc.

II. Procedure employed for analysis

First analysis was performed on the individual datasets thereafter all datasets were combined together to get more insight into the data.

III. Scientific Questions/Hypotheses

A. *Is there a feature correlated to the number of unforced errors committed by the player who wins the match? (Australian Open Men's section 2013)*

B. *Is there a relation between first serve and second serve for a player who wins the match? (All tournaments combined Men's section)*

C. *How does the performance of the winner of a tournament vary with the progress of the tournament? (Australian Open Women's section 2013)*

D. *Is there a trend in the number of unforced errors committed by players as the tournament progresses? (French Open Men's section 2013)*

E. *Assumption: If a person is winning a lot of first serves, they must also be hitting a lot of aces. (French Open women's section 2013)*

F. *Hypothesis: The player who wins more number of break points also wins the match. (US Open men's section 2013)*

G. *Do women and men serve equally well? (Wimbledon 2013)*

H. *Amongst the the three types of courts used to play tennis, which court is the most difficult to play on?*

IV. Details of Libraries and Functions

The various libraries used in this analysis are as follows:

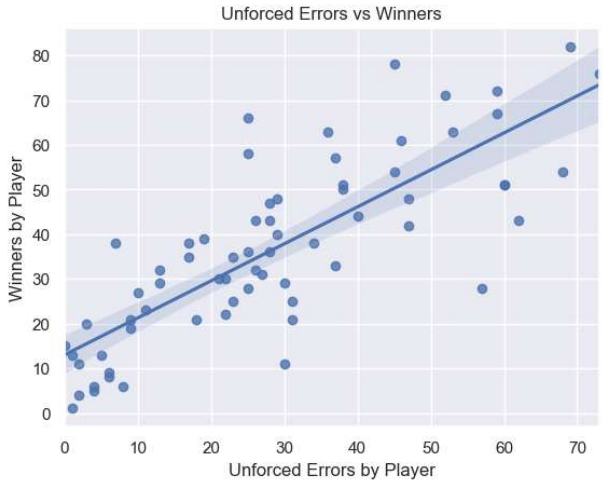
1. Pandas – This library is used to process and analyse large datasets.
2. Matplotlib – This library is used to get various types of plots using python language.
3. Numpy – This library is used to perform quick and efficient calculations using arrays, in python.
4. Seaborn – This library is used to get more detailed and modern looking plots than Matplotlib using python language.
5. Plotly – This library is used to make more advanced interactive graphs using python language.
6. Scikit-learn – This library is used to perform predictive data analysis. It was used to find correlation between the data of two columns.

V. Answers to the Questions

A. *A feature which is greatly correlated with the number of unforced errors committed by a player who wins the match (Australian Open Men's section 2013)*

Number of winners earned by a player shows a high correlation with the number unforced errors committed by him. A possible reason for this behaviour can be the frustration of committing an error on their own, which helped the player to play better to earn more winners and eventually win the match.

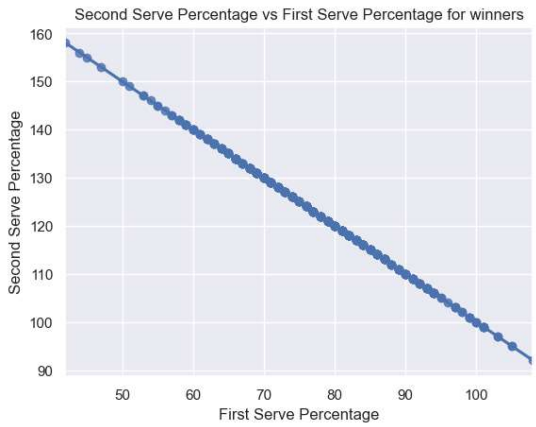
To obtain this result, first, a covariance matrix was formed to find out how various features are inter-related with each other. After observing a high correlation between the aforesaid two features, a regression plot was used to visualise their relationship.



Graph depicting the relationship between the number of unforced errors committed and number of winners earned by a winning player

B. *Relation between first serve and second serve for winning players*

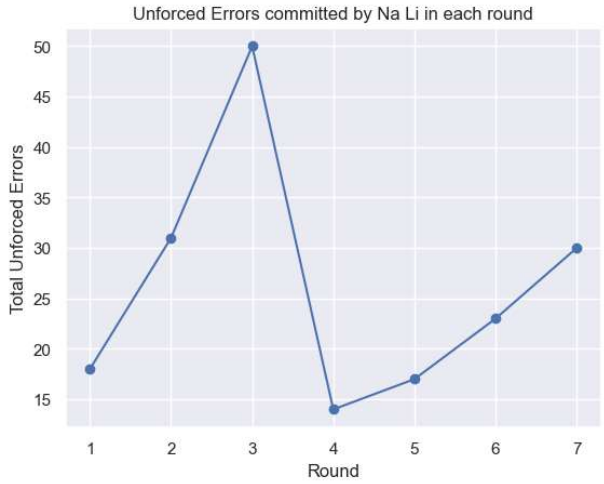
The first serve percentage for a winning player is highly negatively correlated with the second serve percentage. This might be an indication towards the higher reliance of winning players on their first serve. And the players become lousy in their second serve.



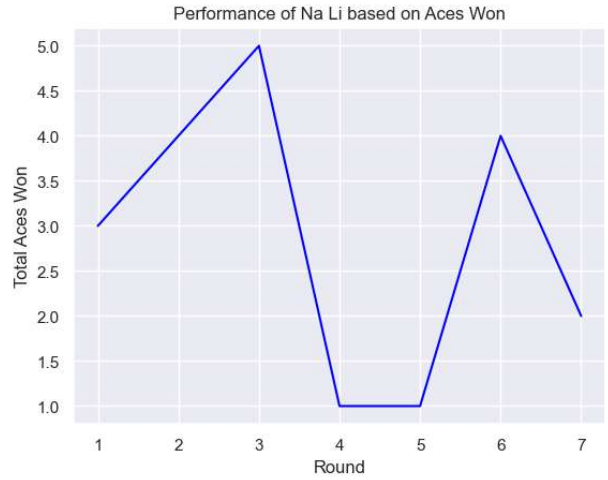
Graph depicting the negative correlation between first-serve percentage and second-serve percentage for winning players

C. *Performance of the winner of the tournament over the rounds*

The performance of the winner is gauged by the number of unforced errors committed and the number of aces won for a given round. It turns out that there does not exist an observable pattern in the variation of performance of the winner. The player's performance suddenly improved but abruptly drops in the next round which cannot be explained.



Line plot depicting the number of unforced errors committed in various rounds by the champion

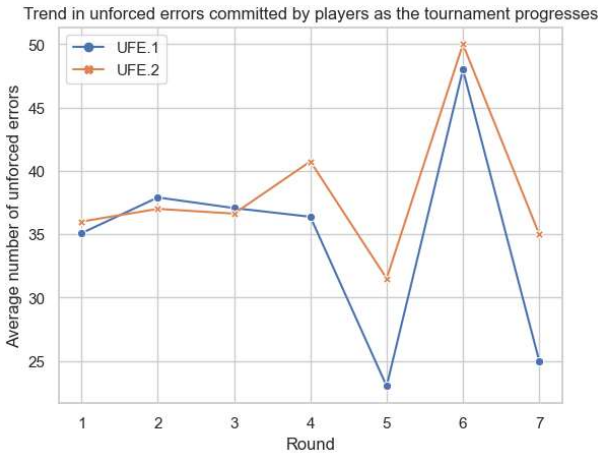


Line plot depicting the number of aces won in various rounds by the champion

D. Number of unforced errors committed by players in each round

There is no observable trend in the number of unforced errors committed by the players as the tournament progresses. It is interesting to note that, on average, in a given match, both players commit almost the same number of errors.

To answer this question, the average number of unforced errors committed by each player for all matches in a given round was calculated.



Line plot depicting the average number of unforced errors (UFE) committed by both players (in blue and orange) in each round

E. Correlation between number of aces and first serves won

The assumption is incorrect as the correlation coefficient between aces hit and the number of first serve won comes out to be just 0.26.

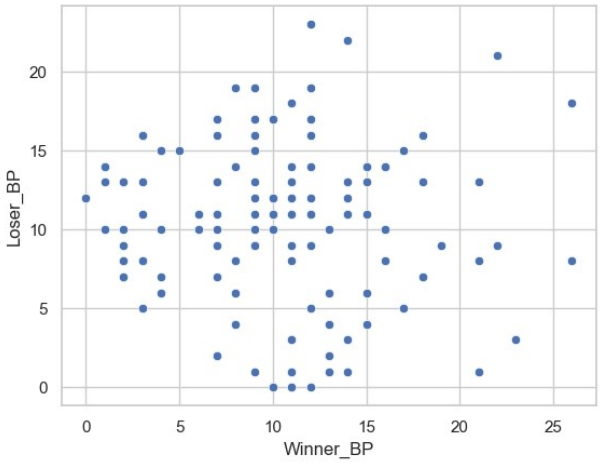
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

The formula used to calculate the correlation coefficient

F. Hypothesis: The player who wins more number of break points also wins the match.

The hypothesis does not hold true as the loser and the winner of a match seem to win equal number of break points on average. The correlation coefficient between the number of break points won by the winner and the loser also comes out to be -0.04.



Scatter plot depicting number of break points won by winners and losers

G. Do women and men serve equally well? (Wimbledon 2013)

It was found that women and men do not perform the same way on serving front. The average number of aces for men is almost double the average number of aces for women.

Average number of aces by men = 14.39
Average number of aces by women = 6.24

Assumption: The metric used to judge serving ability was number of aces by the player.

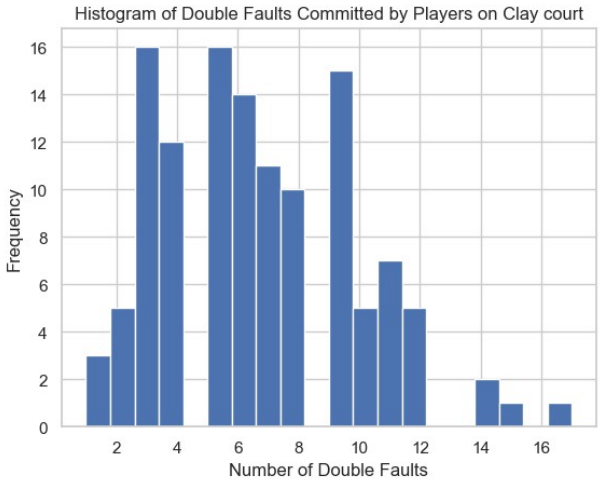
H. Difficulty level of various tennis court surfaces

There are 3 types of surfaces used to play tennis. We consider the men's section of the following tournaments:

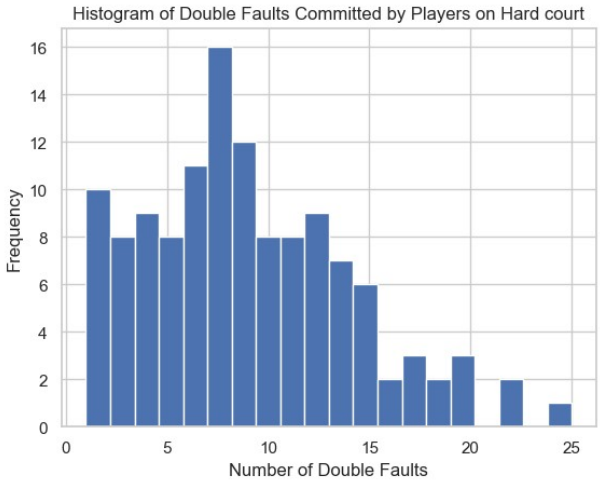
- French Open – Clay court
- Australian Open – Hard court
- Wimbledon – Grass court

We consider the number of double faults committed by players as metric for the difficult of play i.e., a greater number of double faults implies more difficulty and vice-versa.

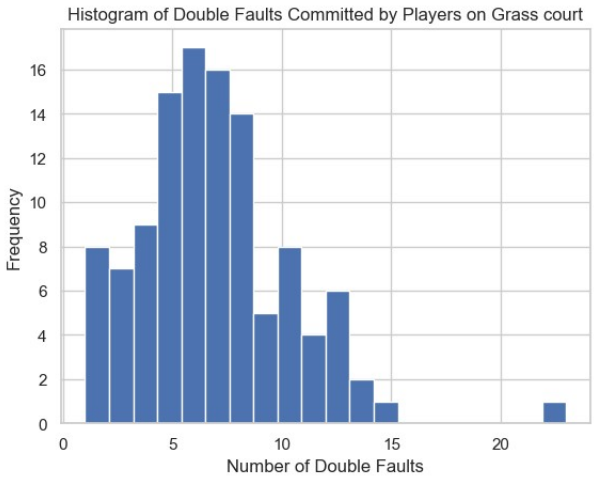
Looking at the histograms, it seems that clay court is the most difficult to play on, followed by hard court and thereafter grass court.



Histogram depicting the number of double faults on clay court



Histogram depicting the number of double faults on hard clay



Histogram depicting the number of double faults on hard clay

VI. Summary of the Observations

Tennis is a difficult sport to predict as finding correlation in this dataset was quite difficult. Nonetheless, there were some correlations present. It was found that the difficulty of play varies on different court types and one should not try to directly compare men's and women's performances as the two do not complete with each other.

Acknowledgment

I would like thank Prof. Shanmuganathan Raman for providing us with this opportunity of experiencing Data Analysis in our 1st year of BTech program. I also thank the creators of all the free resources available on the internet.

References

- Pandas user guide. April 20, 2023. https://pandas.pydata.org/docs/user_guide/index.html
- Matplotlib user guide. April 20, 2023

  https://matplotlib.org/stable/users/index.html
- Seaborn user guide. April 20, 2023. https://seaborn.pydata.org/tutorial.html
- Data Analysis with Python Course. April 21, 2023.

  https://jovian.com/learn/data-analysis-with-python-zero-to-pandas