Probability , Statistics and Data Visualisation

Tarun meena
Department of Electrical Engineering
Indian Institute of Technology Gandhinagar
Gujarat, India

***Abstract*— The "Tennis Major Tournament Match Statistics" dataset provides detailed information about tennis matches played in Grand Slam tournaments from 2000 to 2016, for both men's and women's singles matches in different countries. The dataset includes various match statistics, player information, and tournament information, making it useful for researchers and analysts interested in studying tennis match statistics and trends over time. This dataset can be used to develop predictive models or to gain insights into player performance and strategies.**

I. OVERVIEW OF THE DATASET

The dataset contains detailed information about tennis matches played in Grand Slam tournaments from 2000 to 2016. The dataset includes data for both men's and women's singles matches.

The data in this dataset includes various match statistics, such as the number of aces, double faults, first serve percentage, etc. It also includes information about the players, such as their names, countries, and rankings. Additionally, the dataset includes information about the tournament. Also there are 8 datasets , 4 for mens and 4 for womens of four different countries.

Overall, this dataset can be useful for researchers and analysts who are interested in studying tennis match statistics and trends in Grand Slam tournaments over time. It can also be used to develop predictive models or to gain insights into player performance and strategies.

II. SCIENTIFIC QUESTIONS AND HYPOTHESIS

Q1 From the given dataset AusOpen-men-2013 , if a player scores a higher percentage of First Serve Points than the opponent, what is the probability that he won that game? Also, give a pie chart of the players who scored more % of first serve points categorized by winner vs loser.

Q2 If a player can score an ace won, then there is a chance that there is less competition for that player. But with increasing the round level, the competition increases. Then, find the avg ace won in each round for the winning players and compare those using a bar graph. (from dataset AusOpen-women-2013).?

Q3 From the dataset, FrenchOpen-men-2013

Lists the players who are constantly performing better in each round till the final round. That means the no. of points the player earned increases in each round. Also, plot the line chart for no. of improving players for each round.

Q4 In the dataset FrenchOpen-women-2013

If a player attempts to net points, it shows that the player is trying to play aggressively. On the other, if a player attempts the least net points, this indicates that she is playing defensively. From the given dataset for each match, the winner finds whether she has an aggressive or defensive style. Also, plot the pie chart for the winner w.r.t aggressive and defensive styles.

Q5 In the dataset USOpen-men-2013

, what is the probability that a player wins the match if he does not commit double faults? Also, find the probability that if a player commits less double faults than his opponent and he wins the match.

Q6 In dataset USOpen-women-2013

If we choose a player randomly for any match, then what is the probability that the Player won the match when their first serve winning percentage is greater than 50%?

Q7 In the dataset, Wimbledon-men-2013

find the top 4 players, i.e. players who qualify for round 6. Compare Avg first serve won, Avg net points won, Avg ace points won, and Avg break points won using a bar graph.

Q8 An unforced error committed by any player may affect the result of the match. At a higher level of competition, there is less chance for any mistake. Find the avg unforced error done by players in each round. Also, plot the graph for round vs avg unforced error for tournament Wimbledon-women-2013

III. DETAILS OF LIBRARIES AND FUNCTIONS USED

In the field of data analysis and machine learning, Python is a widely-used programming language that

provides a wide range of libraries for data manipulation ,analysis, and visualization. There are various libraries in python to make the data appeal more informative and interesting.

I. Numpy - a library for numerical computing that provides fast and efficient array operations, including linear algebra, Fourier transform , and random number generation.

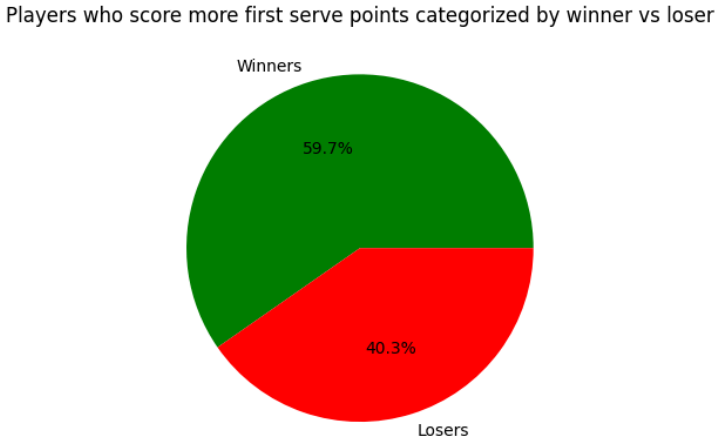
II. Matplotlib - a plotting library that provides a wide range 2D and 3D visualization tools, including line plot, scatter plot ,histograms, pie-charts , and heat maps.

III. Pandas - a library for data manipulation and data analysis that provides tools for cleaning, merging and reshaping, as well as data visualization and statistical analysis.

IV. ANSWERS TO QUESTIONS (WITH APPROPRIATE ILLUSTRATIONS)

Ans1.

The probability that the player who scores more no. of first serve point won is: 0.59



Here we can see that if a player scores more no. of first serve points then there is a probability of 0.59 that he can win the match.

Ans2.

Avg ace points scored by winners in 1 round is 3.8125
Avg ace points scored by winners in 2 round is 3.15625
Avg ace points scored by winners in 3 round is 3.0625
Avg ace points scored by winners in 4 round is 2.75
Avg ace points scored by winners in 5 round is 1.0
Avg ace points scored by winners in 6 round is 2.0
Avg ace points scored by winners in 7 round is 2.0

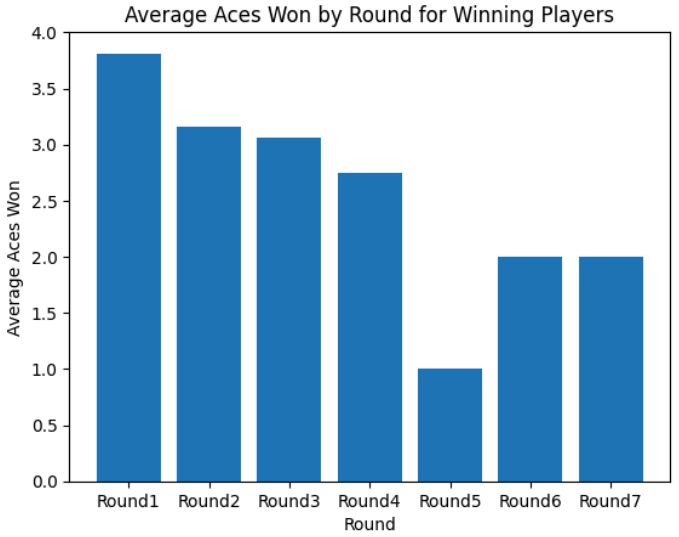
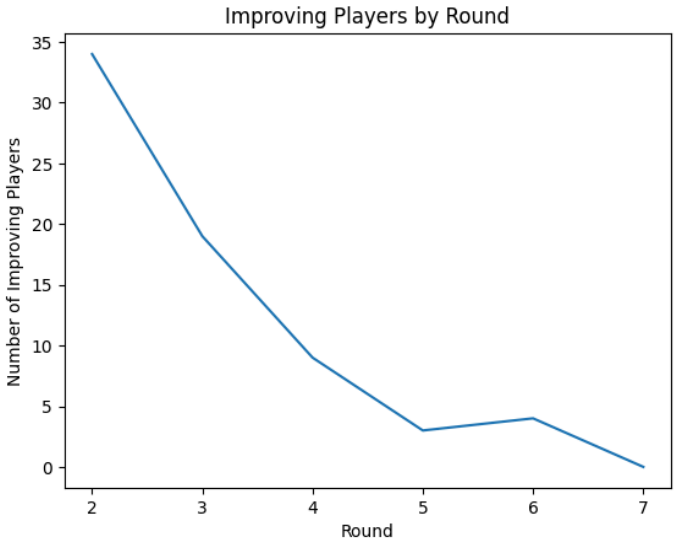


fig2: Bar graph

Here we can see that.Avg ace points are decreasing with increasing round level.(Exception Round 5 to Round 6)

Ans3.

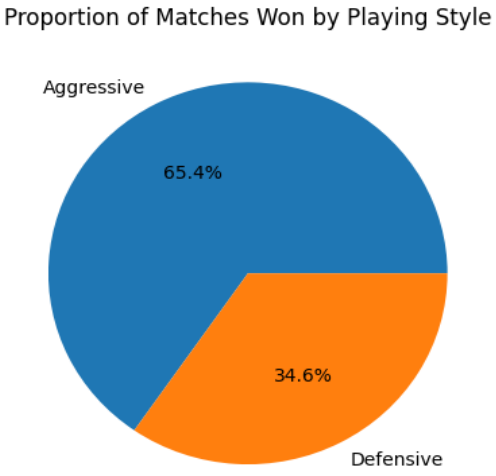
Round		Players
0	2	[Julien Benneteau, Roberto Bautista Agut, Jo-W...
1	3	[Gilles Simon, Gael Monfils, Nikolay Davydenko...
2	4	[Gilles Simon, Jo-Wilfried Tsonga, Tommy Robre...
3	5	[Jo-Wilfried Tsonga, Rafael Nadal, Tommy Haas]
4	6	[David Ferrer, Novak Djokovic, Jo-Wilfried Tso...
5	7	[ ]



Here we can see that players are improving themselves but as soon as the rounds are increasing they can't score more points than the previous round. This may be because of the increasing competition.

Ans4.

	NPA_Winner	NPA_Loser	Winner_Style
0	3	1	Aggressive
1	16	4	Aggressive
2	4	14	Defensive
3	5	2	Aggressive
4	1	3	Defensive
.	.	.	.
.	.	.	.
.	.	.	.



Here we can see that players who play with an aggressive style have more chances to win.

Ans5.

From the information of USOpen-men-2013 tournament we can say that

The probability of winning the match with no double faults is 0.75

The probability of winning the match with fewer double faults is 0.66

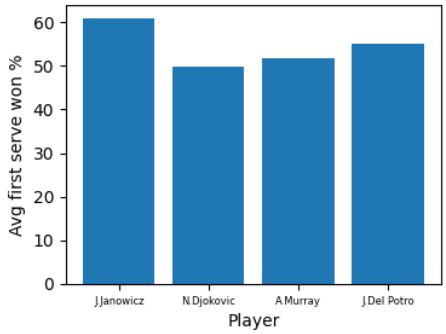
Ans6.

From the data given for USOpen-women-2013 tournament we can say that
The probability of winning the match when first serve won percentage greater than 40% is 0.94

Ans7.

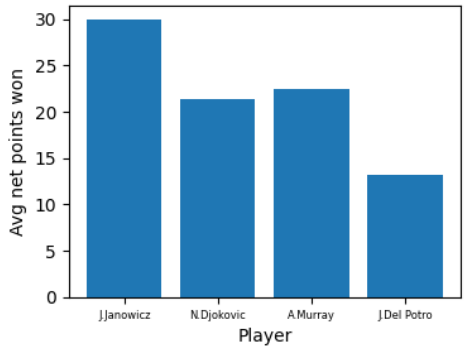
Avg first serve % of top 4 players are

J.Janowicz 61.0
N.Djokovic 49.714285714285715
A.Murray 51.714285714285715
J.Del Potro 55.0



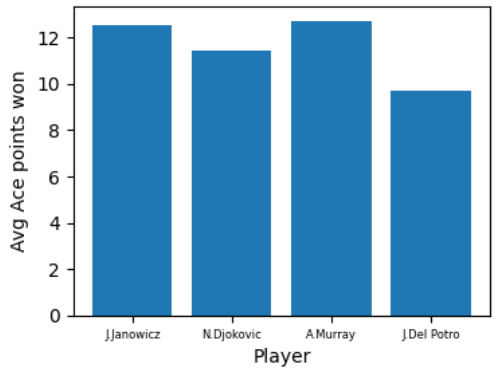
Avg net points won by top 4 players

J.Janowicz 30.0
N.Djokovic 21.42
A.Murray 22.42
J.Del Potro 13.16



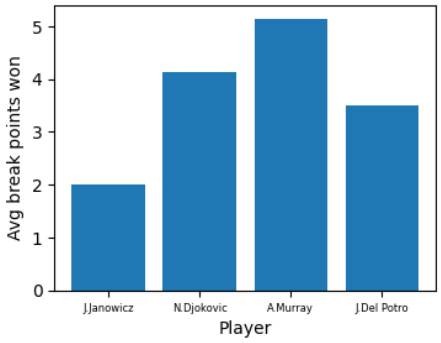
Avg ace points won by top 4 players

J.Janowicz 12.5
N.Djokovic 11.42
A.Murray 12.71
J.Del Potro 9.66



Avg break points won by top 4 players

J.Janowicz 2.0
N.Djokovic 4.142857142857143
A.Murray 5.142857142857143
J.Del Potro 3.5



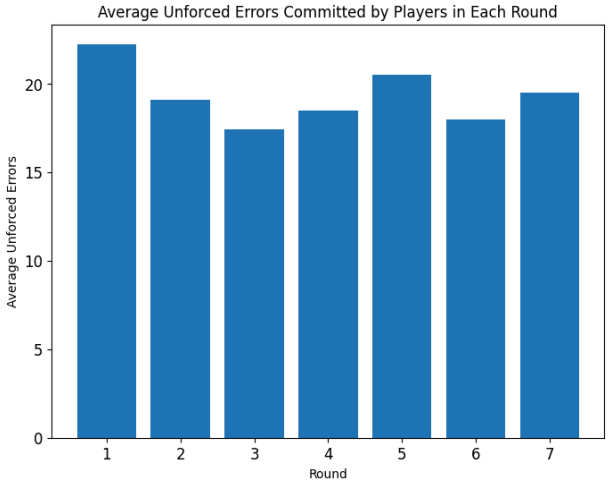
This data for top4 players may be helpful to make strategies against them.

Ans8.

Round

1 22.234375
2 19.092593
3 17.437500
4 18.500000
5 20.500000
6 18.000000
7 19.500000

This is the data of avg unforced error in each round



IV. SUMMARY OF THE DATA

1. The probability that the player who scores more no. of first serve point won is: 0.59
2. The avg no. of ace points scored in a round are decreasing with increasing the round level
3. Players are improving themselves but as soon as the rounds are increasing they can't score more points than the previous round. This may be because of the increasing competition
4. Players who play with an aggressive style have more chances to win.
5. The probability of winning the match with no double faults is 0.75
   The probability of winning the match with fewer double faults is 0.66
6. for USOpen-women-2013 tournament we can say that
   The probability of winning the match when first serve won percentage greater than 40% is 0.94
7. There is comparison for avg FSW, avg BPW, avg Ace points and avg net points between top4 players
8. In general avg UFE in each round will start decreasing with increase in round level.

V. REFERENCES

- https://numpy.org/
- https://matplotlib.org/
- https://pandas.pydata.org
- https://seaborn.pydata.org/