DATA NARRATIVE 3

ES 114 PROBABILITY,STATISTICS AND DATA VISUALIZATION

Purva Kaushalbhai Shah
Chemical Engineering
Indian Institute of Technology Gandhinagar
Gandhinagar, India
purva.shah@iitgn.ac.in

***Abstract*—This document is a report on Data Narrative. The data used here Tennis Major Tournament Match Statistics [5]. This data contains 8 different files. It shows us with the data of 4 Major Tournaments: The French Open, The Australian Open, Us Open and Wimbledon Open. Each tournament's data is recorded in two files, one for women and one for men.**

I. Introduction

The data provided in the dataset consists of 8 files. It has data of 4 tournaments and each tournament has 2 files, one for men and one for women. This dataset shows us the whole statistics of the matches played. It contains 42 columns and 76 rows.The data is of 2013. These 8 files had data of 4 Major Tournaments the Australian open, The French open, US open and Wimbledon Open and these four tournaments have 2 files one for men and other for women. It contains 42 columns and 76 rows.The data is of 2013.

II. Overview

In this task, we are reading the data provided using pandas[2]. Tennis Major tournament Match Statistics [5] contains 8 files. These 8 files had data of 4 Major Tournaments the Australian open, The French open, US open and Wimbledon Open and these four tournaments have 2 files one for men and other for women. It contains 42 columns and 76 rows..

III. Questions/ Hypotheses

A. *How does the mean distribution of points of players differ in all sets? Approximately what is the average number of points scored by player1 in all rounds and scored by player 2 in all rounds in the Australian Open Men (2023)? If a set is chosen at random, what is the probability that player 2 scored more (mean as considering the whole set) than player 1 in the Australian Open Men (2023)?*

B. *Who won more matches in each round in the Australian open women(2013)?If a match is chosen at random, what is the probability that player1 has won it? Show the distribution of matches won by players.*

C. *Show the graph of final numbers of matches won by players in each round in the French Open Men (2013)? What is the maximum number of matches a player can win in each round? What is the highest number of matches that a player won?*

D. *Show the distribution of first serve percentage players get in the French Open Women(2013). What range of percentage is most common? Show the first serve win percentage. What is the maximum first serve win percentage?*

E. *Show the distribution of total points scored by players in the US Open Men(2013). What is the average of points scored by player1 and player2?*

F. *Which player has more second serve percentage? Who has more chances to win the second serve?*

G. *How are break points distributed among players? Who won more break points? When a break point created created by player 1 is chosen at random, what is the probability that it is below 10 given that it is above 5?*

H. *Show the average distribution of Aces won by players in the Wimbildon Open Women (2013). Who scored more aces in round 2? If a round is chosen at random, what is the probability that in that round player2 has scored more aces?*

I. *Is the performance of a player changing in different Tournaments?*

J. *Who is the best tennis player?*

IV. Software/Libraries Used

A. *Google colab [1]*

A platform provided by google where we can code. In this task google colab was used to read the uploaded file, code and write some texts in between. The code in google colab is written in python language.

Fig 1 - Google colab [1]

B. *Pandas [2]*

Pandas is a software which provides us with various inbuilt functions. In this task, we are using pandas to read the csv file and operate on it. The csv files are viewed and are operated like we operate the Data frames.

1. *Functions used:*
    a. *apd.read_csv(): to read the csv file.*
    b. *dtypes: to check the type of columns present in the dataframe*
    c. *groupby : to group the data by a particular column*
    d. *pd.count(): to count the number of values*

Fig 2 - Pandas[2]

C. *Matplotlib [4]*

Matplotlib is the software used to plot the graphs. In the task of data narrative, it was used to plot the various types of graphs like bar graphs, pie charts etc.

1. *Functions used:*
    a. *figure(): to open a panel where graph can be plotted*
    b. *plot(): to plot the given graph*
    c. *xlabel(): to label x axis*
    d. *ylabel(): to label y axis*
    e. *bar(): to plot the bar graph*
    f. *pie(): to plot the pie chart*
    g. *show(): to show the plotted graph*
    h. *plot.scatter() : to plot the scatter graph*
    i. *grid(): to show the grid in the graph*
    j. *explode : to make one of the wedge present in pie chart stand out of others*
    k. *unstack(): it returns a DataFrame having a new level of column labels*

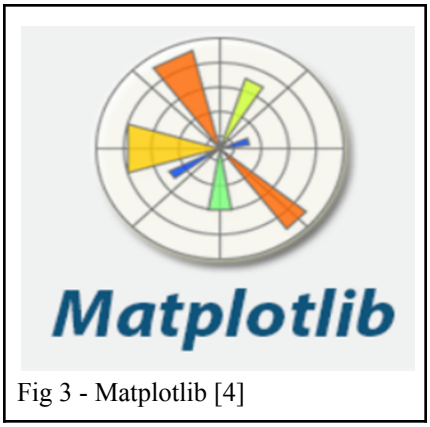

Fig 3 - Matplotlib [4]

D. *Seaborn [6]*

1. *Functions used:*
    a. *boxplot(): to show the boxplot*
    b. *kdeplot() : to plot the kernel density*
    c. *histplot(): to plot the histogram*

Fig - 4 Seaborn [6]

V. Answers

A. *The graph below shows the mean of points scored by player1 and player2 in different sets. Each color in the bar graph represents a different set played by the player. Average points scored by player1 in all sets is 4.936554 whereas average point scored by player2 in all sets is 4.810325000000001. By the*

graph we can see that player 2 has scored more than player 1 in the 5th set only.
Probability of player2 scoring more than player1 is 0.2

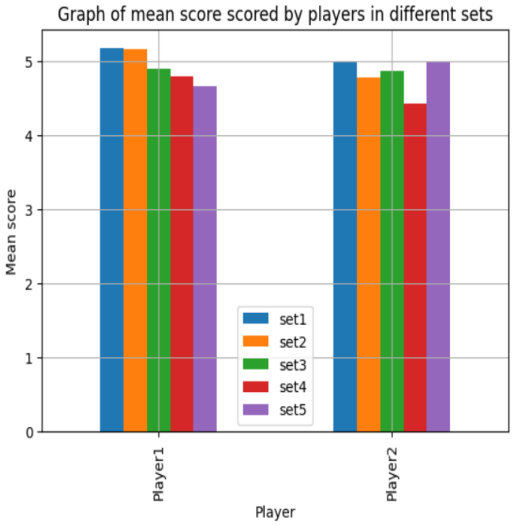

Fig 5 - Graph showing mean score scored by players in different sets.

From the above graph we can also say that player1 scored most points at average set2.

B. *Player2 won more matches than player 1 in rounds 1 and 4. Player1 won more matches than player 2 in rounds 2,3, and 7. Both players won the same number of matches in rounds 5 and 6.*
If the match is chosen at random, the probability that player1 has won the match is 0.48031496062992124 The graph shows the distribution of matches won by player 1 and 2 in each round. The graph below shows the win distribution of players in total.

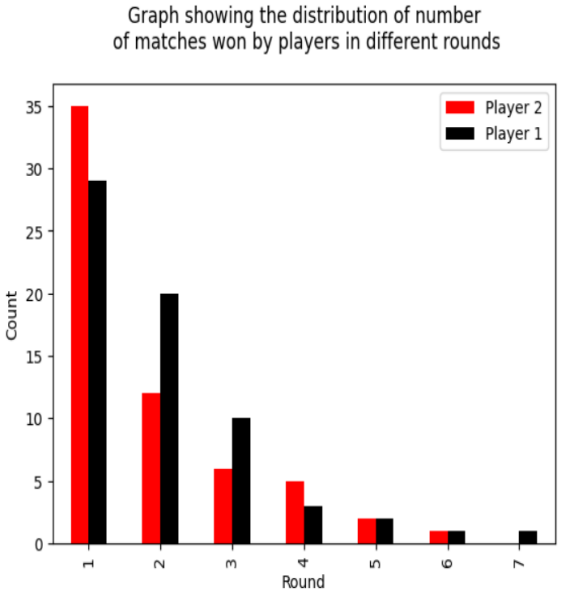

Fig 6 - Graph showing the distribution of number of matches won by players in different rounds

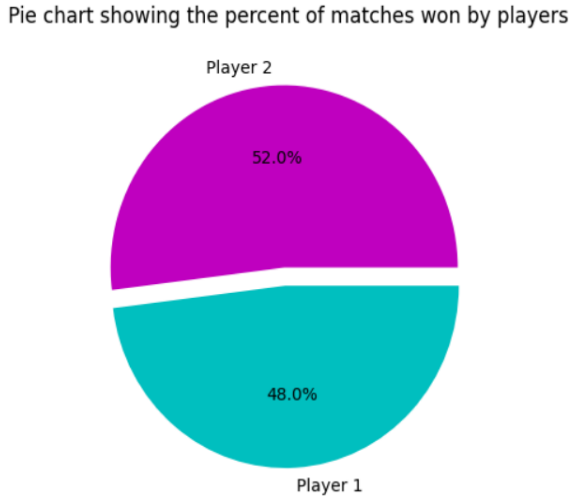

Fig 7 - Pie chart showing the percent of matches won by players

C. *The graph below shows the final number of matches won by players in each round. Here Player1 and Players2's match won is taken in count irrespective of who the player is. If the player plays as player1 his data would be marked in player1. If he plays as player2 then data would be marked in player2 irrespective of who the player is.*
The maximum number of matches played in each round is 3. So, the maximum number of matches a player can win in a round is 3. The maximum number of matches won by a player in a round is 3.

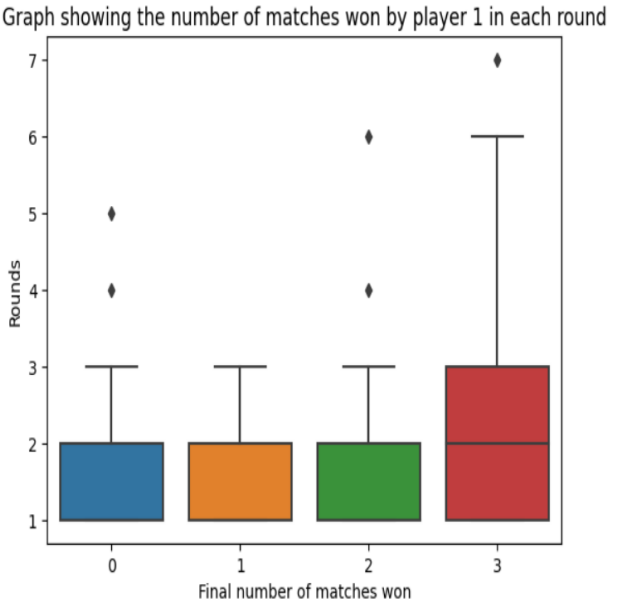

Fig 8 - Graph showing the number of matches won by player 1 in each round

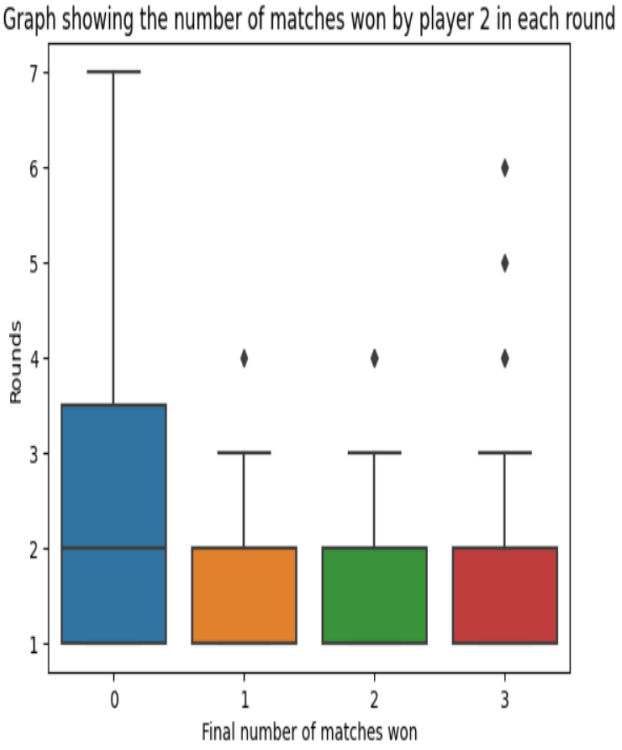

Fig 9 - Graph showing the number of matches won by player 2 in each round

From the graph above we can say that player2 in round 2,3 has won 0 matches in majority and player 1 has won 3 matches in majority in round 2. The points present in the graph says that in this round these numbers of matches are won by the player.

D. *The graph below shows the first serve percentages of player1 and player2.*

The graph below shows the first serve win percentages of player1 and player2. We can see that player 2 marked in magenta color has more first serve percentage. If we choose the first serve percentage of a player2 then there are high chances that it lies between 50%-80%. If we choose player 1 at random then there are high chances that the first serve percentage lies between 60-70%.

If we see the graph of win percentage of the first serve of player 1, then we can see that the highest first serve win percentage is 70%. Player 1 has a 30% win percentage as the majority.

If we choose the graph of win percentage of the first serve of player 2, we can see that the highest first serve win percentage is nearly 70%. Player 1 has a 30-40% win percentage as the majority.

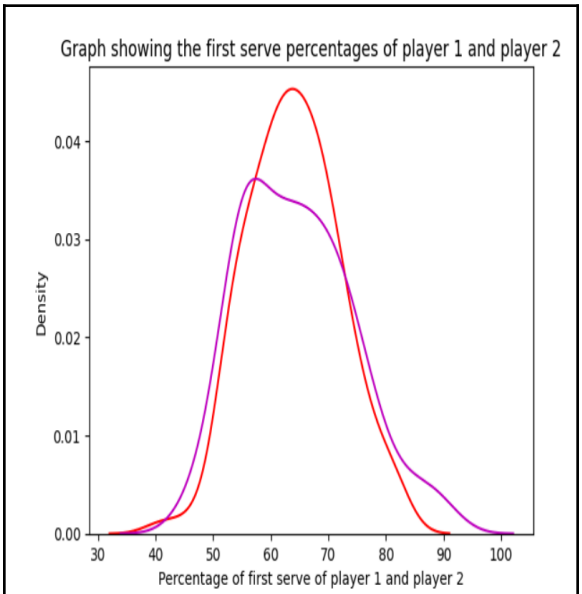

Fig 10 - Graph showing the first serve percentages of player1 and player2

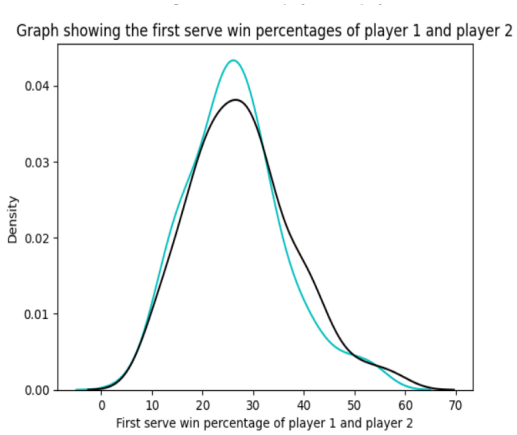

Fig 11 - Graph showing the first serve win percentages of player1 and player2

The most common range of percentages appearing when a player serves is 50-70%. The maximum first serve percentage for player 1 is 82% and for player 2 is 91%.
The maximum first serve win percentage for player1 is 54% while for player 2 it is 58%. So the maximum first serve win percentage is 58%. We can see that the first serve win percentage of player 1 is higher than player 2. It means that there are very high chances that player 1 will win the first serve.

E. *From the graph provided below we can say that player 1 has majority points scored in range 80-120 whereas player 2 has majority points scored in range 100-120.*

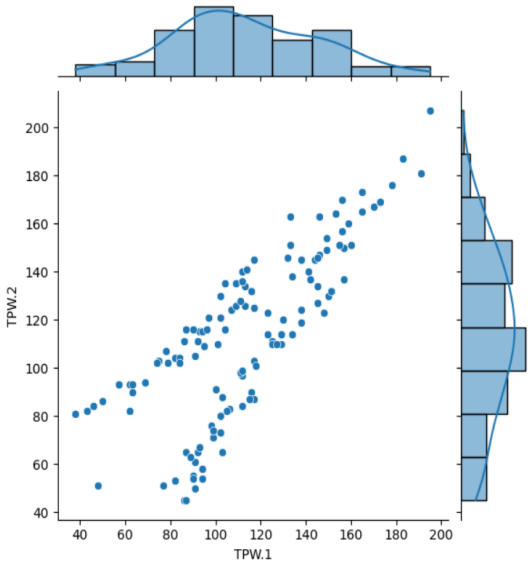

Fig 12 - Graph showing the distributions of total number of points scored by player 1 and player2.

The average total points scored by player 1 is 112.93 and average total points scored by player 2 is 113.18. If we select a player from player 1 and see its total score there are high chances he has scored around 100 points. If we select a player at random then there are very high chances that he scored between 80-120.

F. *Player 2 has more seconds serve percentages. This means that player 2 got more chances to serve second. Player 1 has more second serve win percentage, the majority of win percentage lies around 10% whereas for player 2, it lies around 15%. In the range 40-50% if choose the player there is a higher possibility that the person chosen is player 1.*

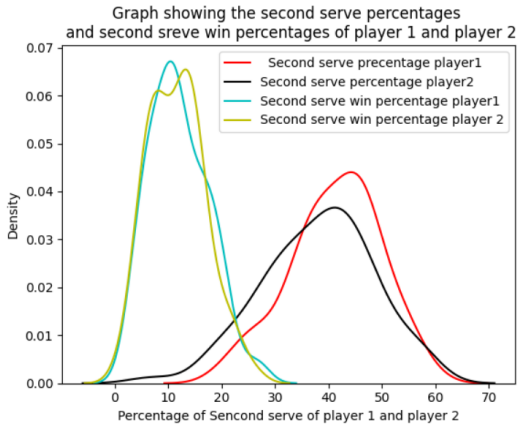

Fig - 13 Graph showing the second serve percentages \n and second serve win percentages of player 1 and player 2

G. *The graph (14) below shows the distribution of break points created by player 1 and player 2. From the graph we can say that the majority of the breakpoints generated by both the players lies in range 5-15 points. Player 2 has created more break points in the range 5-10 points. Player1 has created a total of 988 break points. Player 2 has created a total of 902 break points.*

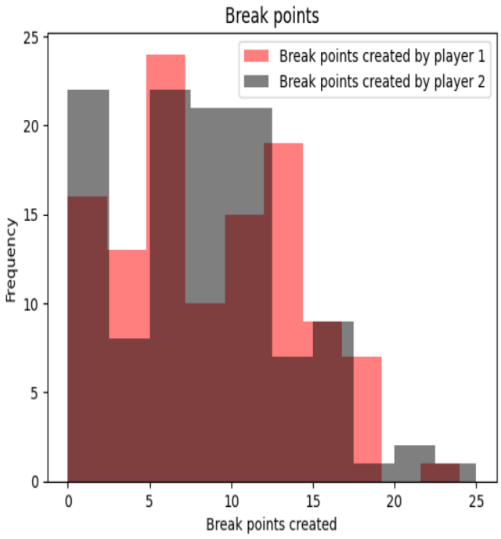

Fig - 14 Break points
(Graph showing the distribution of break points)

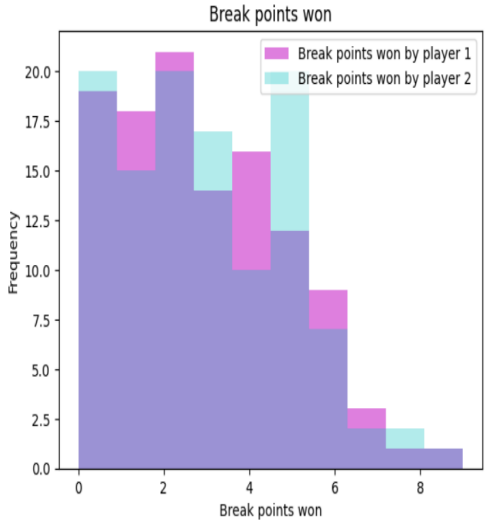

Fig - 15 Break points won
(Graph showing the distribution of breakpoints won)

The graph (15) shows the distribution of break points won by the player. We can see that winning points 3 and 5 have occurred with the highest frequency by player 1 and 2 respectively. Only Player 2 has won more than 8 break points. Player 1 has won 318 points in total and player 2 has won 327 points in total. Though player 1 created more break points but player 2 scored more points.
The probability of that when a break point created by player1 is chosen at random, it lies below 10 and above 5 is 0.23920265780730895

H. *The graph below shows the average distribution of the aces scored by players in each round.*
From the graph, we can see that player 1 has scored the maximum aces in round 6 whereas player 2 has scored maximum aces in round 5. The overall distribution of aces scored by player 2 is confined in a particular range with less variation. In almost every round player 2 has won more aces on average. Player 2 scored more aces in round 2. The probability that when a round is chosen at random, player 2 has scored more aces than player 1 is 0.7142857142857143 . In round 7, the aces won by player 2 on an average is twice the aces won by player 1. Both the players have their avg highest aces won same 15.5

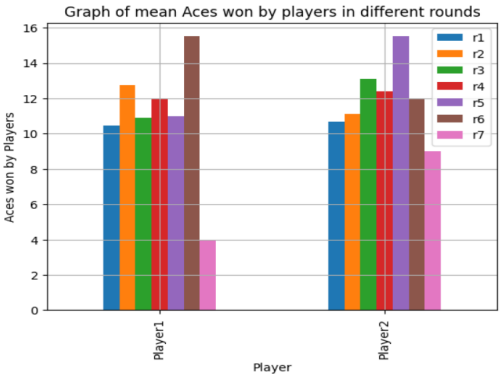

Fig - 16 Graph showing mean of aces won by players in different rounds

VI. Observations

A. *The average points scored by player 1 is more than player2 in all sets. Player 1 scored more points in all rounds except set 5.*
The probability of of player2 scoring more average points in a set than player1 is 0.2

B. *Player2 has won more matches than player1 in rounds 2,3, and 7. player2 won more matches than player 1 in round 1 and 4. Round 5 and 6 had a tie in the number of matches won by player1 and player2.*
If a match is chosen at random the probability that player1 wins the match is 48%.

C. *Player 1 has won 3 matches in majority in round 2 and 3. Player 2 has won 0 matches in majority in rounds 2 and 3.*

D. *From the graph we can see that the majority of first serve percentages of person 1 are between 60-70% and majority of first serve percentages of player 2 is between 50-60%. So we can say that the majority of players have the first serve percentages between 50-70%. We can see that the first serve win percentage of player 1 is higher than player 2. It means that there are very high chances that player 1 will win the first serve. The first serve win percentage of both players is between 20-40%.*

E. *From the graph we can see that player1 has the majority of total points scored around 100 points whereas player2 has scored the majority of total points in range 100-120. Both the players have an average of total points scored above 100 points. If we select a player at random, then there are high chances that his total points scored will be around 100.*

F. *Player 2 has a maximum second serve percentage. The second serve win percentage of player 2 is higher.Player 1 has more second serve win percentage, the majority of win percentage lies around 10% whereas for player 2, it lies around 15%.*

G. *Both the players have the break points with highest frequency in range 5-15 points. Player 1 has created more break points in total and player 2 has won more break points in total. When a break point created by player 1 is chosen at random, then its probability to be greater than 5 but less than 10 is 0.23920265780730895*

H. *Both player 1 and player 2 have highest aces won as 15.5, Player 1 has won the maximum ace point in round 6 whereas player 2 scored it in round 5. In rest all rounds their average aces won are almost the same but in round 7 there is a lot of variation. The aces won by player 2 in round 7 is almost two times aces won by player 1.*

I. *We can't say anything about the performance of a player because it changes in all tournaments, all matches. Also all players are not playing all tournaments.*

J. *We can't say who is the best tennis player because winners of all tournaments are different. Also there is no data to link the men and women. So, because of variations given in data, we can't say who is the best tennis player.*

VII. Unanswered Questions

Is the performance of a player changing in different tournaments can not be answered because not all players played all tournaments. There are few players who did not play all tournaments. Also the data given does not say which tournament has occurred at what time, So we can not say the performance has changed over time.

Who is the best tennis player is an unanswered question because winners of tournaments for women are all different players and for men also winners of all tournaments are 3 different players. We can't say who is the best player by this data. So, this question remains unanswered.

VIII. Acknowledgement

I would like to thank prof. Shanmuga for providing me with the data to do the analysis. The links provided were very helpful.

IX. References

[1] Google colab. Google. Accessed March 30, 2023. https://colab.research.google.com/.

[2] Pandas. "Python Data Analysis Library — Pandas: Python Data Analysis Library." Pydata.org, 2018, pandas.pydata.org/.

[3] Python. "Welcome to Python.org." Python.org, Python.org, 29 May 2019, www.python.org/.

[4] Matplotlib. "Matplotlib: Python Plotting — Matplotlib 3.1.1 Documentation." *Matplotlib.org*, 2012, matplotlib.org/.

[5] Jauhari,Shruti, Morankar,Aniket & Fokoue,Ernest. (2014). Tennis Major Tournament Match Statistics. UCI Machine Learning Repository. https://doi.org/10.24432/C54C7K.

[6] seaborn. "Seaborn: Statistical Data Visualization — Seaborn 0.9.0 Documentation." *Pydata.org*, 2012, seaborn.pydata.org/.