Data Narrative 3

Name: Mrugank Patil
Department: Computer Science and Engineering
Roll No: 22110158

I. OVERVIEW OF THE DATASET

The dataset named "Tennis Major Tournament Match Statistics" comprises information on matches played in the four major tennis tournaments. It contains various data points such as the names of players, match results, and diverse match statistics including first serve percentage, aces, winners, and more. The dataset includes 50 columns in total.

II. SCIENTIFIC QUESTIONS

A. Question 1 / (Hypothesis)

The probability of a winner having a first serve percentage of over 60% and a second serve percentage of over 30% in a randomly selected match from the dataset should lie in the range of 0.2 to 0.6.

B. Question 2

What is the probability that a player wins a match when they have hit more aces (ACE) than their opponent?

C. Question 3

If we assume that the distribution of total points won by a player (TPW) is approximately normal, what is the probability that a player will win the match if they have a TPW of 100?

D. Question 4

What is the probability that a women's match in the 2013 French Open will end with the most common final game score, and how many matches actually ended with that score?

E. Question 5

Is there a pattern in the distribution of break points created and won by players that could indicate clusters? (use KMeans to find the clusters)

F. Question 6

What is the joint probability that a player wins the match given that she wins at least 70% of her first serves and hits at least 10 winners?

G. Question 7

What is the probability of players who first name starts with A or M and have committed at least 3 double faults in a single match in Wimbledon Men's Tournament 2013?

H. Question 8 / (Hypothesis)

Even if the player have less net points than her opponent she have a small chance of winning the match.

III. DETAILS ON LIBRARIES AND FUNCTION

A. read_csv: This function is used to read a CSV file and load it into a pandas Data Frame.

B. Pandas: A Python package for data analysis and manipulation is called Pandas[1].

C. matplotlib.pyplot: Creating static, animated, and interactive visualisations in Python is made possible by the Matplotlib.pyplot module[2].

D. sum(): Gives back the total of all values in the given array or along a particular axis in a multi-dimensional array.

E. groupby(): Enables the application of aggregate functions to data grouped by one or more columns in a Pandas Data Frame.

F. nunique(): Pandas has a function called nunique() that gives the total number of distinct values in a column or series of a Data Frame

G. merge(): The Pandas merge() function combines two or more Data Frames based on a common key and supports several join types, including inner, outer, left, and right joins.

H. tight_layout(): The subplot parameters are automatically adjusted by tight_layout so that they fit in the figure area.

I. mean(): The mean/average of input values or a data set is determined using the statistics. mean() method.

J. std(): Used to calculate standard deviation of the input data.

K. scipy.stats: Numerous probability distributions, frequency and summary statistics, correlation functions and statistical tests, masked statistics, kernel density estimation, and other features are included in this module[3].

L. *norm.cdf(): It is used to calculate the normal cdf of the data priovided.*

M. *matplotlib.pyplot.vlines(): This function deals with the plotting of the vertical lines across the axes.*

N. *np.linspace(): It is used to create an evenly spaced sequence in a specified interval.*

O. *idxmax(): The idxmax() method returns a Series with the index of the maximum value for each column.*

P. *Scikit-learn: Scikit-learn is an open source data analysis library, and the gold standard for Machine Learning (ML) in the Python ecosystem[4].*

Q. *KMeans(): K-means is an unsupervised learning method for clustering data points. This function is used to calculate KMean of the input dataset provided.*

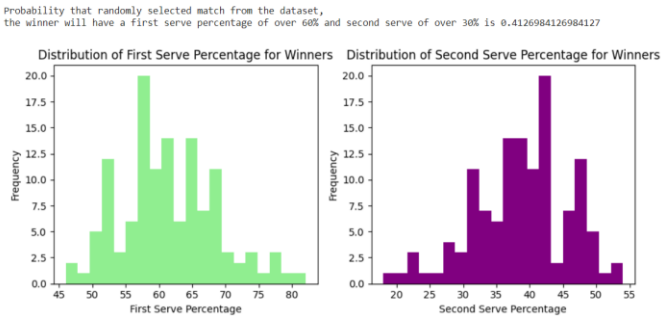
R. *KMeans.labels_:It receives a label as the index of the cluster it gets assigned to.*

S. *Seaborn : Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics[5].*

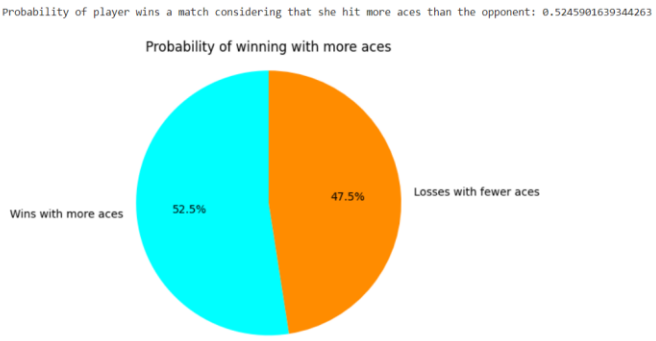
T. *colorbar(): Colorbars are a visualization of mapping from scalar values to colors.*

IV. ANSWER TO THE QUESTIONS

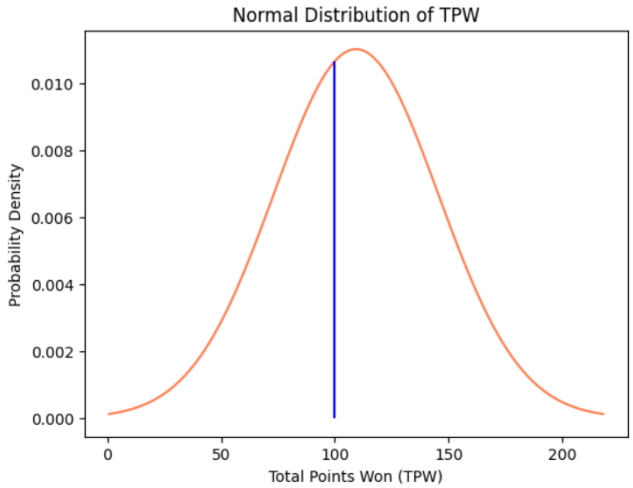
A. *Answer 1: As the probability of a winner having a first serve percentage of over 60% and a second serve percentage of over 30% in a randomly selected match from the dataset is approximately 0.4. It lies between 0.2 to 0.6 so our hypothesis is correct. From the graphs we can visualize that there are large number of winners above 60% and 30% in the first serve and second serve percentage winner respectively.*



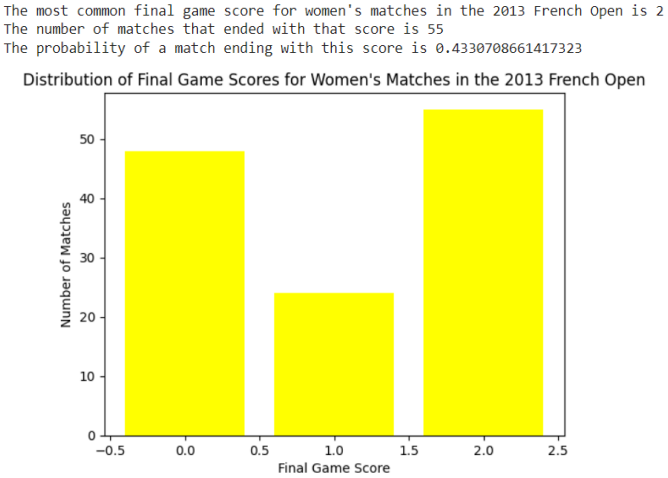
B. *Answer 2: The probability of the player wins a match considering that she have hit more aces than the opponent is nearly 0.52. It can be visualise from the pie chart provided below .*



C. *Answer 3: The probability of winning with total point won by player is 100 is approx 60.14% . The graph of the following can be observed below. The graph is of the normal distribution provided.*



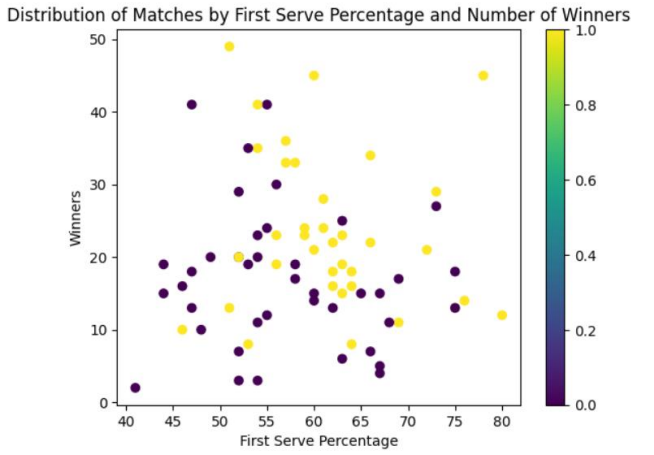
D. *Answer 4: The most common final game score for women's matches in the 2013 French Open is 2. This can be visualize by the graph provided below. The number of matches that ended with that score is 55. The probability of a match ending with this score is 0.43.*



E. *Answer 5: Based on the players' created and won break points, we can see from the scatterplot that there are three different player groups. Players with few break points won and few break points made are represented by the cluster in the lower left corner of the plot. Players with a high break point creation rate and a high break point win rate are shown by the cluster in the top right corner. A player with high break points created but low break points won is represented by the cluster in the bottom right corner.*

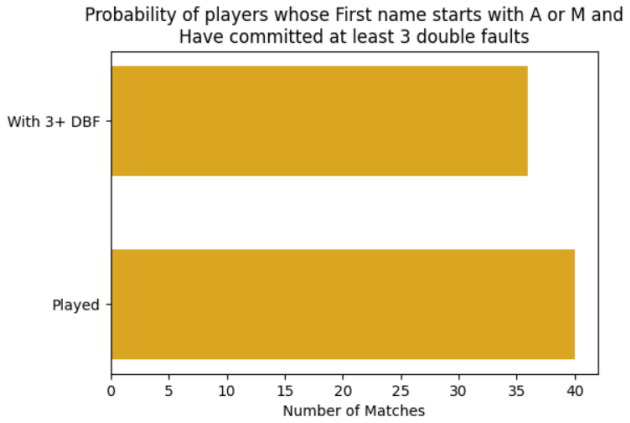
F. *Answer 6: The joint probability that a player wins the match given that she wins at least 70% of her first serves and hits at least 10 winners is 0.0789 which is very low stating that only few players are best players among them. This could be visualize by the graph given below. From the graph it is clearly be seen that there are more winners in the range of 10 to 40 who have more than 40% of the first serve percentage.*

```
Joint probability of winning the match given winning at least 70% of first serves and hitting at least 10 winners: 0.0789
```



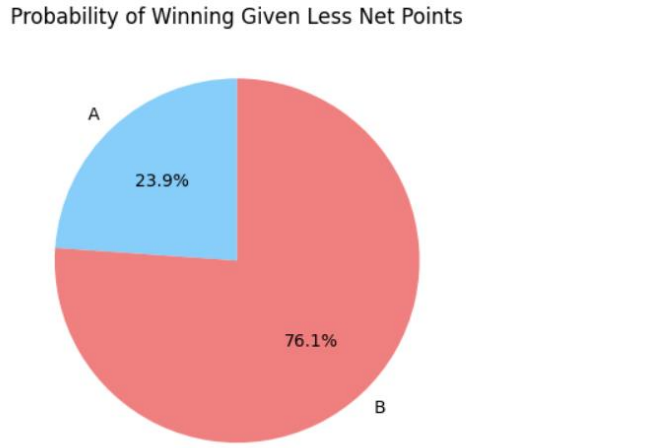
G. *Answer 7: The probability of players whose first name starts with A or M and have committed at least 3 double faults is 0.9 which is very high. This could be visualize from the graph provided below.*

```
Probability of players whose first name starts with A or M and
Have committed at least 3 double faults: 0.9
```



H. *Answer 8: The probability of winning the match even if the player have less net points than her opponent is nearly 0.23. This could be visualize by the pie chart provided. As the probabilty is less but not very low states that a player should not loss hope of winning the match taking into consideration that she have less net points.*

```
Probability of Winning Given Less Net Points is: 0.23893805309734514
```



V. SUMMARY OF THE OBSERVATION

A. *Observation 1: From the answer we can state that having a higher first serve percentage and a higher second serve percentage gives a player an advantage in a tennis match, as it allows them to start the point in a favorable position and avoid giving free points to their opponent. So, it is reasonable to say that the results of the probability calculation suggest that having a strong serve can give a player an advantage in winning a tennis match*

B. *Observation 2: From the pie chart we can observe that even if the player hits more aces the probability of winning is still the same. So there is no coreation between hitting more aces and winning the match.*

C. *Observation 3: The probability of winnining with TPW 0f 100 is 60.14% states that if the player have more TPW score then he have more chances of winning the match.*

D. *Observation 4: The most common final score for women's matches in 2013 is 2 which states that final score can only exceed till 2. And it is little easiler to score 2 points in the game.*

E. *Observation 5: The scatterplot can be useful in analyzing a player's performance in terms of their ability to create and win break points, which is an important aspect of the game. It can also provide insights into the playing style of different players and how they approach their service games and return games.*

F. *Observation 6: From observing the graph we can say that there are many player who have less than 70% of the first serve and who have hit more than 10 winners. This states that first serve hit is less correlated to winning the match who have hit more than 10 winners.*

G. *Observation 7: It is interesting to note that the proportion of players who have committed double faults and there name starts with A or M is generally high and can be related of their first name.*

H. *Observation 8: A player shouldn't give up on winning the game despite having less net points because the probability is smaller but still not very low.*

VI. REFERENCES

A. *[1] "Pandas Documentation." pandas documentation - pandas 1.5.3 documentation. Accessed February 23, 2023. https://pandas.pydata.org/*

B. *[2] "Matplotlib 3.7.0 Documentation." Matplotlib documentation -Matplotlib 3.7.0 documentation. Accessed February 23, 2023. https://matplotlib.org/*

C. *NumPy documentation. Accessed February 23,2023 . https://numpy.org/*

D. *[3] "SciPy." SciPy, https://scipy.org/. Accessed 22 Apr. 2023.*

E. *[4] Scikit-learn. 2019. "Scikit-Learn: Machine Learning in Python." Scikit-Learn.org. 2019. https://scikit-learn.org/stable/*

F. *[5] Seaborn. 2012. "Seaborn: Statistical Data Visualization — Seaborn 0.9.0 Documentation." Pydata.org. 2012. https://seaborn.pydata.org/.*

VII. ACKNOWLEGEMENT

A. *The Python Software Foundation, which is responsible for the creation and upkeep of the Python programming language.*

B. *The designers and collaborators of the Pandas, Matplotlib, and Numpy libraries for their support of the Python ecosystem.*

C. *Thank you to each and every one of my teachers for teaching me and assisting me in reaching my goals. Thank you to my friends and other students for creating moments with me that I will always treasure.*

D. *I would like to thank GeeksforGeeks and w3schools for offering such a fantastic environment for learning and developing programming skills. That is very fantastic that you are so committed to make computer science available to everyone.*

E. *I want to express my gratitude to my mother for her unwavering love, unending help, and constant presence in my life. I appreciate everything you have done for me and for being my life's inspiration.*