Data Narrative 3

Bhavik Patel, Btech Mechanical Engineering
Roll no.-22110047
Course: ES 114
Prof.- Shanmuga R , IIT Gandhinagar

I. OVERVIEW OF THE DATASET

The dataset is about tennis major tournament match statistics, including various features such as the players' names, set results, first serve to win, the total number of matches, the tournament, the round, the score, the number of aces, and many more. The dataset is a collection of 8 files containing the match statistics for women and men at the four major tennis tournaments of 2013. Each file has 42 columns, a minimum of 76 rows, and a maximum of 127. The dataset includes data from four major tennis tournaments, the Australian Open, the French Open, Wimbledon, and the US Open.

II. SCIENCTIFIC QUESTION/HYPOTHESES

A. Can we predict who will commit more no. of double faults, men or women? In which tournament the number of double faults committed were maximum?

Approach:

1. First, we will add the double faults committed by player 1 and player 2.
2. Then, we will plot the barplot for all 4 tournaments in pairs
3. We will also plot a pie chart to visualize and compare double faults men and women commit.

Solution:

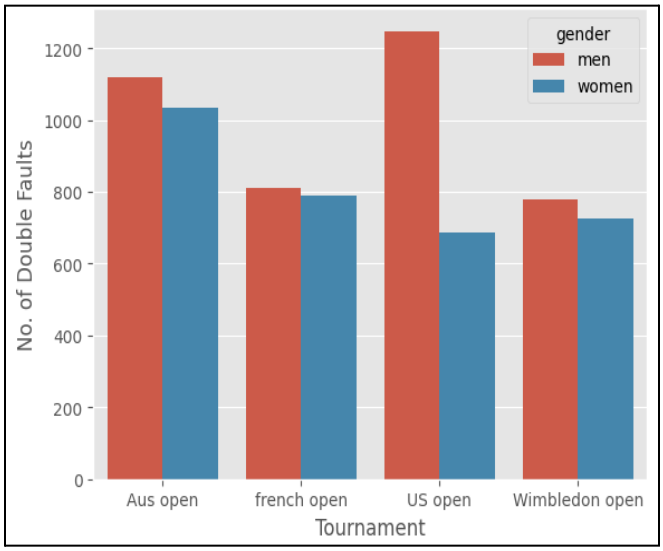

Fig. 1- No. of double faults

- From the above figure, we can see the difference between no. of double faults committed by men and women.
- The maximum no. of double faults was committed in Australian Open.
- It can be seen that men have committed more no. of double faults than women, but the difference is not much more significant except for the US Open, where men have committed as much as twice the no. of double faults.

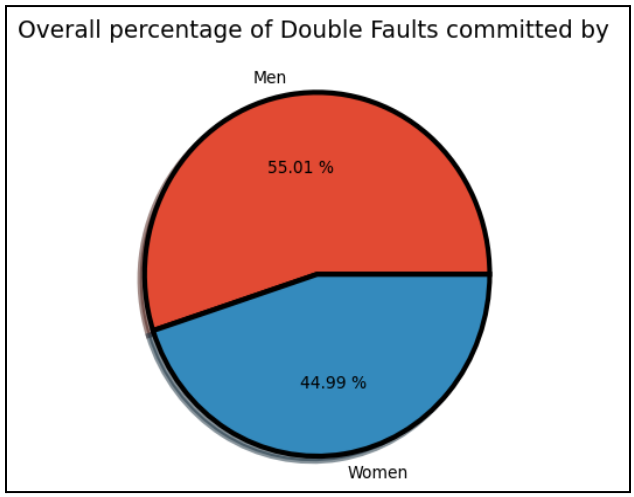

Fig. 2- Pie chart representing overall double faults committed by Men and Women

- Figure 2 also shows that there is not much difference between men and women. Therefore, we can't predict who will commit more no. of double faults.
- But in the case of the US Open, we may predict that men will commit more double faults than women.

B. What can we say about the number of double faults and aces increasing or decreasing as the tournament proceeds to further rounds?

Approach:

1. We will merge all the data frames first to answer the question for all the tournaments together.
2. Before merging, we will add two columns named 'gender' and 'tournament' in each data frame to distinguish them in the merged data frame.
3. For merging, we will use the 'concat' function of pandas.

```
df_merge=pd.concat([aus_men,french_men,...],axis=0)
df_merge=df_merge.reset_index(drop=True)
```

4. Then, we will calculate the number of aces and double faults for each round in the merged dataset.
5. We will use Seaborn's barplot to compare different tournament rounds.
6. We will use the Aces to Double Faults ratio to see the correlation between Aces and double faults.

Solution:

- Figure 3 shows the variation in the number of double faults and aces in different tournament rounds.
- We can see that double faults and aces decrease as the tournament progresses to further rounds.
- After round 4, the double faults committed increased, showing that the players were trying to win an ace at the tournament's final stage.
- Still, the trend is not the same for aces. The number of aces decreased continuously as the tournament progressed, except in round 6, when there was a huge increase in the number of aces.

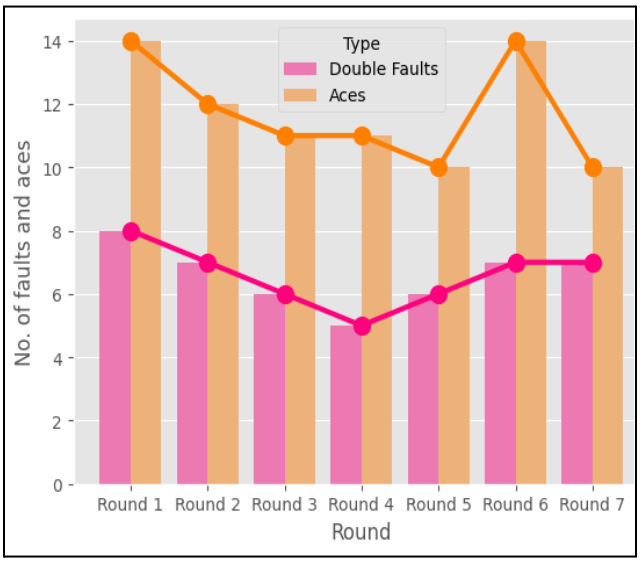

Fig. 3: Number of aces and faults in different rounds

- The number of aces is the least at the 7th round, i.e., in finals
- In Figure 4, we can't see any strong relationship between the number of aces and double faults. But it is essential to notice that ratio of aces to double fault is the least in finals.
- This shows that players in finals don't try to attempt many Aces to minimize the risk of double faults.

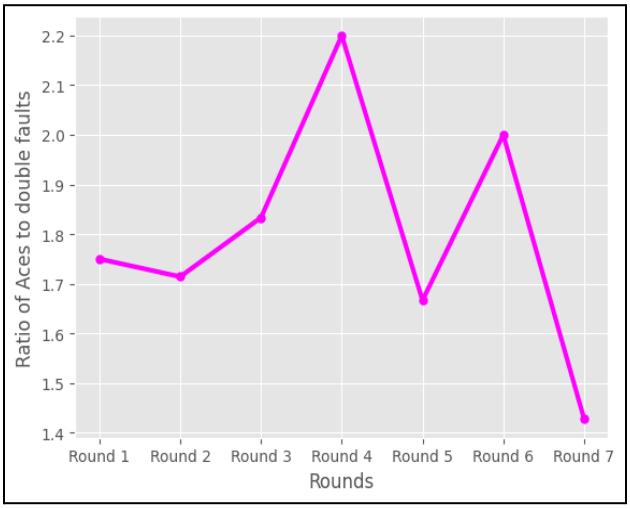

Fig. 4: Ratio of the number of Aces to Double Faults in different rounds.

C. *Take any one tournament and the two finalists of that tournament based on their individual skills and capabilities. Does the finalist who performed better in league matches was able to win the title, or the other finalist won it?*

Hypothesis: The player who performed better would more likely win the final.

Approach:

1. We will use the women's French Open for the above question.
2. We will find the two finalists.
3. To compare them, we will then use the number of Aces, Double Faults, and unforced errors.

Unique approach: The above-mentioned features for comparison are based on individual skills, not opponents.

Solution:

- The below three figures show the line plots of no. of Aces, Double Faults, and Unforced Errors of the two finalists, Serena Williams and Maria Sharapova.
- If we see Figure 5, almost all the rounds, Serena Williams served more Aces than Maria Sharapova.
- The average number of Aces served by Williams was about 6, and that of Maria was around 4 per match.

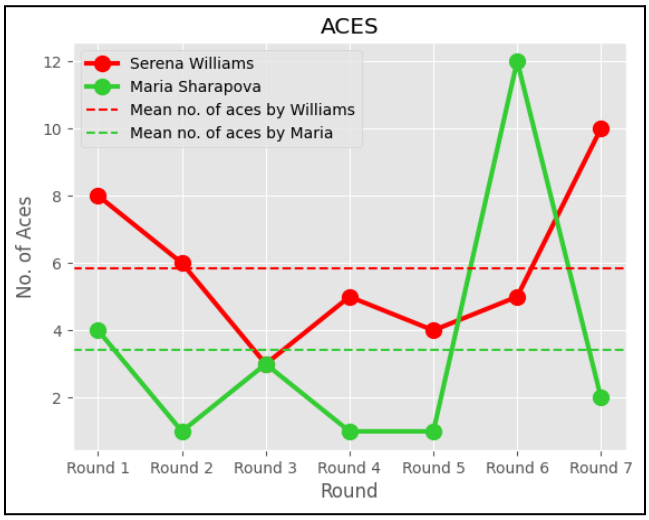

Fig. 5- No. of Aces in different rounds

➢ Figure 6 shows the number of double faults, and Williams has committed much fewer double faults than Maria.

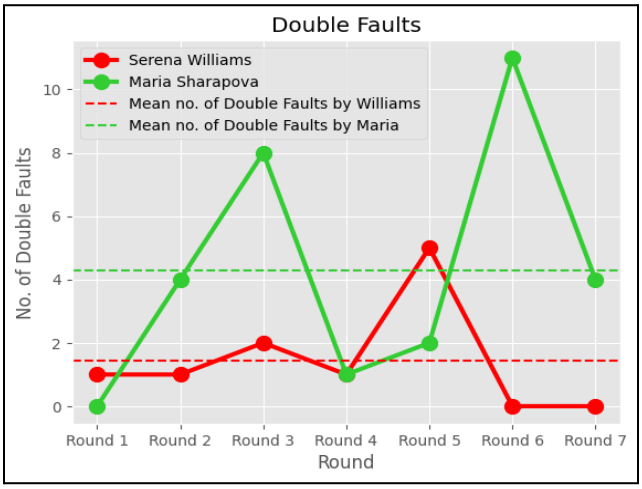

Fig. 6- No. of Double Faults in different rounds

➢ Figure 7 also shows that Williams has committed fewer unforced errors than Maria.

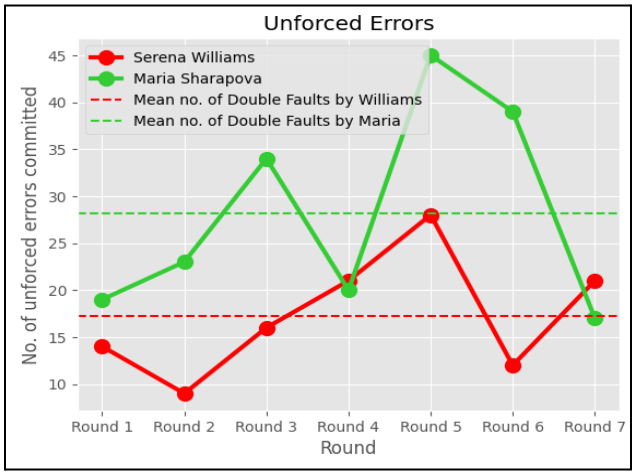

Fig. 7- Scatter plot b/w avg. salary and no. of faculty

➢ So from the above three plots depicting a comparison between the two finalists, Williams has performed better than Maria in terms of individual skills in the tournament.

➢ We can say that the hypothesis we considered at the start of the question is true, as Serena Williams won the tournament.

D. Can we predict the result of the match by just seeing the statistic of the match? Make a model with that dataset that can also help most accurately predict the results for other datasets. Tell the dataset which is used to train the model.

Unique Approach:

1. We will use a linear regression model from sklearn to make a model that can predict the result of the match.
2. The features of our model would be first serve percentage, aces, double faults, and break points won.
3. The target of the model is the result.
4. We will use the merged dataset in which all the datasets are combined for testing the model and predicting the results.
5. First, we will train the model with all 8 datasets one by one and the mean squared error and store it in a list.
6. We will use that dataset to train the model, which gives us the minimum error.

Solution:

➢ Figure 8 shows the mean square error of the models trained from different datasets.

➢ The men's and women's Australian Open and men's US Open datasets have the least mean squared error.

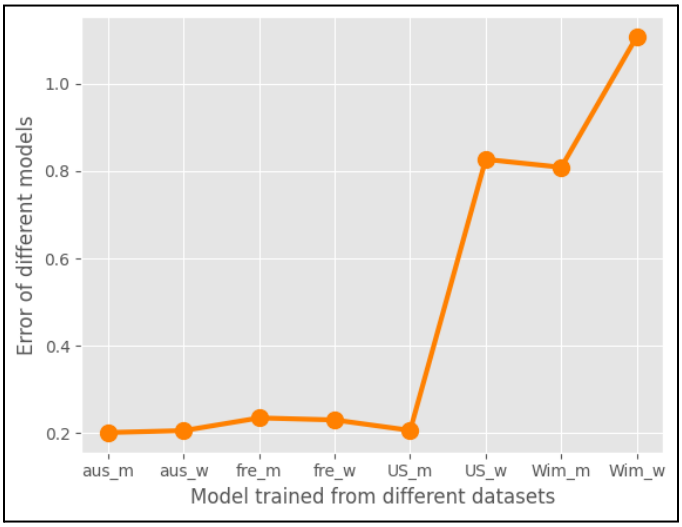

Fig.8: Mean squared error of different models.

➢ So, we will use the men's Australian Open dataset to train the model.

➢ So, to predict the result, we will use model coefficients corresponding to the feature matrix and model intercept in the below formula to predict the result.

```
x1= First serve percentage; x2= Aces
x3= Double Faults; x4= Unforced Errors
pred_result = model.intercept_ + model.coef_[0]*x1 + model.coef_[1]*x2
                               +model.coef_[2]*x3 + model.coef_[3]*x4
```

- The model coefficients are array([0.0021, 0.0383, -0.0167, 0.1508])
- The model intercept is -0.1762.

NOTE: The pred_result will not give us 0 or 1 but will give the probability. The more the pred_result is closer to one, the more the probability of winning the match.

E. What is the probability that a player will lose the match after losing a particular set? Which set is the most crucial set of the match that the player must win otherwise, he/she may lose the match?

Approach:

1. We will use the merged dataset that contains all the datasets for this question.
2. First, we will find the total number of matches in which the player lost the match after losing the particular set.
3. We will divide it by the total number of matches to get the probability.

Solution:

- From Figure 9, we can see that the probability of losing the match is highest when the player has lost set 2.
- So we can infer that the most crucial set of the match is set 2, which the player must win to increase his chances of winning.
- The next crucial set is set1, where if we lose the set, the probability of losing the match is around 0.65.

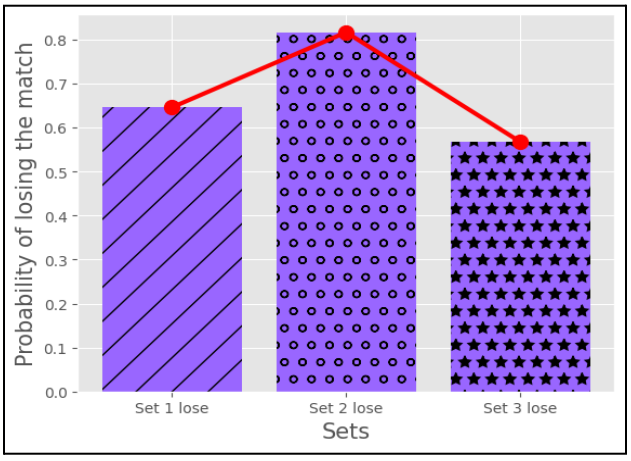

Fig. 9

F. What are the probabilities of the match to get over in Set3, Set4, and Set5 in the Men's US Open and Wimbledon Open?

Approach:

1. We will first find the number of matches that ended in set 3, set 4, and set 5 for the respective tournaments.
2. Then divide them by the total number of matches played in the tournament to get the probability.

Solution:

- Figure 10 shows the probability of the match getting over in set3, set4, and set5 in the Men's US Open and Wimbledon Open.
- We can see the probability of a match going to higher sets is decreasing. So we can infer that most matches get over in the 3rd or 4th set, and very few matches go up to set 5.
- Comparing the two tournaments we can see that the Wimbledon Open has more chances of getting the match over in the 3rd set with a probability around 0.6. This tells us that more than half of the matches ended in 3rd set in the Wimbledon Men's Open.

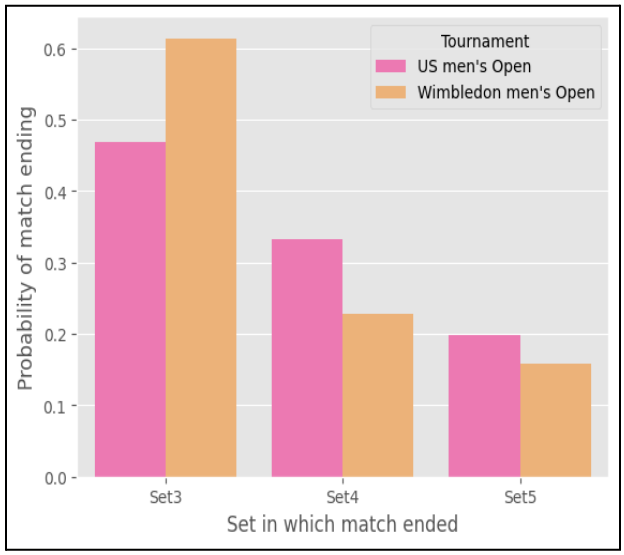

Fig. 10

G. Which are the top 10 players who have won the most matches combining all the tournaments in 2013? How many of them were able to win any of the tournaments?

Approach:

1. We will use the merged dataset that contains all the datasets for this question.
2. Using the value_counts( ) function, we will find the top 10 players who have won the most matches.
3. Then, we will check how many of them have won the tournament.

Solution:

- Figure 11 shows the player with the most wins. It is clearly seen that Rafael has won the most number of matches, that is 20.
- The figure also shows highlights the player who was able to win at least 1 tournament.
- The players who won the tournament were Rafael Nadal, Stanislas Wawrinka, and Serena Williams. The rest of the players in Figure 11 also outperformed in 2013 but, unfortunately, could not win any tournaments.

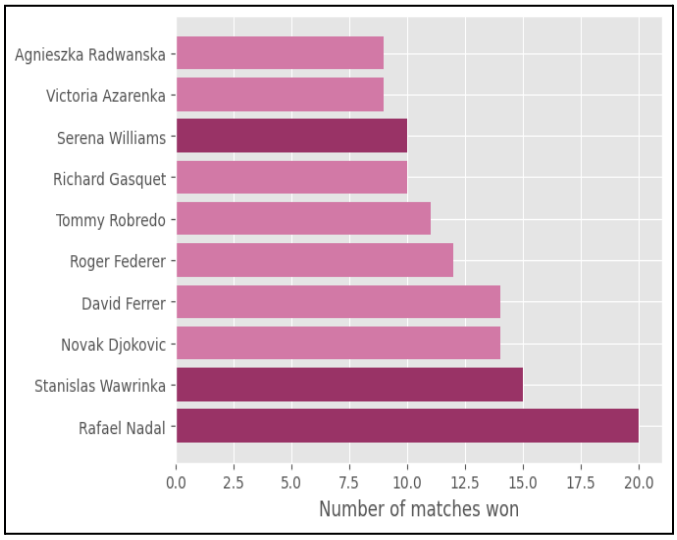

Fig. 11

H. Compare the performance of Rafael Nadal in the Australian and French Men's Open. Can you tell which tournament is more challenging from the comparison?

Approach:

1. First, we will take Australian Open and find the number of Aces, Double Faults, and breakpoints won and the first serve percentage of Rafael Nadal.
2. Then we will also do the same to find for French Open.
3. We will use the matplotlib bar plot to compare both tournaments.

Solution:

- Figure 12 compares Rafael Nadal's performances in the Australian and French Open.
- We can see that Nadal performed well in French Open in terms of winning breakpoints from opponents and percent of first serve won.
- The number of Aces and Double Faults are more or less the same in both tournaments.

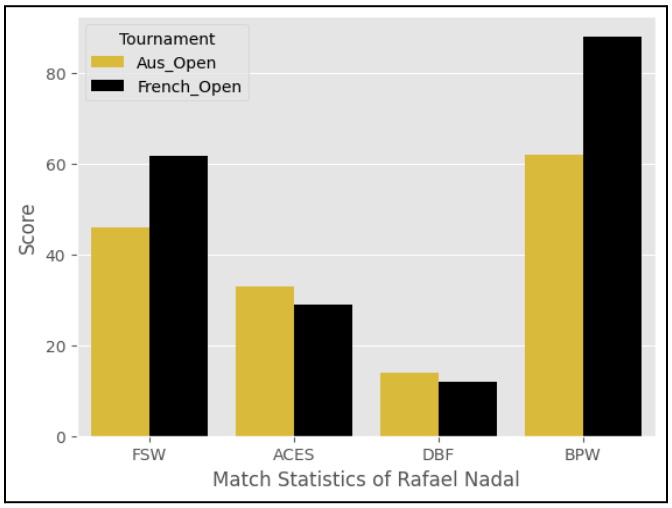

Fig. 12

- Now, to compare both tournaments to find which was challenging, we will break point won and first serve won criteria as these criteria also depend upon the opponents.
- We can see from Figure 12 that it is relatively difficult to win break points and first serve in Australian Open.
- Hence, Australian Open was more challenging than French Open.

III. DETAILS OF LIBRARIES

1. PANDAS: Pandas is a popular open-source Python library for data manipulation, analysis, and preparation. It provides fast and flexible data structures for structured data, such as tables, time series, and matrices, and offers tools for data cleaning, exploration, and visualization. The two primary data structures in pandas are Series and DataFrame.[2]

2. MATPLOTLIB: Matplotlib is a well-known open-source Python library for creating high-quality data visualizations. Line plots, scatter plots, bar charts, histograms, and other plot types are available. Matplotlib is built to work with NumPy arrays and can create static and interactive visualizations. Some major features of this library are easy customization supports, multiple output formats, etc.

3. NUMPY: NumPy is a widely used open-source Python library for numerical computation. It provides a quick and efficient way to work with numeric data arrays and matrices, as well as various tools for mathematical operations, random number

generation, linear algebra, and more. Some key features are multidimensional array support, broadcasting, mathematical operations, and random number generation.

4. SEABORN: Seaborn is a popular open-source Python data visualization library. It is based on Matplotlib and offers a high-level interface for creating informative and appealing statistical graphics. Seaborn is built to work well with Pandas data structures and can easily handle large datasets.

IV. DETAILS OF FUNCTIONS

1. *read_csv*: The read_csv function is a method provided by the Pandas library in Python for reading and parsing data from a comma-separated values (CSV) file. It is a flexible function that can handle various input data formats and options.

2. *plt.plot:* The plt.plot function is a Python method provided by the Matplotlib library for creating line plots. It is a simple, flexible function that handles various input data formats and options. The plt.plot function takes one or more arrays of x and y values and creates a data line plot. It provides various options for customizing the appearance of the plot, such as setting the color, line style, marker style, and label.

3. *value_counts():* The value_counts() function is a Python method the Pandas library provides for calculating the frequency of unique values in a Pandas Series. It returns a new Series object that contains the count of each unique value in the original Series.

4. *drop_duplicates():* The drop_duplicates() function is a Python method the Pandas library provides for removing duplicate rows from a data frame. It is a powerful tool for cleaning and preprocessing data and is commonly used in data analysis and exploration tasks.

5. *merge():* The merge() function is a method provided by the Pandas library in Python for combining two or more data frames based on a common set of columns. It is a powerful tool for merging and joining data from different sources and is commonly used in data analysis and exploration tasks.

6. *Scatterplot:* A scatter plot is a type of data visualization that displays the relationship between two variables in a dataset. It is a graph that uses dots or markers to represent individual data points, with the position of each dot representing the value of the two variables it corresponds to.

7. *Kernel density plot:* A kernel density plot, also known as a density plot, is a graph showing the probability density function of a continuous variable. It is used to visualize the distribution of a dataset by estimating the probability density function of the underlying variable. A smooth curve is drawn over the data points in a kernel density plot, with the area under the curve representing the variable's probability density at each point. The shape of the curve depends on the bandwidth, or smoothing parameter, used in the kernel density estimation.

V. REFERENCE LIST

1. "UC Irvine Machine Learning Repository." https://archive-beta.ics.uci.edu/dataset/300/tennis+major+tournament+match+statistics

2. "Pandas Documentation — Pandas 1.5.3 Documentation," n.d. https://pandas.pydata.org/docs/

3. "Matplotlib — Visualization with Python," n.d. https://matplotlib.org/.

4. "NumPy Documentation — NumPy v1.24 Manual," n.d. https://numpy.org/doc/stable/

5. "Seaborn: Statistical Data Visualization — Seaborn 0.12.2 Documentation," n.d. https://seaborn.pydata.org

VII. ACKNOWLEDGEMENT

I sincerely thank Professor Shanmuga R. for his invaluable assistance in completing this assignment. His guidance and support were instrumental in enabling me to develop a deeper understanding of the course material. His willingness to share his time, knowledge, and expertise with me was inspiring.