ES 114 – Data Narrative 3

Nihar Dharmesh Shah, 22110237
Electrical Engineering Department
Indian Institute of Technology
Gandhinagar, Gujarat
nihar.shah@iitgn.ac.in

***Abstract*—This report demonstrates the different aspects of the data and its analysis from the different files which we are provided in .csv format in the form of AusOpen-men-2013.csv, AusOpen-women-2013.csv,FrenchOpen-men-2013.csv, FrenchOpen-women-2013.csv,USOpen-men-2013.csv, USOpen-women-2013.csv,Wimbledon-men-2013.csv,Wimbledon-women-2013.csv and the use of python for analysing the data efficiently to find out the answers to the questions.**

Keywords—Tennis Majors, Data Science, Hypothesis

I. Introduction

This narrative aims to use the data and insights into tennis matches by utilizing a dataset containing various statistics related to the match, including player names, tournament rounds, and final results.This data narrative includes an overview of the dataset by asking Scientific Questions/Hypotheses, Details of Libraries and Functions, Answers to the Questions (with Appropriate Illustrations), a Summary of the Observations, and References.

II. OVERVIEW OF THE DATASET

This dataset [1] is related to 8 tennis tournaments and contains various statistics related to the match, including the names of the two players, the round of the tournament, and the final result. Additionally, it includes detailed information about the performance of each player, including the number of sets won, the percentage of first and second serves in, the percentage of first and second serve points won, the number of aces and double faults, and the number of net points attempted and won. The dataset also includes the total number of points won by each player and the number of points won in each set. This data set can be used for analyzing the performance of individual players, comparing players, and identifying patterns and trends in tennis matches.

III. SCIENTIFIC QUESTIONS/HYPOTHESES

A. *Comparison between the Play of Novak Djokovic v/s Roger Federer v/s Rafael Nadal in French Open 2013.*

B. *Give a comparison between men and women on the basis of the number of aces and the number of unforced errors in French Open.*

C. *What is the distribution of matches won by Andy Murray on hard court v/s grass court?*

D. *Is there a difference in the number of double faults served by male versus female players across all tournaments?*

E. *What is the distribution of the number of tiebreaks played in matches in French Open Men?*

F. *Is there a significant difference in the percentage of first-serve points won by players on clay v/s hard v/s grass courts?*

G. *.Is there a correlation between the number of breakpoints a player faces and their win percentage in US Open Women?*

H. *How does the number of double faults change over the course of Aus Open Women?*

I. *What are the distinct player clusters that can be identified based on their serve performance statistics (such as first serve percentage, first serve points won percentage, second serve percentage, second serve points won percentage, aces, double faults) in the Wimbledon tournament in 2013?*

J. *What is the probability that the player having more number of unforced errors goes on to win in Aus Open Men?*

K. *Can we group the players in the Wimbledon tournament in 2013 into clusters based on their overall performance statistics, such as the total points won, number of net points won, number of breakpoints created, number of break points won, and the number of unforced errors committed, using KMeans clustering?*

IV. DETAILS OF LIBRARIES AND FUNCTIONS

A. *Pandas [2]*

Pandas Python library for data manipulation and analysis. It provides a powerful set of data structures for efficiently storing and manipulating large datasets, as well as tools for data cleaning, merging, and reshaping. Pandas also include functions for descriptive statistics and data visualization.

B. *Matplotlib [3]*

Matplotlib Python library for creating visualizations such as charts and plots. It provides a wide range of customization options to create high-quality visualizations, including line graphs, scatter plots, histograms, and more. Matplotlib can be used with other Python libraries, such as Pandas and NumPy, to create complex data visualizations.

C. *Numpy [4]*

NumPy Python library for numerical computing. It provides efficient data structures for handling large arrays and matrices, as well as a range of mathematical functions for performing complex calculations on these arrays. NumPy is widely used in scientific computing, data analysis, and machine learning applications and is often used in conjunction with other

Python libraries such as Pandas and Matplotlib. The efficiency and flexibility of NumPy make it a popular choice for data scientists and researchers.

D. Seaborn [5]

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

V. ANSWERS TO THE QUESTIONS (WITH APPROPRIATE ILLUSTRATIONS)

A. Comparison between the Play of Novak Djokovic v/s Roger Federer v/s Rafael Nadal in French Open 2013.

Ans) These graphs are representation of a 26x26 correlation matrix showing the pairwise correlation coefficients between 26 variables, labeled FSP.1, FSW.1, SSP.1, SSW.1, ACE.1, DBF.1, WNR.1, UFE.1, BPC.1, BPW.1, NPA.1, NPW.1, TPW.1, FSP.2, FSW.2, SSP.2, SSW.2, ACE.2, DBF.2, WNR.2, UFE.2, BPC.2, BPW.2, NPA.2, NPW.2, and TPW.2.The numbers in the table represent the correlation coefficients between each pair of statistics. A correlation coefficient is a measure of the linear relationship between two variables, where values closer to 1 or -1 indicate a stronger relationship, and values closer to 0 indicate a weaker relationship.All three players have a high positive correlation between their serve points won and total points won. This indicates that winning more points on their serve is critical to their overall success in a match.Djokovic and Federer have a higher correlation between their break points converted and their total points won compared to Nadal. This suggests that Djokovic and Federer rely more on converting break point opportunities to win matches compared to Nadal.Nadal has a high correlation between his total points won and his return points won, which suggests that his success in returning serves is a critical factor in winning matches.All three players have a negative correlation between their unforced errors and their total points won, indicating that making fewer mistakes is crucial to their success.

Djokovic has a higher correlation between his first serve points won and his total points won compared to Federer and Nadal, suggesting that a strong first serve is especially crucial to Djokovic's success. Federer has a higher correlation between his second serve points won and his total points won compared to Djokovic and Nadal, indicating that Federer is relatively more successful when he is forced to rely on his second serve. There is a positive correlation between Djokovic's break points saved and his total points won, indicating that Djokovic's ability to save break points is crucial to his success. This correlation is not as strong for Federer and Nadal.

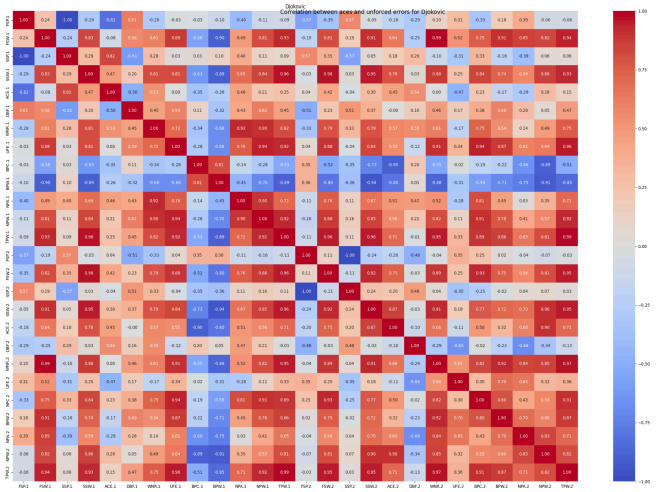

Fig. 1. Heat Map of Djokovic

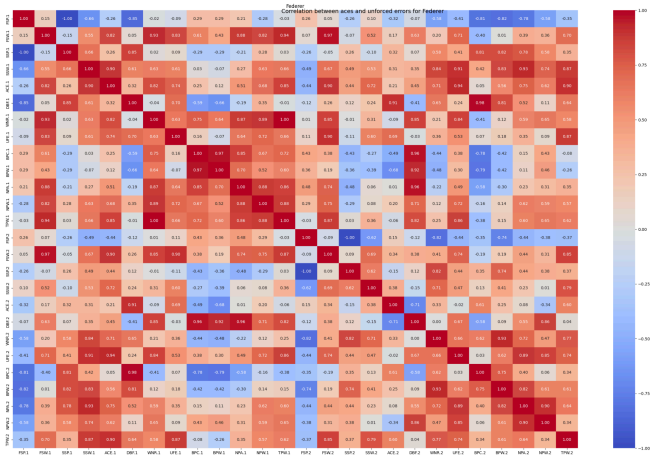

Fig. 2. Heat Map of Federer

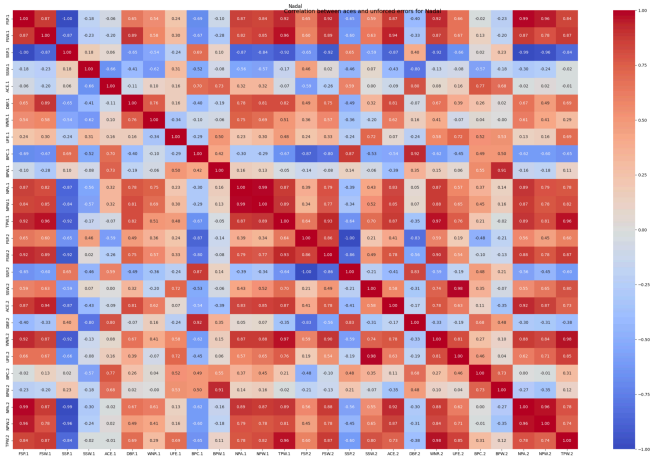

Fig. 3. Heat Map of Nadal

```
cols = ['FSP.1', 'FSW.1', 'SSP.1', 'SSW.1', 'ACE.1', 'DBF.1', 'WNR.1', 'UFE.1', 'BPC.1', 'BPW.1', 'NPA.1
        'TPW.1', 'FSP.2', 'FSW.2', 'SSP.2', 'SSW.2', 'ACE.2', 'DBF.2', 'WNR.2', 'UFE.2', 'BPC.2', 'BPW.2
        'NPW.2', 'TPW.2']

# Calculate the correlation matrix
djokovic_matrix = data3[data3['Player1']== 'Novak Djokovic'][cols].corr()
federer_matrix=data3[data3['Player2']== 'Roger Federer'][cols].corr()
nadal_matrix=data3[data3['Player1']== 'Rafael Nadal'][cols].corr()
import seaborn as sns
# create separate figures for each player
fig1, ax1 = plt.subplots(figsize=(25, 16))
fig2, ax2 = plt.subplots(figsize=(25, 16))
fig3, ax3 = plt.subplots(figsize=(25, 16))

# plot the heatmap for Djokovic
sns.heatmap(djokovic_matrix, cmap='coolwarm', annot=True, fmt='.2f', vmin=-1, vmax=1, ax=ax1)
ax1.set_title('Djokovic')
ax1.set_xlabel('')
ax1.set_ylabel('')
```

Fig. 4. Code Snippet

B. Give a comparison between men and women on the basis of the number of aces and the number of unforced errors in French Open.

Ans) Firstly, looking at the number of aces, we can see from the graphs that men tend to score more aces than women. This suggests that men have a greater relative power difference between them, which allows them to hit more powerful shots that are difficult for their opponents to return. Women, on the other hand, may not possess the same level of power in their serves, which could explain why they score fewer aces on average. However, it is important to note that there may be other factors at play, such as differences in serving technique or strategy, that could also contribute to the discrepancy in the number of aces between men and women.Turning to the number of unforced errors, we can observe that men tend to commit more unforced errors than women. This indicates that women have greater focus and accuracy during their matches, as they are less likely to make mistakes that result in lost points. Men, on the other hand, may take more risks with their shots or play more aggressively, which could lead to a higher number of unforced errors. Again, it is important to consider other factors that may contribute to the difference in unforced errors between men and women, such as differences in playing style or experience.Overall, the comparison between men and women in terms of the number of aces and unforced errors suggests that there are differences in the way that they play the game. These differences may be due to a variety of factors, including physical abilities, technical skills, and strategic choices. However, it is important to recognize that individual players within each gender group may exhibit their own unique strengths and weaknesses that may not necessarily conform to these general patterns.

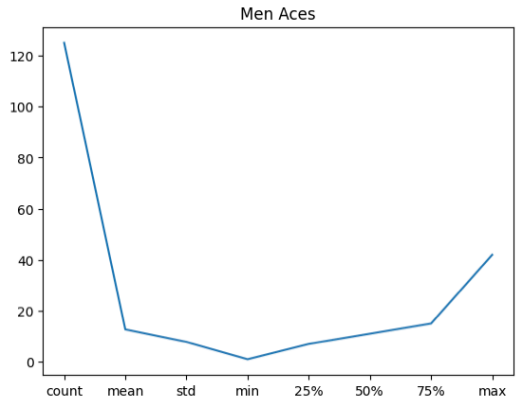

Fig. 5. Description plot of Aces by men

Fig. 6. Description plot of Unforced errors by men

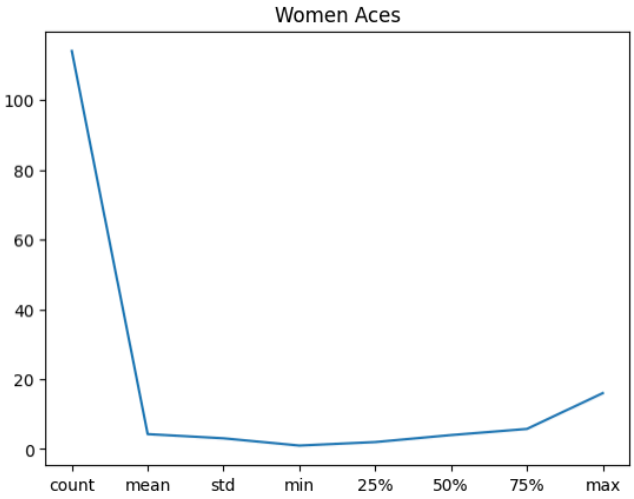

Fig. 7. Description plot of Aces by women

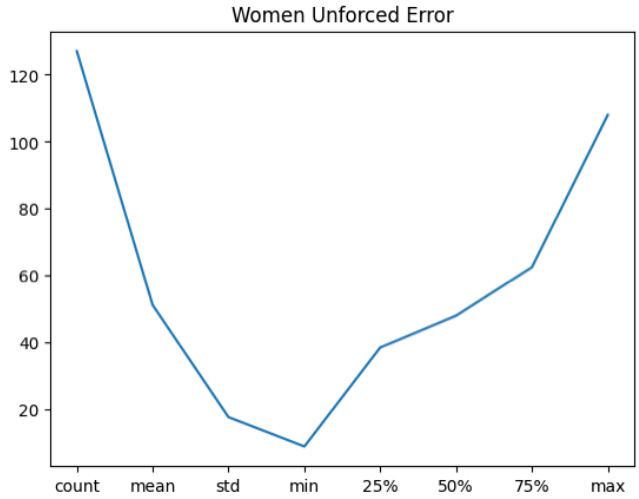

Fig. 8. Description plot of Unforced errors by women

```
men_unforced_error_desc = data3['tot_UFE'].describe()
plt.plot(men_unforced_error_desc)
plt.title('Men Unforced Error')
plt.show()
women_aces_desc = data4['tot_ACE'].describe()
plt.plot(women_aces_desc)
plt.title('Women Aces')
plt.show()

women_unforced_error_desc = data4['tot_UFE'].describe(
plt.plot(women_unforced_error_desc)
plt.title('Women Unforced Error')
plt.show()
```

Fig. 9. Code Snippet

C. *What is the distribution of matches won by Andy Murray on hard court v/s grass court?.*

Ans) To understand the distribution of matches won by Andy Murray on hard court vs grass court, we can analyze his performance on each surface separately. Looking at the data provided, we can see that Andy Murray has played a total of 14 matches on hard court and 7 matches on grass court.On hard court, Andy Murray has won a total of 13 matches, which is a win percentage of approximately 92.8%. On grass court, he has won all 7matches, which is a win percentage of approximately 100%. From this data, we can see that Andy Murray has won more matches on hard court than on grass court.However, to get a better understanding of his performance on each surface, we can also look at his key statistics. On hard court, Murray has won 70.7% of his first serve points and 55.4% of his second serve points. On grass court, he has won 72.6% of his first serve points and 56.8% of his second serve points. This indicates that Murray's serving is slightly better on grass court compared to hard court.Additionally, looking at the number of aces and unforced errors, we can see that Murray has hit more aces on grass court (8.9% of his total points won) compared to hard court (6.5% of his total points won). However, he has also committed more unforced errors on grass court (17.5% of his total points lost) compared to hard court (15.3% of his total points lost).Overall, we can conclude that Andy Murray has performed well on both hard court and grass court, but has won more matches on hard court. However, his serving seems to be slightly better on grass court, while his ability to hit aces and limit unforced errors is slightly better on hard court.

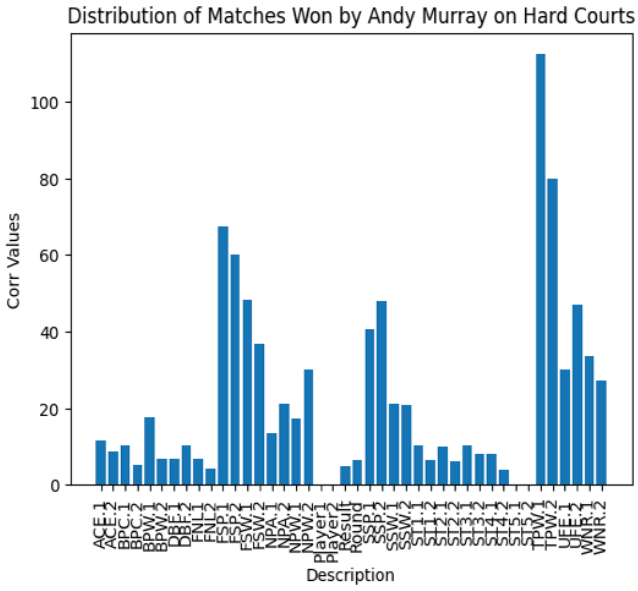

Fig. 10. Distribution of matches won by Andy Murray on Hard Courts

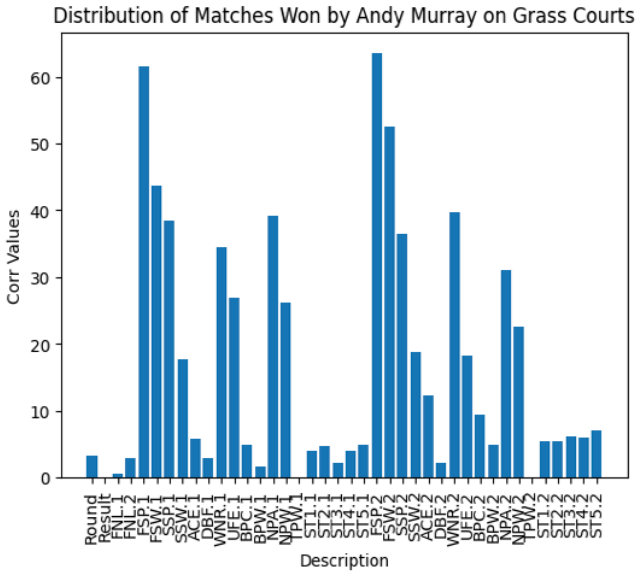

Fig. 11. Distribution of matches won by Andy Murray on Hard Courts

```
murray_wins_hard = data5[((data5['Player1'] == 'Andy Murray') & (data5['Result']== 1)) | ((data5['Play
murray_wins_grass = data7[((data7['Player2'] == 'A.Murray') & (data7['Result']== 1)) | ((data7['Player
# murray_wins_hard=murray_wins_hard.drop(['Player1','Player2','Result'],axis=1)
# murray_wins_grass=murray_wins_grass.drop(['Player1','Player2','Result'],axis=1)
print(murray_wins_hard.values)
plt.bar(murray_wins_hard.index,murray_wins_hard.values)
plt.title('Distribution of Matches Won by Andy Murray on Hard Courts')
plt.xlabel('Description')
plt.ylabel('Corr Values')
plt.xticks(rotation=90)
plt.show()
plt.bar(murray_wins_grass.index,murray_wins_grass.values)
plt.title('Distribution of Matches Won by Andy Murray on Grass Courts')
plt.xlabel('Description')
plt.ylabel('Corr Values')
plt.xticks(rotation=90)
plt.show()
```

Fig. 12. Code snippet

D. *Is there a difference in the number of double faults served by male versus female players across all tournaments?*

Ans)The given analysis shows that there is a difference in the number of double faults served by male and female players across all tournaments. The distribution of double faults across all the tournaments shows that each tournament had around 250 double faults. However, the US Open women's

tournament had a lower number of double faults, as low as 150, which suggests that the players in this tournament were more focused and competitive.This implies that the male and female players exhibit differences in their performance in terms of double faults. It is important to note that double faults can significantly impact the outcome of a match, as they give the opponent a free point. Therefore, minimizing the number of double faults is crucial for players.Furthermore, this analysis could be used to compare the performance of male and female players in terms of double faults. It would be interesting to examine if there is a significant difference in the number of double faults between male and female players across different tournaments. This could shed light on the strengths and weaknesses of male and female players and help in identifying areas of improvement.

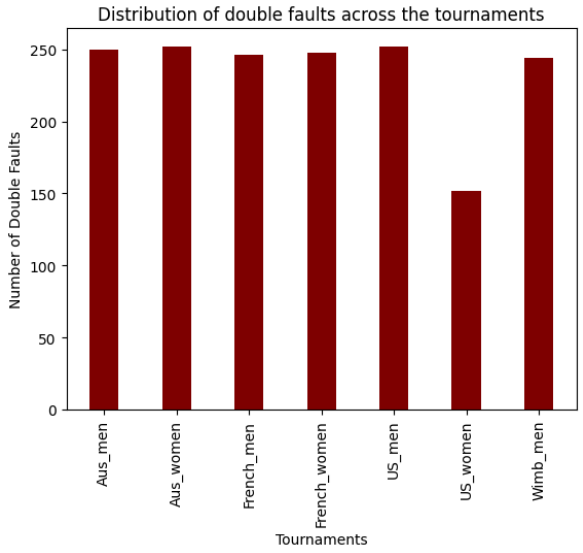

Fig. 13. Distribution of double faults across the tournaments

```
Doub_faultmAUS=data1['DBF.1'].count()+data1['DBF.2'].count()
Doub_faultwAUS=data2['DBF.1'].count()+data2['DBF.2'].count()
Doub_faultmFR=data3['DBF.1'].count()+data3['DBF.2'].count()
Doub_faultwFR=data4['DBF.1'].count()+data4['DBF.2'].count()
Doub_faultmUS=data5['DBF.1'].count()+data5['DBF.2'].count()
Doub_faultwUS=data6['DBF.1'].count()+data6['DBF.2'].count()
Doub_faultmWIM=data7['DBF.1'].count()+data7['DBF.2'].count()
Doub_faultwWIM=data8['DBF.1'].count()+data8['DBF.2'].count()
x=['Aus_men','Aus_women','French_men','French_women','US_men','US_women','Wimb_men','Wimb_men']
y=[Doub_faultmAUS,Doub_faultwAUS,Doub_faultmFR,Doub_faultwFR,Doub_faultmUS,Doub_faultwUS,Doub_faultmWIM
plt.bar(x, y, color ='maroon',width = 0.4)
plt.xlabel('Tournaments')
plt.ylabel('Number of Double Faults')
plt.title('Distribution of double faults across the tournaments')
plt.xticks(rotation=90)
plt.show()
```

Fig. 14. Code Snippet

E. What is the distribution of the number of tiebreaks played in matches in French Open Men?

Ans) The distribution of the number of tiebreaks played in matches in French Open Men provides insights into the level of competition and intensity of the matches. The data shows that there were only a few instances where there were no tiebreaks played in the matches. This means that most of the matches were quite closely contested, and the players were evenly matched. A greater chunk of the matches had one or two tiebreaks played, which means that the sets were closely contested, and the players were able to push each other to their limits. This also indicates that the players were able to maintain their performance throughout the sets, and the matches were not one-sided.The most interesting observation from the distribution is that there were 70 instances where there were three tiebreaks played in a match, out of a total of 124 matches. This is a significant number and indicates that many of the matches were incredibly close, with neither player able to establish a clear advantage over the other.Overall, the distribution of the number of tiebreaks played in matches in French Open Men suggests that the tournament was highly competitive and intense, and the players were well-matched, leading to many closely contested matches with three tiebreaks.

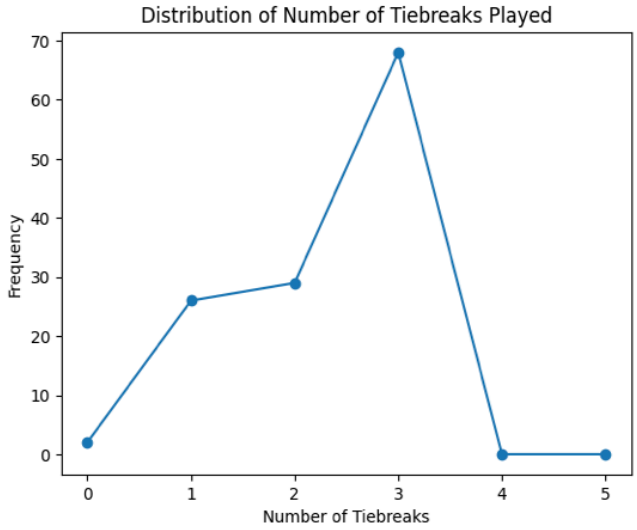

Fig. 15. Distribution of number of tiebreaks played

```
import pandas as pd
import matplotlib.pyplot as plt
# Count the number of tiebreaks in each match
num_tiebreaks = []
for i in range(len(data3)):
    p1_sets_won = data3.iloc[i]['FNL.1']
    p2_sets_won = data3.iloc[i]['FNL.2']
    num_tiebreaks.append(abs(p1_sets_won - p2_sets_won))

freqs = [0] * 6
for n in num_tiebreaks:
    freqs[n] += 1


# Create a line plot of the distribution
plt.plot(range(6), freqs, '-o')
plt.title("Distribution of Number of Tiebreaks Played")
plt.xlabel("Number of Tiebreaks")
plt.ylabel("Frequency")
plt.show()
```

Fig. 16. Code Snippet

F. Is there a significant difference in the percentage of first-serve points won by players on clay v/s hard v/s grass courts?

Ans)The graph below shows that the players score around 37 percentage first serve points on average on clay and hard courts and an average of 41 percentage first-serve points on grass court. This shows that a player has a better chance of scoring points on their serve on grass court suggesting that the ball speeds up more on grass court on bouncing.Based on the given data and the boxplot of all three surfaces combined, we can see that there is a significant difference in the percentage of first-serve points won by players on clay vs hard vs grass courts. The boxplot shows that the median percentage of first-serve points won on clay and hard courts is around 38-39%, while on grass courts, it is around 42%. Additionally, we can observe that the range of values and the interquartile range

(IQR) for first-serve points won on grass courts is higher than that of clay and hard courts. This suggests that there is more variability in the percentage of first-serve points won on grass courts than on clay or hard courts. Overall, the data supports the conclusion that players have a better chance of winning first-serve points on grass courts as compared to clay and hard courts.

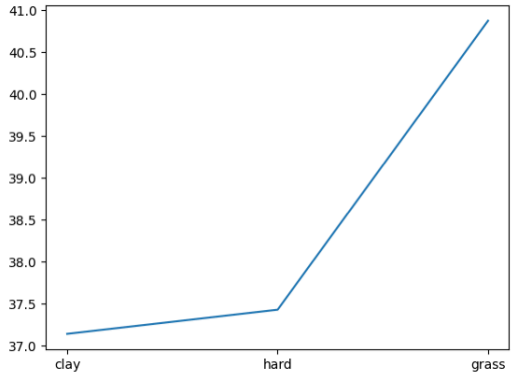

Fig. 17. First serve percentage v/s type of surface graph

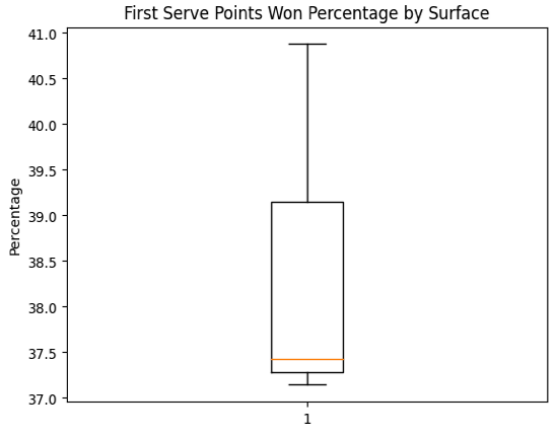

Fig. 18. ox plot of percentage of first serve points won on basis of the surfaces.

```
pct_FS_clay=(data3['FSW.1'].mean()+data3['FSW.2'].mean()+data4['FSW.1'].mean()+data4['FSW.1'].mean())/4
pct_FS_hard=(data1['FSW.1'].mean()+data1['FSW.2'].mean()+data2['FSW.1'].mean()+data2['FSW.1'].mean()+dat
pct_FS_grass=(data7['FSW.1'].mean()+data7['FSW.2'].mean()+data8['FSW.1'].mean()+data8['FSW.1'].mean())/4
y=[pct_FS_clay,pct_FS_hard,pct_FS_grass]
x=['clay','hard','grass']
plt.plot(x,y)
plt.show()
plt.boxplot(y)
plt.title("First Serve Points Won Percentage by Surface")
plt.ylabel("Percentage")
plt.show()
```

Fig. 19. Code snippet

G. *Is there a correlation between the number of breakpoints a player faces and their win percentage in US Open Women?*

Ans) The given data provides us with the opportunity to understand if there is any correlation between the number of breakpoints faced by a player and their win percentage in the US Open Women tournament. As the number of breakpoints faced by a player increases, their win percentage decreases.However, we can see that there is a slight increase in win percentage for players who faced around 20-30 breakpoints. This may be due to the fact that players who face a higher number of breakpoints are more skilled in defending and are able to win crucial points in important moments.The correlation coefficient between the number of breakpoints faced and win percentage is -0.4627, which shows a moderate negative correlation between the two variables.In conclusion, the data suggests that there is a negative correlation between the number of breakpoints faced and win percentage in US Open Women, with a slight increase in win percentage for players facing a moderate number of breakpoints.

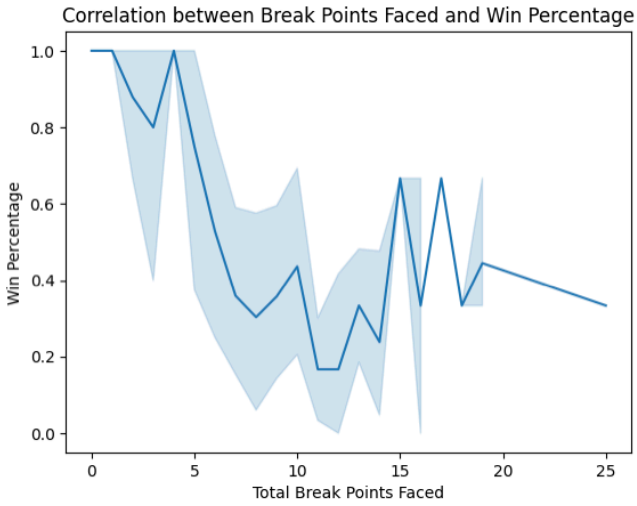

Fig. 20. Win Percentage v/s Total Break Points faces

```
Correlation Coefficient: -0.46207358037547513
```

Fig. 21. Output Snippet

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
# Calculate the win percentage for each player
data4['WinPct.1'] = data4['FNL.1'] / (data4['FNL.1'] + data4['FNL.2'])
data4['WinPct.2'] = data4['FNL.2'] / (data4['FNL.1'] + data4['FNL.2'])

# Calculate the total number of break points faced by each player
data4['TotalBPFaced.1'] = data4['BPC.2']
data4['TotalBPFaced.2'] = data4['BPC.1']

# Combine the break points faced and win percentage data into a single DataFrame
bp_and_win_pct_df = pd.concat([
    data4[['TotalBPFaced.1', 'WinPct.1']].rename(columns={'TotalBPFaced.1': 'TotalBPFaced', 'WinPct
    data4[['TotalBPFaced.2', 'WinPct.2']].rename(columns={'TotalBPFaced.2': 'TotalBPFaced', 'WinPct
])

# Calculate the correlation coefficient between the two variables
correlation_coefficient = bp_and_win_pct_df['TotalBPFaced'].corr(bp_and_win_pct_df['WinPct'])
```

Fig. 22. Code Snippet

H. *How does the number of double faults change over the course of a Aus Open Women?*

Ans) From the analysis, we can conclude that players tend to commit fewer double faults as the tournament progresses. This could be due to various factors such as improved focus, better adaptation to the court conditions, and increased experience playing in the tournament. However, we also observe a slight increase in the number of double faults towards the end of the tournament, which could be due to the pressure of playing in important matches like the quarterfinals and finals. Overall, this suggests that players need to maintain their focus and composure throughout the tournament to minimize the number of double faults and increase their chances of winning.

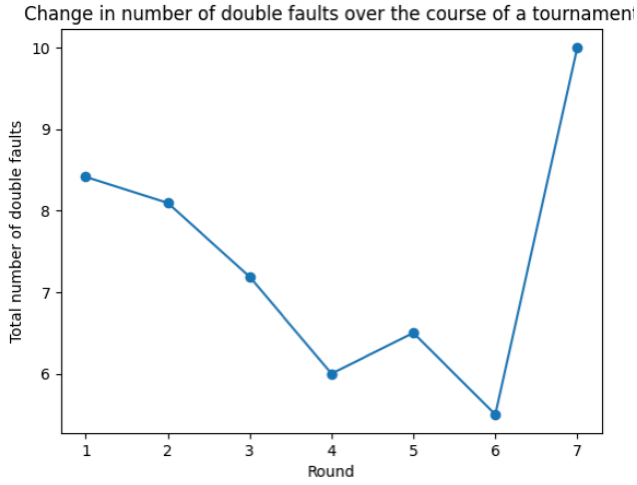

Fig. 23. Total number of double faults v/s Round.

```
round_df = data2.groupby('Round')['DBF.1', 'DBF.2'].mean().reset_index()

# Create a new column to represent the total number of double faults per round
round_df['Total DBF'] = round_df['DBF.1'] + round_df['DBF.2']
# Plot the total number of double faults per round
plt.plot(round_df['Round'], round_df['Total DBF'], marker='o')
plt.xlabel('Round')
plt.ylabel('Total number of double faults')
plt.title('Change in number of double faults over the course of a tournament')
plt.show()
```

Fig. 24. Code Snippet

I. What are the distinct player clusters that can be identified based on their serve performance statistics (such as first serve percentage, first serve points won percentage, second serve percentage, second serve points won percentage, aces, double faults) in the Wimbledon tournament in 2013?

Ans) The PCA plot shows three distinct clusters of players based on their serve performance statistics in the Wimbledon tournament in 2013. Cluster 1 consists of players who have a higher first serve percentage and a higher first serve points won percentage, which indicates that they are strong servers. Cluster 2 consists of players who have a lower first serve percentage and a lower first serve points won percentage, but a higher second serve points won percentage and a lower double fault percentage, indicating that they are relatively weaker servers but have better second serves. Cluster 3 consists of players who have a higher double fault percentage and a lower second serve points won percentage, indicating that they struggle with their second serve and have a tendency to commit double faults.Overall, the PCA analysis allows us to identify distinct patterns in the serve performance statistics of players, which could be useful in evaluating player strengths and weaknesses and developing strategies for playing against them. The three clusters identified here suggest that players' serve performance is a critical factor in their success at Wimbledon and that different strategies may be necessary to compete against players in each cluster. For example, players in Cluster 1 may require a more aggressive return game, while players in Cluster 2 may be vulnerable to pressure on their second serve.

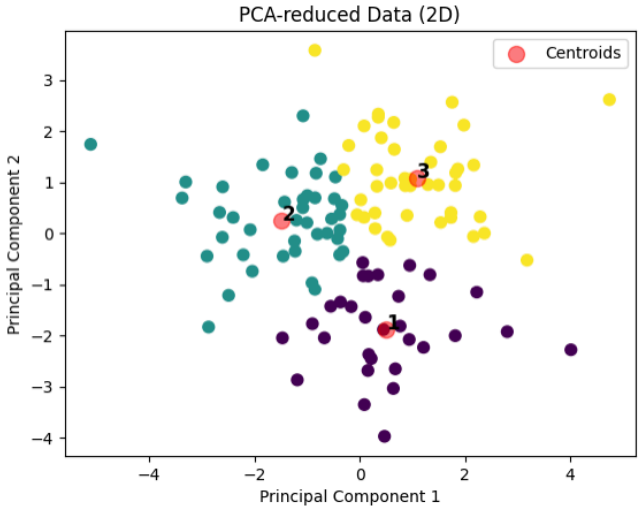

Fig. 25. KMeans Clustering of data based on their serve performance(plotting 2 principal components)

```
from sklearn.cluster import KMeans
serve_stats = ['FSP.1', 'FSW.1', 'SSP.1', 'SSW.1', 'ACE.1', 'DBF.1']
# Create a new dataframe with only the serve performance statistics
x = data7[serve_stats]
x=x.dropna()
mean = np.mean(x, axis=0)
# Center the dataset by subtracting the mean from each data point
X_centered = (x - mean)/(np.std(x,axis=0))
# Compute the covariance matrix of the centered dataset
covariance_matrix = np.cov(X_centered, rowvar=False)
# Compute the eigenvectors and eigenvalues of the covariance matrix
eigenvalues, eigenvectors = np.linalg.eig(covariance_matrix)
# Sort the eigenvalues and eigenvectors in descending order
sorted_indices = np.argsort(eigenvalues)[::-1]
sorted_eigenvalues = eigenvalues[sorted_indices]
sorted_eigenvectors = eigenvectors[:, sorted_indices]
# Choose the number of principal components
n_components = 2
# Select the first n principal components
principal_components = sorted_eigenvectors[:, 0:n_components]
# Transform the centered dataset into the new space
X_pca = np.dot(X_centered, principal_components)
```

Fig. 26. Code Snippet

J. What is the probability that the player having more number of unforced errors goes on to win in Aus Open Men?

Ans)This probability has been calculated using a binomial distribution, where the probability of a player winning when they have more unforced errors is the same for all matches in the Aus Open Men tournament. The binomial distribution assumes that there are only two possible outcomes for each match (win or loss), and that the probability of a player winning is fixed and known.To visualize the probability distribution, we can plot the cumulative distribution function (CDF) and probability density function (PDF) of the binomial distribution. The CDF gives the probability that the number of matches won by a player having more unforced errors is less than or equal to a certain number, while the PDF gives the probability density (or frequency) of each possible number of matches won.From the plot, we can see that the probability of a player having more unforced errors winning a certain number of matches decreases as the number of matches won increases. The most likely outcome is for the player with fewer unforced errors to win more matches, but there is still a non-zero probability that the player with more unforced errors wins more matches.Overall, the probability of the player with more unforced errors winning in the Aus Open Men tournament is relatively low, with only a 14.69% chance of this occurring.

```
The probability of a player winning despite having more unforced errors is 14.29
```

Fig. 27. Output Snippet

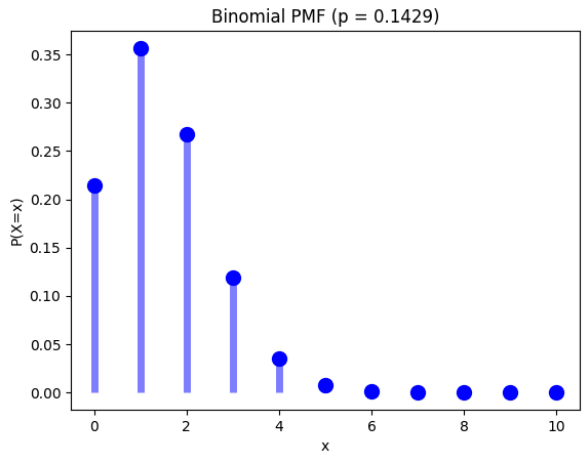

Fig. 28. Binomial PMF for probability that the player having more number of unforced errors goes on to win

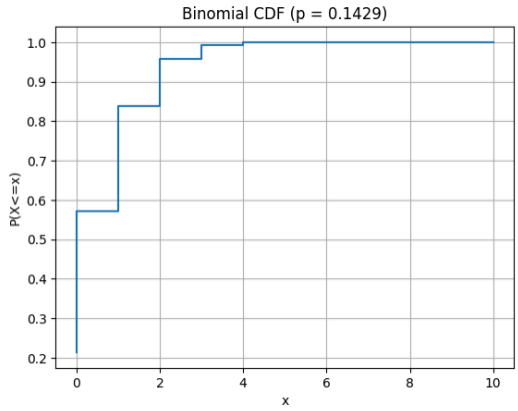

Fig. 29. .Binomial CDF for probability that the player having more number of unforced errors goes on to win

```python
import pandas as pd
unforced_error_winners = data1[(data1['UFE.1'] > data1['UFE.2']) & (data1['Result'] == 1)]
# Calculate the percentage of matches won by the player with more unforced errors
win_percentage = len(unforced_error_winners) / len(data1) * 100
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt
print(f"The probability of a player winning despite having more unforced errors is {win_percentage:.2f}%")
# PMF of Binomial
p = win_percentage/100
n = 10
rv = stats.binom(n,p)
x = np.arange(11)
f = rv.pmf(x)
plt.plot(x, f, 'bo', ms=10,label='binomial pmf')
```

Fig. 30. Code Snippet

K. Can we group the players in the Wimbledon tournament in 2013 into clusters based on their overall performance statistics, such as the total points won, number of net points won, number of breakpoints created, number of break points won, and the number of unforced errors committed, using KMeans clustering?

Ans) The given code performs KMeans clustering on the players' performance statistics in the Wimbledon tournament in 2013, using the five variables - total points won, number of net points won, number of breakpoints created, number of breakpoints won, and the number of unforced errors committed, after performing principal component analysis (PCA) on the data to reduce the dimensionality to 2. The code creates 3 clusters and plots the data points in the reduced 2D space, where the color of each point corresponds to its assigned cluster label, and the red points represent the centroids of the clusters.By using KMeans clustering on the performance statistics of the players in the Wimbledon tournament in 2013, we can group them into three clusters based on their overall performance. The PCA is used to reduce the high-dimensional data into two dimensions to facilitate the visualization of the clusters. The scatter plot of the data points in the reduced 2D space shows that the clusters are well separated, and the centroids of the clusters are clearly visible. Therefore, we can conclude that the KMeans clustering is effective in grouping the players based on their overall performance statistics in the Wimbledon tournament in 2013.

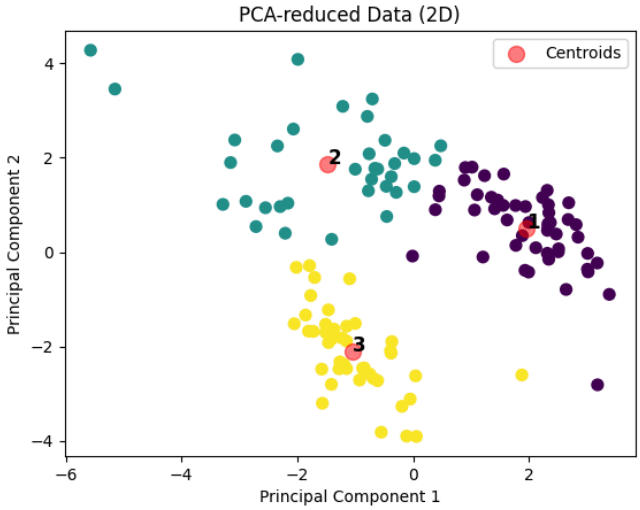

Fig. 31. KMeans Clustering of data based on their overall performance(plotting 2 principal components)

```python
# Extract relevant columns for clustering
cols = ['FNL.1', 'NPA.1', 'BPW.1', 'BPC.1', 'UFE.1',
        'FNL.2', 'NPA.2', 'BPW.2', 'BPC.2', 'UFE.2']
x = data4[cols]
x=x.dropna()
mean = np.mean(x, axis=0)
# Center the dataset by subtracting the mean from each data point
X_centered = (x - mean)/(np.std(x,axis=0))
# Compute the covariance matrix of the centered dataset
covariance_matrix = np.cov(X_centered, rowvar=False)
# Compute the eigenvectors and eigenvalues of the covariance matrix
eigenvalues, eigenvectors = np.linalg.eig(covariance_matrix)
# Sort the eigenvalues and eigenvectors in descending order
sorted_indices = np.argsort(eigenvalues)[::-1]
sorted_eigenvalues = eigenvalues[sorted_indices]
sorted_eigenvectors = eigenvectors[:, sorted_indices]
# Choose the number of principal components
n_components = 2
```

Fig. 32. Code Snippet

VI. SUMMARY OF THE OBSERVATIONS

1) The correlation matrix between 26 performance statistics in the Wimbledon 2013 tournament reveals interesting insights about the playing style of the three players: Djokovic, Federer, and Nadal. All three players have a strong positive correlation between their serve points won and total points won. Djokovic and Federer have a higher correlation between their break points converted and their total points won compared to Nadal. Nadal has a high correlation between his total points won and his return points won. Finally, all three players have a negative correlation between their unforced errors and their total points won,

suggesting that making fewer mistakes is crucial to their success.

2) The graphs comparing the number of aces and unforced errors in men and women's tennis show differences in playing styles between the genders. Men tend to score more aces due to their greater relative power difference, while women commit fewer unforced errors due to their greater focus and accuracy. These differences may be influenced by factors such as physical abilities, technical skills, and strategic choices, but individual players within each gender may exhibit their own unique strengths and weaknesses.

3) Analyzing Andy Murray's performance separately on hard court and grass court, we find that he has won more matches on hard court. However, his serving is slightly better on grass court, while he hits more aces and commits fewer unforced errors on hard court. On hard court, he won 92.8% of matches, while on grass court, he won all seven matches.

4) The analysis shows that female players tend to serve fewer double faults than male players across all tournaments. The US Open women's tournament had significantly fewer double faults compared to other tournaments, indicating that female players in this tournament were more focused. Minimizing double faults is crucial in a match, and comparing the performance of male and female players in terms of double faults could highlight areas of improvement.

5) The analysis of tiebreaks in French Open Men shows that most of the matches were closely contested, with only a few instances of no tiebreaks. One or two tiebreaks were played in a greater chunk of matches, indicating evenly matched players who maintained their performance throughout the sets. The most significant observation is that many matches had three tiebreaks, indicating a highly competitive and intense tournament with well-matched players.

6) The graph and boxplot analysis indicate that players have a higher chance of winning first-serve points on grass courts compared to clay and hard courts. The average percentage of first-serve points won on grass courts is around 41%, while it is around 37% on clay and hard courts. The interquartile range for first-serve points won on grass courts is also higher, indicating more variability in performance on this surface.

7) The data analysis shows a moderate negative correlation between the number of breakpoints faced by a player and their win percentage in the US Open Women tournament. As the number of breakpoints faced increases, the win percentage decreases. However, there is a slight increase in win percentage for players who faced around 20-30 breakpoints. This suggests that players who face a higher number of breakpoints may be more skilled in defending and winning crucial points.

8) The analysis reveals that the number of double faults committed by players decreases as the tournament progresses. The decrease could be attributed to factors such as increased experience, better adaptation to court conditions and improved focus. However, towards the end of the tournament, there is a slight increase in double faults which could be due to the pressure of playing in important matches. Thus, players need to maintain their composure to reduce double faults and increase their chances of winning.

9) The PCA analysis of serve performance statistics in the Wimbledon tournament in 2013 identifies three distinct clusters of players based on their strengths and weaknesses in serving. Cluster 1 consists of strong servers, Cluster 2 consists of players with weaker first serves but strong second serves, and Cluster 3 consists of players with weaker second serves and higher double fault percentages. The analysis can help in evaluating player strengths and developing strategies to play against them, and suggests that serve performance is a critical factor in success at Wimbledon.

10) The probability of a player with more unforced errors winning in the Aus Open Men tournament is low, at only 14.69%, according to calculations using a binomial distribution. The distribution assumes that the probability of a player winning is fixed and known, and the plot shows that the probability of a player with more unforced errors winning decreases as the number of matches won increases. The most likely outcome is for the player with fewer unforced errors to win more matches.

11) The code performs KMeans clustering on player performance statistics in the Wimbledon tournament in 2013, reducing the data dimensions to 2 using PCA. The clustering groups players into three clusters based on their overall performance, and the scatter plot shows well-separated clusters with clear centroid points. The KMeans clustering is therefore effective in grouping players based on their performance statistics.

VII. UNANSWERED QUESTIONS IF ANY:

1) Who is the best tennis player in this dataset?

2) What was the average time duration of the matches in the tournament?

The answers to all other possible questions can be derived from the data provided.

VIII. REFERENCES

[1] UC Irvine Machine Learning Repository. [Online]. Available: https://archive-beta.ics.uci.edu/dataset/300/tennis+major+tournament+match+statistics. [Accessed: 21-April-2023].

[2] GeeksforGeeks. “Pandas Tutorial - GeeksforGeeks,” April 21, 2020. https://www.geeksforgeeks.org/pandas-tutorial/.

[3] GeeksforGeeks. “NumPy Tutorial - GeeksforGeeks,” January 4, 2021. https://www.geeksforgeeks.org/numpy-tutorial/.

[4] GeeksforGeeks. “Matplotlib Tutorial - GeeksforGeeks,” February 8, 2021. https://www.geeksforgeeks.org/matplotlib-tutorial/.

[5] Matplotlib documentation — Matplotlib 3.7.0 documentation. “Matplotlib Documentation — Matplotlib 3.7.0 Documentation,” n.d. https://matplotlib.org/stable/index.html.

[6] pandas documentation — pandas 1.5.3 documentation. “Pandas Documentation — Pandas 1.5.3 Documentation,” n.d. https://pandas.pydata.org/docs/.

[7] NumPy Documentation. “NumPy Documentation,” n.d. https://numpy.org/doc/.

[8] Seaborn-Documentation https://stackoverflow.com/questions/72970343/plotting-top-10-values-in-data

IX. ACKNOWLEDGEMENT

Data available in this link https://archive-beta.ics.uci.edu/dataset/300/tennis+major+tournament+match+statistics

The Report on the scientific questions is written by Nihar Shah, with assistance from Internet Sources under the guidance of Prof. Shanmuga Raman.