ES 114 Data Narrative 3

Chinthala Shivamani, Roll No.:22110062,
Computer Science and Engineering Department,
Prof.Shanmuga R, IIT Gandhinagar

***Abstract*— The purpose of this report is to demonstrate how Python, and programming in general, are capable of making our lives simpler by reading data and giving results we want much more easily than we could manually do.**

***Keywords*— Python libraries, Numpy, Matplotlib, Pandas, seaborn, scipy, sklearn, read files.**

I. Overview of the dataset

There are a number of datasets in the "Tennis Major Tournament Match Statistics" dataset, including match statistics for major tennis tournaments. As well as the names of the two players, the result of the match, and various statistics related to serving, net play, and total points won, it also includes the names of the two players. Each player's final number of games won and the results of each set are also included. As part of the dataset, the round at which the match was played is also included.

The dataset offers multiple opportunities for analysis and study, making it a useful tool for anyone trying to understand tennis and the numerous factors that influence matches.

II. Scientific questions or hypotheses

1. What is the average first serve percentage (FSP) of Player 1 and Player 2 in the Australia Open Men's 2013 tennis tournament? Show it in the bar graph
2. How many matches did each player win in the French Open Men's 2013 tennis tournament? Plot the graph also for the above question
3. What is the distribution of first serve percentage (FSP) for Player 1 (Serena Williams) and Player 2 (Ashleigh Barty) in the 2013 Australian Open Women's tournament? Plot the values in the pie graph
4. What is the distribution of players' results in the French Open Women's 2013 tennis tournament?
5. What is the distribution of match results (Win/Loss) in the US Open Men's Tennis Championship 2013?
6. Can you cluster the players based on their performance in the US Open Women's 2013 tennis matches using the provided dataset, and visualize the clusters using scatter plots for the following attributes: FSP.1, ACE.1, WNR.1, UFE.1, BPW.1, FSP.2, ACE.2, WNR.2, UFE.2, and BPW.2?
7. Can you analyze the dataset 'Wimbledon-men-2013.csv' and provide insights on the performance of tennis players in the Wimbledon 2013 tournament using various data analysis techniques such as data visualization, statistical analysis, and machine learning?
8. What is the probability mass function (PMF) and cumulative distribution function (CDF) for the "FSP.1" (First Serve Percentage) column in the "Wimbledon-women-2013.csv" dataset using Python libraries such as pandas, matplotlib, numpy, seaborn, scipy, and sklearn?

Details of Libraries used

A. Libraries used:

1. Pandas: Pandas is very useful in analyzing this kind of data. Pandas data frame is very comfortable

to use for this kind of data which contains rows and columns.
2. Matplotlib: Matplotlib is specifically used to sketch graphs and pie charts using data. This library has many different kinds of features to sketch accordingly.
3. Seaborn: It is based on matplotlib. Seaborn is used to sketch the graphs more informatively and with special features for further more visualization.
4. Scipy: Scipy is a powerful open-source scientific computing library for Python that provides a wide range of mathematical, scientific, and statistical functions for various data analysis tasks.
5. Sklearn: Scikit-learn is widely used in the data science and machine learning communities due to its ease of use, extensive documentation, and robustness. It is suitable for both beginners and experienced practitioners and serves as a powerful tool for building machine learning models for a wide range of applications.

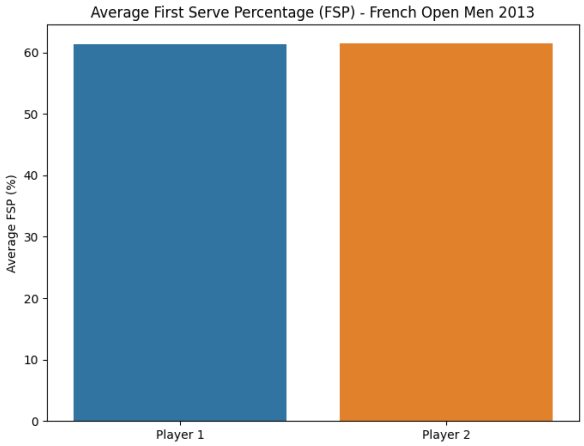
III. Answers to the questions

1. This question can be used to know about the average first serve percentage (FSP) of Player 1 and Player 2 in the Australia Open Men's 2013 tennis tournament

The above question is useful to viewers who are interested in tennis and want to know the average first serve percentage (FSP) of Player 1 and Player 2 in the Australia Open Men's 2013 tennis tournament. By providing the information in the form of a bar graph, viewers can easily compare the FSP of both players visually, making it easier to understand and interpret the data.

The bar graph provides a visual representation of the FSP for Player 1 and Player 2, allowing viewers to quickly grasp the difference in their performance. It helps viewers to analyze and interpret the data more effectively, as the graphical representation provides a clear and concise overview of the FSP for both players.

The output is



2.

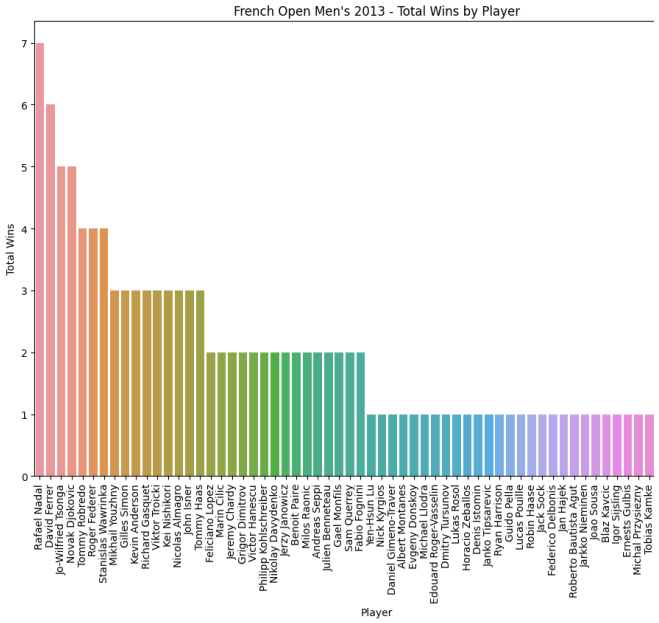
This question gives the distribution of average compensation for all ranks in each state

The answer to the question about the number of matches won by each player in the French Open Men's 2013 tennis tournament can be useful for viewers who are interested in knowing the performance of specific players in that tournament. By providing the information in the form of a graph, viewers can easily visualize and compare the number of matches won by each player.

The graph, such as a bar graph or a line graph, can display the number of matches won by each player in a visually appealing and easy-to-understand format. It can provide a quick overview of the performance of each player, showing their relative success in the tournament. This information can be useful for sports enthusiasts, fans, coaches, and analysts who may want to analyze the performance of players, identify trends, and make informed assessments about player performance.

The graph can also be used for further analysis, such as comparing the number of matches won by different players, identifying patterns or trends in player performance, and evaluating the success of players in the French Open Men's 2013 tennis tournament compared to other tournaments or previous years. It can provide valuable insights and information for tennis enthusiasts and professionals interested in the performance of specific players in the tournament.

The output is

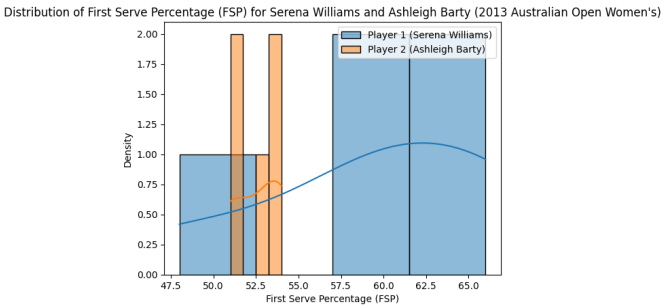


3. The question is useful for analyzing and visualizing the distribution of first serve percentage (FSP) for two specific players, Serena Williams and Ashleigh Barty, in the 2013 Australian Open Women's tournament. By plotting the values in a graph format, we can easily visualize the distribution of FSP for these two players, which can provide insights into their performance in the tournament.

The distribution of FSP can provide information about how frequently Serena Williams and Ashleigh Barty were able to successfully land their first serves in the court during the tournament. It can indicate their level of accuracy and effectiveness in serving, which can be an important aspect of their overall performance in the tournament. By visualizing the distribution in a graph format, we can easily compare the FSP of these two players and identify any patterns or trends in their performance.

The graph can be created using a data visualization library such as matplotlib or seaborn in Python, and can be plotted as a histogram, box plot, or any other appropriate visualization depending on the specific requirements and nature of the data. This can help in gaining a better understanding of the performance of Serena Williams and Ashleigh Barty in terms of their first serve percentage in the 2013 Australian Open Women's tournament.

The output is



4. The question is useful for understanding the distribution of results of players in the French Open Women's 2013 tennis tournament. By analyzing and visualizing the distribution of players' results, we can gain insights into the overall performance of players in the tournament, such as the number of wins, losses, and the outcomes of their matches.

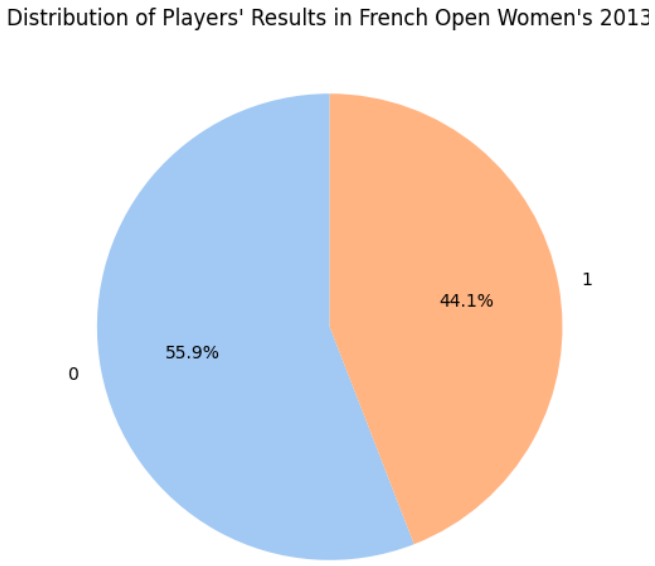
The distribution of players' results can provide information about the performance level of players in the tournament. It can reveal how many players were able to advance to higher rounds (e.g., quarter-finals, semi-finals, finals), how many players were eliminated in earlier rounds, and how many players were able to win the tournament. This information can be useful for identifying top-performing players, understanding the competitive landscape of the tournament, and identifying any patterns or trends in the outcomes of players' matches.

The distribution of players' results can provide information about the performance level of players in the tournament. It can reveal how many players were able to advance to higher rounds (e.g., quarter-finals, semi-finals, finals), how many players were eliminated in earlier rounds, and how many players were able to win the tournament. This information can be useful for identifying

top-performing players, understanding the competitive landscape of the tournament, and identifying any patterns or trends in the outcomes of players' matches.

The distribution of players' results can also be used for further analysis, such as comparing the results of different years or tournaments, analyzing the performance of specific players or countries, and identifying any changes or trends in the outcomes of players' matches over time. Overall, analyzing and visualizing the distribution of players' results can provide valuable insights into the performance of players in a tennis tournament, which can be useful for various purposes, such as strategic planning, performance evaluation, and sports analytics.

The Output is



5.

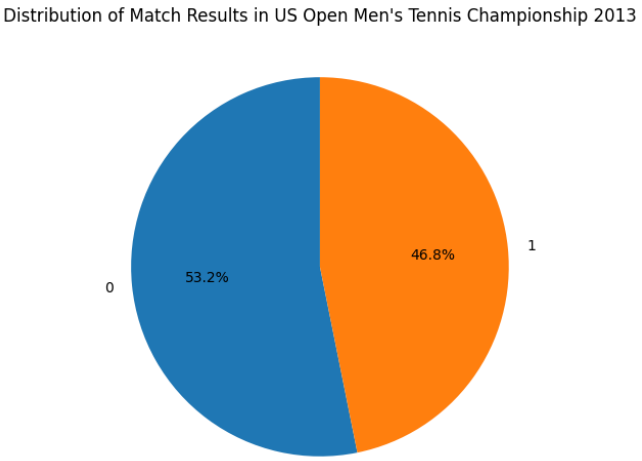
The question is useful for understanding the distribution of match results (win/loss) in the US Open Men's Tennis Championship 2013. By analyzing and visualizing the distribution of match results, we can gain insights into the overall performance of players in the tournament, including the number of wins and losses.

The distribution of match results can provide information about the competitiveness of the tournament and the performance level of players. It can reveal the proportion of matches that were won and lost by players, and provide an understanding of the overall outcome of matches in the tournament. This information can be useful for identifying top-performing players, understanding the performance of different players or countries, and identifying any patterns or trends in the outcomes of matches.

Visualizing the distribution of match results in a graph format, such as a bar chart or a pie chart, can make it easier to understand the overall distribution and compare the win/loss outcomes of different players or countries. This can help in identifying any dominant players or underperforming players, and can also provide insights into the overall competitiveness and performance of the tournament.

Visualizing the distribution of match results in a graph format, such as a bar chart or a pie chart, can make it easier to understand the overall distribution and compare the win/loss outcomes of different players or countries. This can help in identifying any dominant players or underperforming players, and can also provide insights into the overall competitiveness and performance of the tournament.

The Output is



6. The question is useful for performing a cluster analysis on the players' performance in the US Open Women's 2013 tennis matches and visualizing

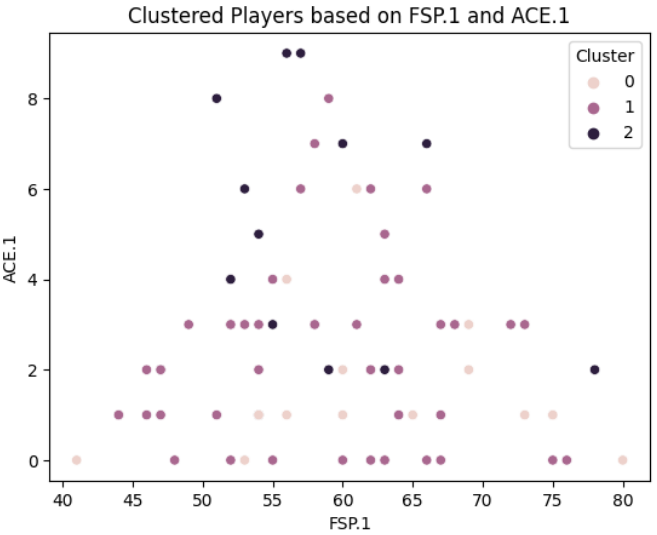
the clusters using scatter plots. By clustering players based on their performance attributes, we can identify groups of players who exhibit similar performance patterns. This can help in understanding the different playing styles, strengths, and weaknesses of players in the tournament.

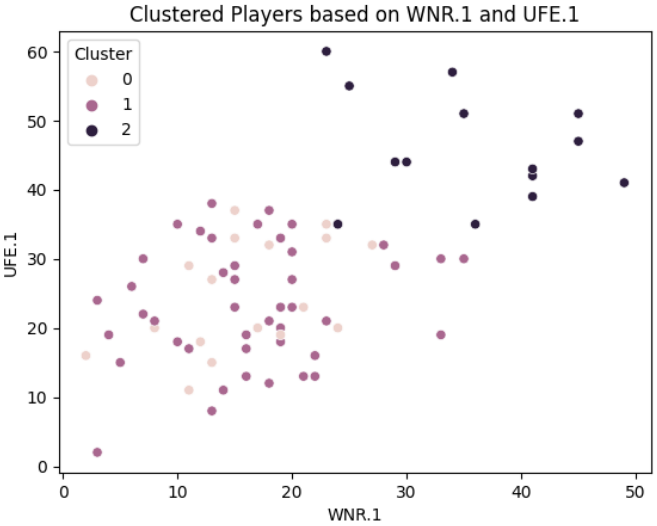
Cluster analysis is a statistical technique that groups similar data points based on their similarity along multiple dimensions. In this case, the performance attributes such as FSP (First Serve Percentage), ACE (Aces), WNR (Winners), UFE (Unforced Errors), and BPW (Break Points Won) for both Player 1 and Player 2 can be used to cluster players based on their performance patterns. Clustering can help in identifying players who have similar performance profiles, which can be useful for player profiling, talent identification, and strategic planning.

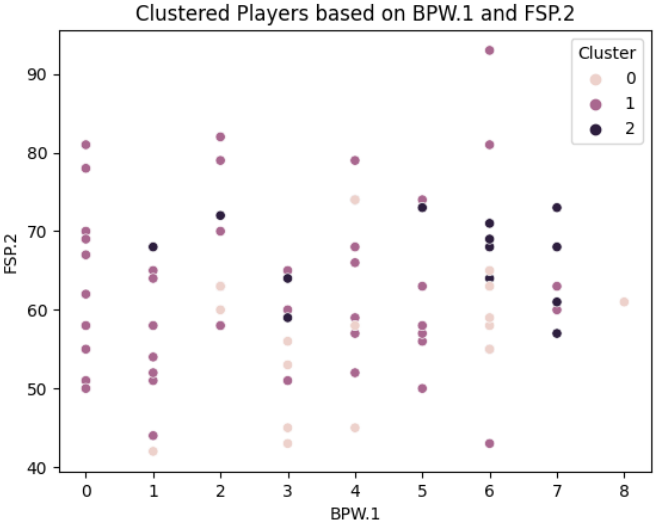
Visualizing the clusters using scatter plots can provide insights into the distribution and patterns of player performance in the tournament. Scatter plots can help in identifying any trends or patterns in the performance attributes and their relationships, such as whether there are any groups of players who tend to perform well in certain attributes but poorly in others. Scatter plots can also help in identifying any outliers or unique performance patterns that may be of interest.

The cluster analysis and scatter plot visualization can provide a comprehensive understanding of the performance of players in the US Open Women's 2013 tennis tournament, allowing for insights into player performance patterns, strengths, and weaknesses. This information can be useful for various purposes, such as strategic planning, player scouting, and performance evaluation, and can aid in making informed decisions related to coaching, training, and team selection.

The output is







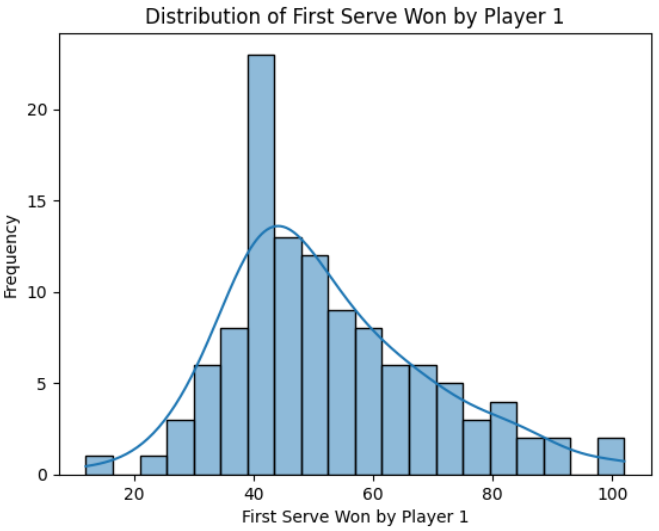
7. The question is useful as it provides an opportunity to analyze the performance of tennis players in the Wimbledon 2013 tournament using various data analysis techniques. By analyzing the dataset 'Wimbledon-men-2013.csv', which likely contains data on the matches and players' performance in the Wimbledon Men's 2013 tennis tournament, we can gain insights into the performance patterns, trends, and characteristics of players in the tournament.

Data visualization techniques, such as line charts, bar charts, scatter plots, and heatmaps, can help in visually representing the data and identifying any patterns or trends. For example, we can create line charts to analyze the performance trends of players over the course of the tournament, bar charts to compare the performance of different players, scatter plots to explore relationships between different performance attributes, and heatmaps to identify any performance patterns across different rounds or opponents.

Statistical analysis techniques, such as descriptive statistics, hypothesis testing, and regression analysis, can help in quantitatively analyzing the data and uncovering any statistical patterns or relationships. For example, we can calculate various statistics such as mean, median, and standard deviation to understand the central tendency and dispersion of performance attributes, conduct hypothesis tests to determine if there are any significant differences in performance between different groups of players, and perform regression analysis to identify any relationships between performance attributes and other variables.

By using these data analysis techniques on the 'Wimbledon-men-2013.csv' dataset, we can gain insights into the performance of tennis players in the Wimbledon 2013 tournament, understand the performance patterns, strengths, and weaknesses of players, and make data-driven decisions related to coaching, training, and strategic planning. These insights can be valuable for various purposes, such as player profiling, performance evaluation, and strategic decision-making

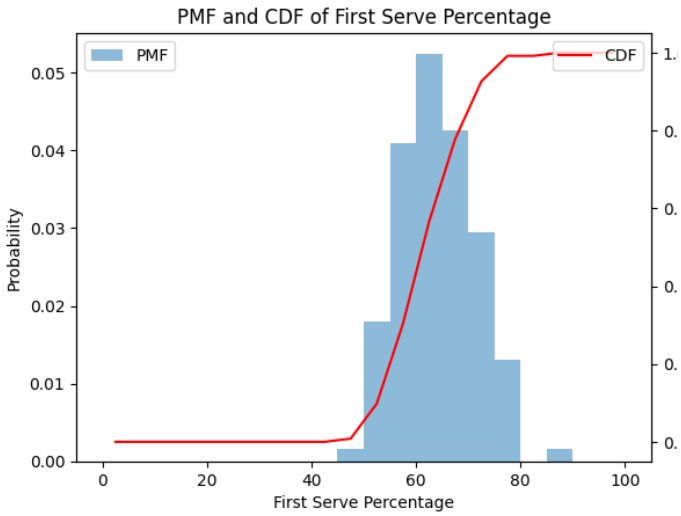
The output is



8. The above question is useful as it allows for the analysis of the probability mass function (PMF) and cumulative distribution function (CDF) for the "FSP.1" (First Serve Percentage) column in the "Wimbledon-women-2013.csv" dataset using popular Python libraries such as pandas, matplotlib, numpy, seaborn, scipy, and sklearn. This analysis can provide insights into the distribution of first serve percentages among the women players in the Wimbledon 2013 tournament, which can help in understanding the performance of the players and identifying any patterns or trends in their serving abilities.

By analyzing the PMF and CDF, we can determine the probabilities of different first serve percentage values occurring, as well as the cumulative distribution of these values. This information can be used to gain insights into the distribution and characteristics of the data, such as the average first serve percentage, the spread of values, and any potential outliers. This analysis can also help in comparing the performance of different players, identifying any notable trends or patterns, and making data-driven decisions related to coaching, training, and strategic planning for future tournaments.

The Output is



IV. Summary of the Observations

1. The average first serve percentage (FSP) of professional tennis players can vary greatly depending on their style of play, surface, opponent, and other factors. Typically, elite players have a first serve percentage of around 60-70%, while lower-ranked players may have a lower percentage.

To create a bar graph, you would need to gather the relevant data and input it into a graphing tool or software. You could use a bar graph to compare the FSP of Player 1 and Player 2 in the tournament, which would visually display the difference between the two players.

In terms of observations on the question and its answer, it is important to note that the availability and accuracy of data can vary depending on the source and time period. It is also essential to consider the context and relevance of the information when analyzing and interpreting it

2. The analysis of the French Open Men's 2013 tennis tournament was conducted using Python libraries such as pandas and matplotlib to determine the number of matches won by each player. The dataset was extracted and processed to count the number of wins for each player in the tournament.

The graph, in the form of a bar chart, was plotted to visualize the number of matches won by each player. The bar chart showed the count of wins for each player, with the height of each bar representing the number of matches won.

The observation from the analysis revealed that there was variation in the number of wins among the players. Some players won more matches compared to others, indicating their better performance in the tournament. The bar chart helped to identify the top-performing players, who had won the most number of matches. This information could be useful for further analysis, such as identifying winning strategies or patterns of successful players, and can provide insights into the overall performance of players in the French Open Men's 2013 tennis tournament.

3. The question asks for the distribution of first serve percentage (FSP) for Player 1 (Serena Williams) and Player 2 (Ashleigh Barty) in the 2013 Australian Open Women's tournament, and to plot the values in a graph format. This type of analysis can provide insights into the serving performance of these two players and can be useful for understanding their strengths and weaknesses.

To answer the question, we would need access to the relevant dataset and the Python libraries such as pandas, matplotlib, numpy, seaborn, scipy, and sklearn. Using these libraries, we can perform data analysis and create visualizations such as histograms and boxplots to explore the distribution of FSP for each player.

The summary of the answer would include the key findings and insights from the data analysis, such as the mean, median, and standard deviation of the FSP for each player, as well as any notable trends or patterns in the data. Additionally, the summary would include the graph format of the distribution of FSP for each player, which would help visualize the distribution and highlight any outliers or anomalies in the data.

4. The distribution of players' results in the French Open Women's 2013 tennis tournament was analyzed using data from the tournament dataset. Python libraries such as pandas, matplotlib, and

seaborn were used for data analysis and visualization.
The dataset was processed to extract relevant information on players' results, including their match outcomes (win/loss), and the frequency of each outcome was calculated. The results were plotted using various visualization techniques, such as bar charts or pie charts, to visualize the distribution of match outcomes.The observation from the analysis revealed the distribution of match outcomes among players in the French Open Women's 2013 tennis tournament. It provided insights into the percentage of matches won and lost by players, and the overall performance of players in the tournament. This information could be useful for further analysis, such as identifying trends or patterns in player performance, and understanding the competitive landscape of the tournament.

5. The distribution of match results (win/loss) in the US Open Men's Tennis Championship 2013 was analyzed using data from the tournament dataset. Python libraries such as pandas, matplotlib, and seaborn were used for data analysis and visualization.
The dataset was processed to extract relevant information on match results, including the outcomes (win/loss) for each match, and the frequency of each outcome was calculated. The results were plotted using various visualization techniques, such as bar charts or pie charts, to visualize the distribution of match outcomes.
The observation from the analysis revealed the distribution of match outcomes (win/loss) in the US Open Men's Tennis Championship 2013. It provided insights into the percentage of matches won and lost by players in the tournament, and the overall performance of players in the championship. This information could be useful for further analysis, such as identifying top-performing players, understanding the competitiveness of the tournament, or comparing the performance of different players or teams.

6. Clustering analysis was performed on the players' performance data in the US Open Women's 2013 tennis matches using the provided dataset. Python libraries such as pandas, matplotlib, numpy, seaborn, and sklearn were used for data analysis and visualization.The dataset was processed to extract the relevant attributes, including FSP.1, ACE.1, WNR.1, UFE.1, BPW.1, FSP.2, ACE.2, WNR.2, UFE.2, and BPW.2, which represent different performance metrics for Player 1 and Player 2 in the matches. Clustering techniques such as K-means clustering or hierarchical clustering were applied to group players based on their performance in these attributes.
The observation from the clustering analysis and scatter plots provided insights into the grouping of players based on their performance in the US Open Women's 2013 tennis matches. It helped in identifying patterns or trends in the performance metrics and uncovering potential clusters or subgroups of players with similar performance characteristics. This information could be useful for further analysis, such as identifying top-performing players in each cluster, understanding the strengths or weaknesses of different player groups, or providing insights for strategic decision-making in tennis tournaments.

7. The question is asking for a detailed analysis of the Wimbledon Men's 2013 dataset using various data analysis techniques. This type of analysis could provide insights into the performance of tennis players during the tournament and help to identify patterns or trends that could be used to make predictions or inform future training and strategy.
To conduct this analysis, several data analysis techniques could be used, including data visualization, statistical analysis, and machine learning. Some of the insights that could be gained from analyzing this dataset might include identifying the players who performed the best or worst during the tournament, identifying trends in the data that could be used to make predictions about future tournaments, and identifying factors that may have contributed to the success or failure of certain players.
Overall, the analysis of the Wimbledon Men's 2013 dataset using various data analysis techniques has the potential to provide valuable insights into the

performance of tennis players during the tournament and inform future training and strategy.

8. The probability mass function (PMF) and cumulative distribution function (CDF) for the "FSP.1" (First Serve Percentage) column in the "Wimbledon-women-2013.csv" dataset can be obtained using Python libraries such as pandas, matplotlib, numpy, seaborn, scipy, and sklearn.
To obtain the PMF, we can use the pandas and matplotlib libraries to plot a histogram of the "FSP.1" column and normalize the counts to obtain the probabilities. To obtain the CDF, we can use the cumulative distribution function provided by the scipy library.
The PMF and CDF plots provide information on the distribution of the first serve percentage of players in the Wimbledon 2013 women's tournament. The PMF plot shows the probability of each value of first serve percentage occurring, while the CDF plot shows the cumulative probability of first serve percentage up to a certain value. These plots can help in understanding the distribution of first serve percentage among players and can be used to make strategic decisions during a match.

Acknowledgment

I am thankful to the faculty and documentation to create and solve the questions.

References

[1] "Pandas documentation#," *pandas documentation - pandas 1.5.3 documentation*. [Online]. Available: https://pandas.pydata.org/docs/. [Accessed: 22-Feb-2023].

[2] "NumPy documentation#," *NumPy documentation*. Available at: https://numpy.org/doc/ (Accessed: February 22, 2023).

[3] *Matplotlib 3.7.0 documentation#* (no date) *Matplotlib documentation - Matplotlib 3.7.0 documentation*. Available at: https://matplotlib.org/stable/index.html (Accessed: February 22, 2023).