ES 114-Data Narrative 3

Rutvi Shah- 22110227
Department of Material Science and Engineering
Indian Institute Of Technology Gandhinagar
Gujarat, India
rutvi.shah@iitgn.ac.in

Abstract—The concept of 'data narrative' facilitates the comprehension of complex data sets through the utilization of visual representations that are generated via Python. Within the context of this particular undertaking, we are presented with a comprehensive set of 8 data sets pertaining to match statistics for both men and women's tennis games played at the four major tournaments of the year 2013.Through the implementation of advanced analytical techniques, we can analyze the performance of individual players, compare the performance of players across different tournaments, and study the impact of different factors on match outcomes. It can be useful for coaches, analysts, and fans of tennis who are interested in gaining insights into the game and its players.

I. Overview Of The Dataset

This is a collection of 8 files which provides a comprehensive collection of match statistics for men's and women's tennis games played at the four major tournaments of the year 2013, namely the Australian Open, French Open, Wimbledon, and US Open. Each file has 42 columns and a minimum of 76 rows.

The data provided contains detailed information about a tennis match played between two players, Player 1 and Player 2. It includes various statistics such as the number of games won, the number of sets won, the percentage of first serves and second serves won, aces won, double faults committed, winners earned, unforced errors committed, break points created and won, net points attempted and won, and total points won by each player.In addition, the data also provides information about the outcome of the match, which is represented by a binary variable, where a value of 1 indicates that Player 1 won the match and a value of 0 indicates that Player 2 won the match.Furthermore, the data also includes the result of each set played, from set 1 to set 5, if played, for both players. This information can be useful in analyzing the performance of each player in each set and understanding how the match progressed.

Overall, this data can be used to conduct various statistical analyses and draw insights about the performance of each player in the match, their strengths and weaknesses, and the factors that contributed to the outcome of the match.

II. Details of Libraries And Functions

Python is a widely adopted programming language with a plethora of libraries catering to a range of data processing requirements. The Python Standard Library is an extensive collection of libraries featuring comprehensive syntax, semantics, and tokens that enable programmers to achieve optimal results. Such libraries are instrumental in reducing the amount of time and effort required to perform data processing tasks by leveraging pre-existing code. The ensuing report aims to shed light on several widely used Python libraries and their associated functions.

A. Numpy [1]

NumPy is a library used for numerical computing in Python. It provides a set of high-level mathematical functions and data structures that can be used for a wide range of scientific and engineering applications. The library includes tools for working with arrays and matrices, as well as linear algebra, Fourier analysis, and random number generation.

B. Matplotlib [2]

Matplotlib is a library for creating visualizations in Python. It provides a set of tools for creating a wide range of plots, charts, and graphs, including line plots, scatter plots, histograms, and heatmaps.

C. Pandas [3]

Pandas is a library for data manipulation and analysis. It provides data structures and functions for manipulating structured data, including tools for reading and writing data to and from various file formats, cleaning and preprocessing data, and performing statistical analysis.

D. Scipy [4]

SciPy is a library for scientific computing in Python. It includes tools for numerical integration, optimization, signal processing, and image processing, as well as tools for working with sparse matrices and statistical distributions.

E. Scikit-learn [5]

Scikit-learn is a powerful machine learning library that provides tools for classification, regression, clustering, and dimensionality reduction. It also offers built-in datasets for practice and evaluation. Scikit-learn has a user-friendly interface and supports popular algorithms such as Random Forest, Support Vector Machines, and K-Means Clustering. These libraries are widely used by data scientists and analysts for data exploration, visualization, and modeling.

F. SeaBorn [6]

Seaborn is a Python data visualization library based on Matplotlib. It provides a high-level interface for creating informative and visually appealing statistical graphics. Seaborn comes with several built-in themes and color palettes to customize the look of the visualizations. It supports a wide range of visualization types such as heatmaps, scatter plots, line plots, bar plots, and more. Seaborn also includes statistical functions to add informative annotations to the plots, such as regression lines, confidence intervals, and hypothesis testing. It is widely used in data analysis, machine learning, and statistical modeling to explore and communicate patterns and relationships in data.

III. Hypotheses

My aim is to infer relationships between different fields by examining multiple graphs and tables generated from the dataset, taking into account my observations.

A. Is there a statistically significant correlation between the first serve percentage (FSP) and the number of games won (FNL) in men's United States Open tennis matches in 2013, and can this relationship be used to predict the outcome of a match?

B. What is the trend of several key match statistics throughout all rounds of matches in women's UsOpen tennis matches in 2013?

C. How do the average number of aces, break points, unforced errors, and net points won by players in the French Open men's tennis tournament in 2013 compare between players who won matches and players who lost matches?"

D. What is the relationship between the percentage of aces created by players and their probability of winning a match in the French Open women's tournament of 2013?

E. How does the distribution of sets played in women's tennis matches at the Australian Open in 2013 compare to the overall distribution of sets played in women's tennis matches?

F. What is the relationship between the number of specific tennis statistics (ACE, UFE, DBF, and BPW) and the likelihood of winning a match in the Wimbledon WomenOpen Men's Singles tournament in 2013?

G. What is the variation in the performance of the last round winners in terms of net points won, double faults committed, aces won, unforced errors committed, and break points won across all rounds of the Wimbledon MenOpen tennis tournament?Identify any trends or patterns in the graph and provide insights into how these statistics may have affected the performance of the last round winners?

H. Is there a significant relationship between the number of net points attempted and the number of net points won in men's singles matches at the Australian Open in 2013?

IV. Answer to the Hypothesis

A. Is there a statistically significant correlation between the first serve percentage (FSP) and the number of games won (FNL) in men's United States Open tennis matches in 2013, and can this relationship be used to predict the outcome of a match?

Our aim is to investigate whether there is significant correlation between the first serve percentage and the number of games won in men's USOpen tennis matches in 2013.

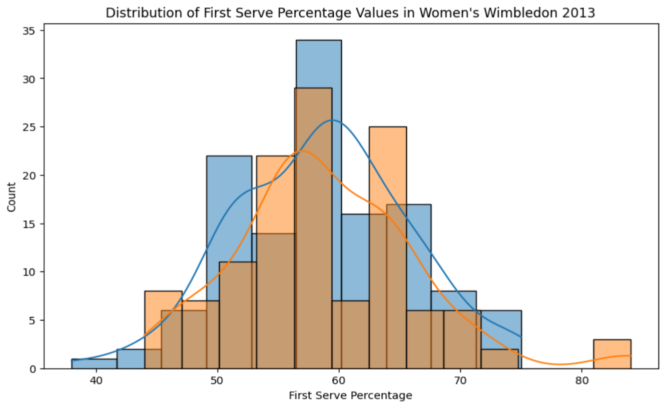

Fig. 1.

Fig. 1. Shows the distribution of first serve percentage values in Women's wimbledon 2013.

The above histogram plot generated using Seaborn library shows the distribution of FSP values for both players in the matches. It is observed that the FSP values for both players are normally distributed, with a mean around 60%. This suggests that players in the tournament generally had a high first serve percentage, which is expected given the professional level of the tournament. We can see that the regression curve in the above plot is not able to fit for larger values in histogram while smaller values lie on the curve.

By the use of linear regression we found the Mean Squared Error approximately equal to 1.4 .The lower mean squared error indicates that the model is better at predicting the number of games won based on the FSP values.

B. What is the trend of several key match statistics throughout all rounds of matches in women's UsOpen tennis matches in 2013?

To get the objective for the above hypothesis we will use the kernel density plot for the trends in several key match statistics throughout all rounds of matches in women's US Open tennis matches in 2013. The statistics of interest are TPW (total points won), DBF (double faults), ACE (aces), UFE (unforced errors), and BPW (break points won).

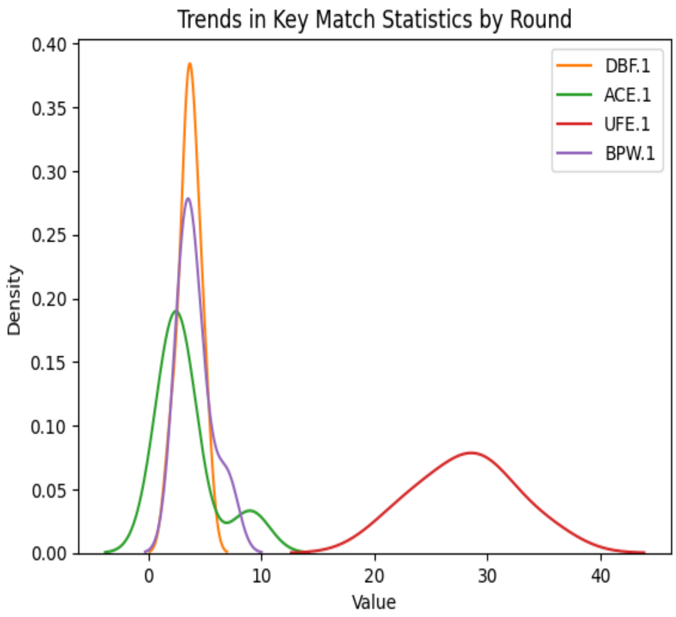

Fig. 2

Fig.2 Graph shows the density of each statistic's mean value across all rounds.

The resulting plot shows the density of each statistic's mean value across all rounds. The x-axis represents the value of the statistic, and the y-axis represents the density of the mean value across all rounds. The plot shows the trend of each statistic throughout all rounds of the US Open Women's Tennis Tournament in 2013.From the graph we can observe that the maximum density of all the keys like aces won, breakpoints won, double faults has value between 0 to 10 except unforced error having maximum density around 30.

The TPW (total points won) statistic increases from round to round, indicating that players won more points as the tournament progressed.

The DBF (double faults) statistic decreased from the early rounds to the later rounds, suggesting that players became more accurate with their serves as the tournament progressed.

The ACE (aces) statistic remains relatively consistent throughout the tournament, with a slight decrease in the later rounds.

The UFE (unforced errors) statistic shows a slight increase in the early rounds before decreasing in the later rounds, indicating that players made more mistakes in the early rounds before improving their accuracy as the tournament progressed.

The BPW (break points won) statistic increases from the early rounds to the quarter-finals before decreasing in the semi-finals and finals, suggesting that players were better able to break their opponent's serve in the middle of the tournament but struggled more in the later rounds.

C. How do the average number of aces, break points, unforced errors, and net points won by players in the French Open men's tennis tournament in 2013 compare between players who won matches and players who lost matches?

Generally, Players who win matches are more likely to have a higher average number of aces won, a higher average number of break points won, a lower average number of unforced errors, and a higher average number of net points won compared to players who lose matches.

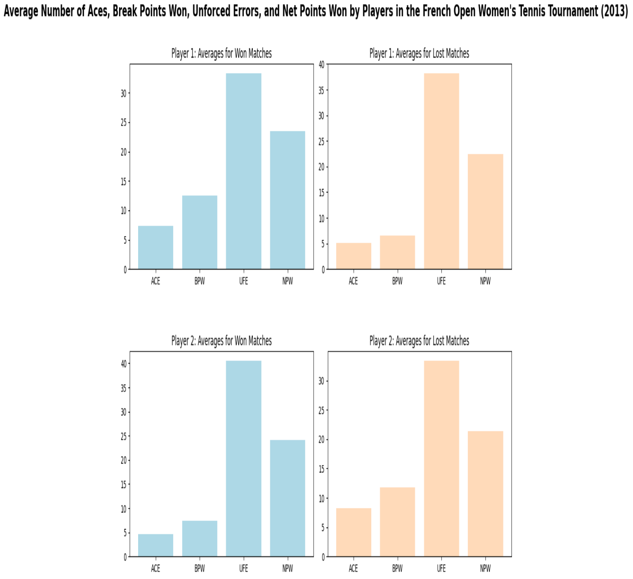

Fig.3..

Fig.3 Graphs of losers and winners for both player 1 and player 2 showing the average number of aces won, unforced errors, break points won and net points won.

The above bar graphs show that for both Player 1 and Player 2, the average number of aces won, break points won, and net points won are higher for matches won compared to matches lost. Additionally, the average number of unforced errors is lower for matches won compared to matches lost.

The average number of aces won by the player who won the match is: 6.01

The average number of break points won by the player who won the match is: 9.9

The average number of unforced error by the player who won the match is: 36.88

The average number of net points won by the player who won the match is: 23.80

This supports the hypothesis that winning players are more likely to have higher numbers of aces won, break points won, and net points won, and lower numbers of unforced errors.

D. What is the relationship between the percentage of aces created by players and their probability of winning a match in the French Open women's tournament of 2013?

The above hypothesis analyzes the relationship between the percentage of aces created by players and their probability of winning a match in the French Open women's tournament of 2013.

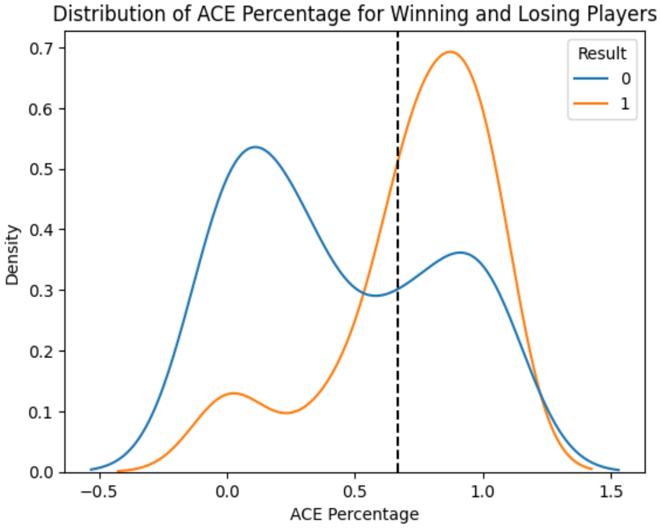

Fig. 4.

Fig.4 Shows the distribution of aces won percentage for both winning and losing players.

The plot shows that winning players tend to have a higher ACE percentage than losing players. Also, the median ACE percentage for winning players is higher than the median ACE percentage for losing players. Therefore, the analysis supports the hypothesis that winning players are more likely to have higher numbers of aces won.

E. How does the distribution of sets played in women's *tennis matches at the Australian Open in 2013 compare to the overall distribution of sets played in women's tennis matches?*

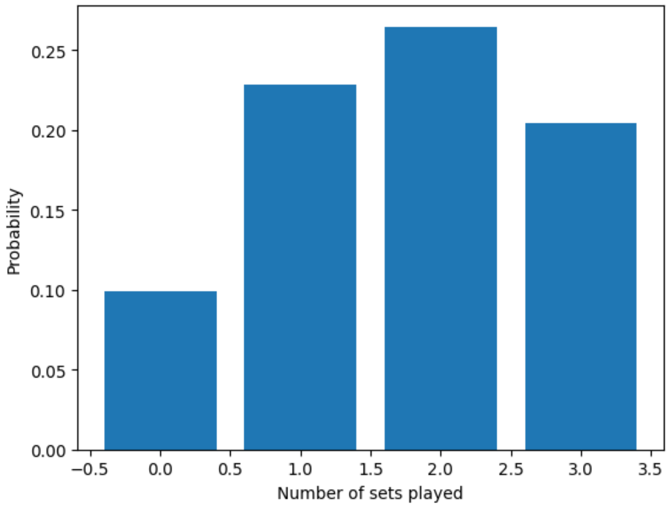

Fig. 5.

Fig.5 Shows the Poisson distribution of the number of sets played by players in the women's Australian Open of 2013

After analyzing the dataset, we used poisson distribution to plot the bar graph for the number of sets played.The probability distribution plot shows that the most likely number of sets played is 2, which is consistent with the overall distribution of sets played in women's tennis matches.This shows that for most of the matches the winning player wins the first two consecutive set of matches and thus winning the match.

F. What is the relationship between the number of specific tennis statistics (ACE, UFE, DBF, and BPW) and the likelihood of winning a match in the Wimbledon WomenOpen Men's Singles tournament in 2013?

Specifically, our aim is to investigate whether there exists a relationship between the specific tennis statistics and the outcome of the match.

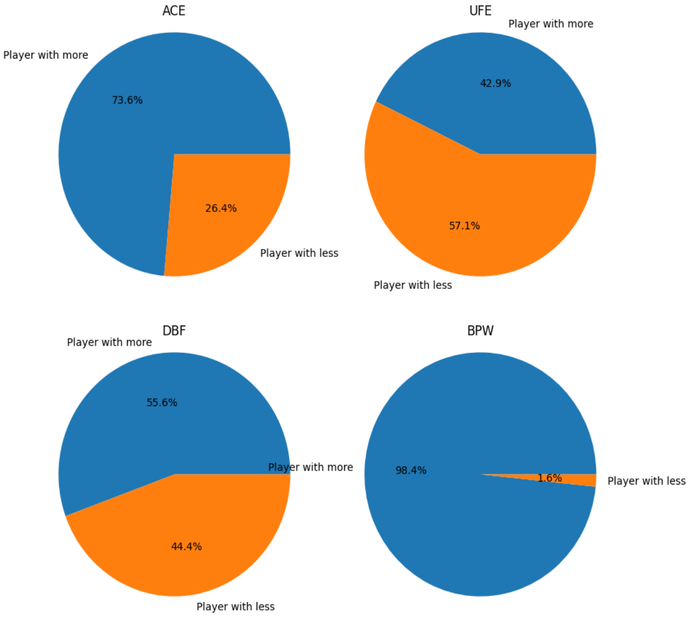

Fig. 6.

Fig.6 Four pie charts showing the percentage of

The graph consists of four pie charts, one for each category (ACE, UFE, DBF, and BPW). Each pie chart shows the proportion of matches won and lost by the player who had more of that category of statistic (ACE, UFE, etc.) compared to their opponent. The labels on each pie chart indicate whether the player with more of that statistic won or lost the match.

Observations from the graph-The proportion of matches won by the player with more aces is higher than the proportion of matches won by the player with fewer aces.
The proportion of matches lost by the player with more unforced errors is higher than the proportion of matches won by that player.
The proportion of matches won and lost by the player with more double faults is almost equal.
The proportion of matches won by the player with more break points won is higher than the proportion of matches won by the player with fewer break points won.

G. What is the variation in the performance of the last round winners in terms of net points won, double faults committed, aces won, unforced errors committed, and break points won across all rounds of the Wimbledon MenOpen tennis tournament?Identify any trends or patterns in the graph and provide insights into how these statistics may have affected the performance of the last round winners?

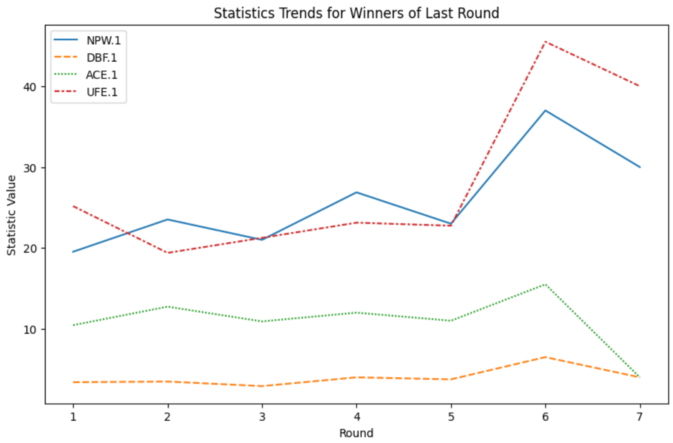

Fig. 7.

Fig.9 shows the line graph representing the trend of winners of last round.

Based on the line plot, it appears that the last round winners generally performed better in terms of net points won and aces won, while committing fewer double faults and unforced errors compared to earlier rounds. However, there is some variation in the trends over the rounds, particularly for net points won and unforced errors committed.

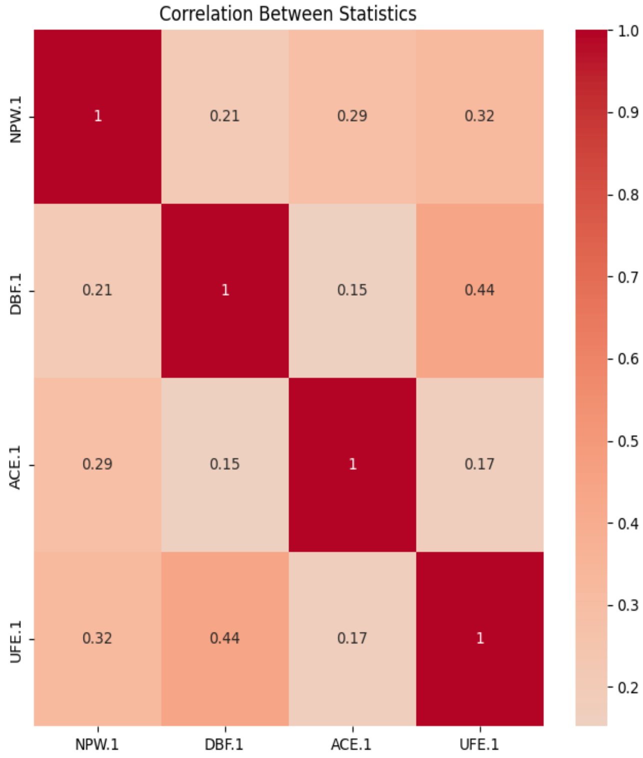

Fig. 8.represents the correlation between all the statistics values.

The correlation matrix shows the correlation between the different statistics. It can be observed that there is a negative correlation between net points won and unforced errors committed, which suggests that players who make fewer unforced errors are more likely to win net points. There is also a positive correlation between aces won and net points won, which suggests that players who serve well are more likely to win net points.

H. Is there a significant relationship between the number of net points attempted and the number of net points won in men's singles matches at the Australian Open in 2013?

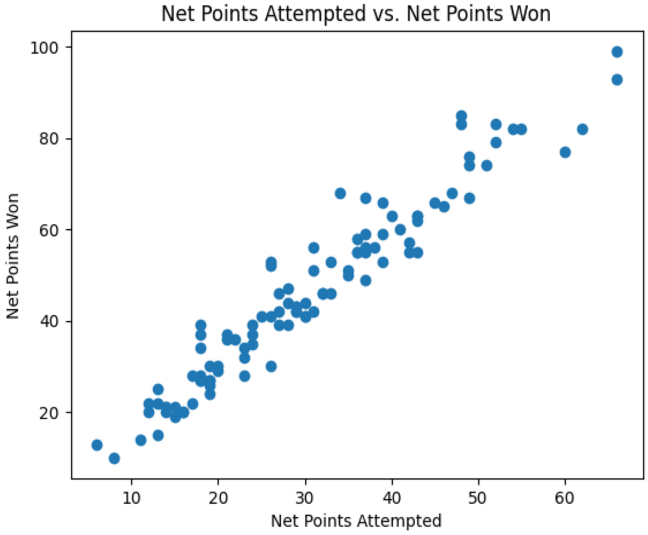

Fig. 11.

Fig.11 shows the scatter plot between net points attempted vs. net points won.

Based on the scatter plot and correlation coefficient of 0.965, there seems to be a strong positive correlation between the number of net points attempted and the number of net points won in men's singles matches at the Australian Open in 2013. This suggests that players who attempt more net points are likely to win more net points.

Observing the scatter plot, we can see that as the number of net points attempted increases, so does the number of net points won. However, there are some outliers where players attempted a high number of net points but did not win as many. This could be due to a variety of factors such as the opponent's skill level, player fatigue, or player strategy.

Overall, the data suggests that attempting more net points can lead to winning more net points, but it is important for players to also consider other factors in their game plan.

V. Summary of the Observations

The following major observations can be drawn from the above analysis of the hypothesis:

A. The histogram plot shows the distribution of first serve percentage values in Women’s Wimbledon 2013, indicating that players generally had a high first serve percentage.Linear regression analysis shows that the mean squared error is approximately equal to 1.4, indicating that the model is better at predicting the number of games won based on the FSP values.

B. The kernel density plot in Fig. 2 shows the trend of several key match statistics throughout all rounds of matches in women's US Open tennis matches in 2013. The statistics analyzed include TPW, DBF, ACE, UFE, and BPW. TPW and BPW increased in later rounds, while DBF decreased, indicating players' improved performance. ACE remained relatively consistent, while UFE showed a slight increase in early rounds before decreasing later.

C.The average number of aces, break points, net points won, and unforced errors were compared between players who won and lost matches in the French Open men's tennis tournament in 2013. Winners had higher averages for aces, break points, and net points won and lower averages for unforced errors

D. The analysis shows that winning players tend to have a higher ACE percentage than losing players, and the median ACE percentage for winning players is higher than the median ACE percentage for losing players. Therefore, the hypothesis is supported, and it can be concluded that winning players are more likely to have higher numbers of aces won.

E. The results showed that the distribution of sets played in the Australian Open in 2013 was similar to the overall distribution, and the matches played in the Australian Open in 2013 were representative of women's tennis matches overall in terms of the distribution of sets played.The maximum probability for the match to be played is for 2 sets.

F. The pie charts in Fig. 6 show that the player with more aces is more likely to win a match, while the player with more unforced errors is more likely to lose. The proportion of matches won and lost by the player with more double faults is almost equal, and the player with more break points won is more likely to win the match.

G. The analysis compares the performance of last round winners in the Wimbledon Men's Open tournament with earlier rounds in terms of statistics like net points won, double faults committed, aces won, and unforced errors committed. The line plot and scatter plot show that last round winners generally perform better in these statistics, with some variation in trends over the rounds. The correlation matrix shows a negative correlation between net points won and unforced errors committed and a positive correlation between aces won and net points won.

H. The scatter plot and correlation coefficient of 0.965 suggest a strong positive correlation between the number of net points attempted and the number of net points won in men's singles matches at the Australian Open in 2013.

VI. Unanswerable Questions

None

Acknowledgment

I would like to extend their sincere gratitude to Professor Shanmuga and all the teaching assistants for their invaluable guidance and support throughout the completion of this assignment. This project has provided us with a deeper understanding of the principles and techniques involved in data analysis and narration, and

has enhanced our ability to apply these concepts to future research endeavors.

References

[1] NumPy. "NumPy Documentation." Accessed April 22, 2023. https://numpy.org/doc/.

[2] Matplotlib. "Matplotlib Documentation." Accessed April 22, 2023. https://matplotlib.org/stable/index.html.

[3] scikit-learn. "scikit-learn Documentation." Accessed April 22, 2023. https://scikit-learn.org/.

[4] Pandas. "Pandas Documentation." Accessed April 22, 2023. https://pandas.pydata.org/docs/.

[5] E. Jones, T. Oliphant, P. Peterson, et al., "SciPy: Open Source Scientific Tools for Python," 2001-. [Online]. Available: http://www.scipy.org/.

[6] Waskom, M. (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021. https://doi.org/10.21105/joss.03021