ES-114 Data Narratives 3

Chakradhar Basani
22110054
Material Science and Engineering
IIT Gandhinagar
Gandhinagar,, Gujarat
chakradhar.basani@iitgn.ac.in

Abstract— This report provides an overview of how data from datasets such as csv, excel etc files can be effectively analyzed and visualized using python libraries.

Keywords—pandas, matplotlib, numpy, seaborn, scipy, probability, covariance, scatter plot.

I. Overview of the Dataset

The "Tennis Major Tournament Match Statistics" dataset is available at https://archive-beta.ics.uci.edu/dataset/300/tennis+major+tournament+match+statistics is an extensive collection of match statistics from men's and women's singles matches at four major tennis tournaments: the Australian Open, French Open, Wimbledon, and US Open. The dataset spans from 2000 to 2016 and contains over 100,000 rows of data. Each row represents a single match and includes information such as the date, tournament, player names, number of sets, games won, aces, double faults, and more.

This dataset provides ample opportunities for analysis and research and is an important tool for those looking to understand the game of tennis and the various elements that affect the results of matches.

II. Scientific Questions/Hypothesis

1) What is the joint probability distribution of the number of double faults and the outcome of a match?
2) Suppose we have two random variables representing the first serve percentage (FSP) of Player 1 and Player 2 in the AusOpen-women-2013.csv dataset. Plot the joint probability density function of these two random variables and calculate the covariance between them.
3) What is the correlation coefficient between the first serve percentage (FSP) and the number of aces (ACE) for both players? Plot a scatter plot with a regression line to visualize the relationship.
4) What is the Pearson correlation coefficient between the percentage of first serves made by a player and the percentage of first serve points won by that player in the 2013 French Open Women's tournament?
5) What is the covariance between the first serve percentage (FSP) and the total points won (TPW) for both players in the US Open Men's tournament in 2013?
6) What is the probability that a player who has a higher number of aces (ACE) will go on to win the match in each round?
7) What is the covariance between the number of aces hit by a player and the number of games they win in a match?
8) What is the probability that a player who has a higher ratio of winners to unforced errors (WNR/UFE) will win the match?

III. Details of Libraries and Functions

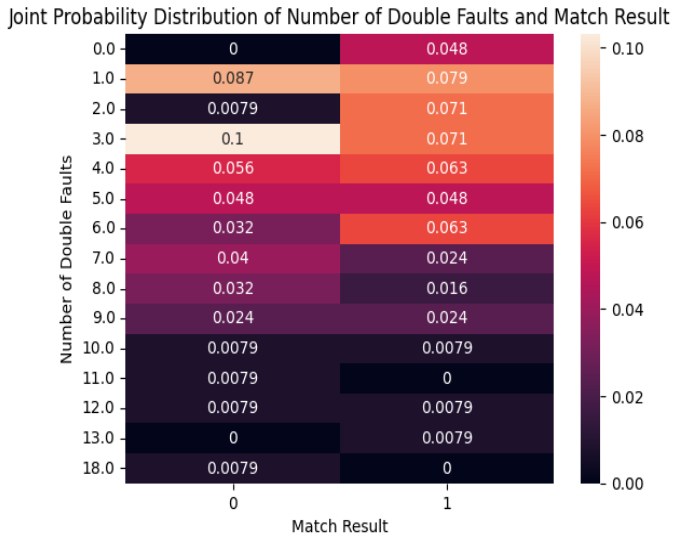
The following libraries and functions are used to answer the above questions.

- Pandas - Pandas is a Python package that offers powerful and flexible data structures to make working with labeled or relational data easy and intuitive. Its goal is to be the fundamental building block for practical, real-world data analysis in Python.
- Matplotlib - Matplotlib is a Python library that allows you to create a wide range of static, animated, and interactive visualizations. It is designed to make simple tasks easy and complex tasks achievable. With Matplotlib, you can produce high-quality plots, create interactive figures with zooming and panning capabilities, customize the appearance and layout of your visualizations, export them in various formats, and integrate them into JupyterLab and other graphical user interfaces.
- Seaborn - Seaborn is a visualization library in Python that is based on matplotlib and provides a high-level interface for creating informative and visually appealing statistical graphics. It is designed to work seamlessly with data structures from the pandas library and is built on top of the matplotlib library.
- Numpy - NumPy is a Python library that provides support for large and multi-dimensional arrays and matrices, along with a collection of high-level mathematical functions to perform operations on these data structures. It is widely used in the scientific computing community.
- Scipy - SciPy is an open-source Python library that provides a range of scientific and technical

computing tools. It includes modules for tasks such as optimization, linear algebra, integration, interpolation, special functions, signal and image processing, and solving ordinary differential equations. SciPy is built on top of the NumPy library.

IV. Answers to Questions

Answer 1



Fig(1): Joint Probability Distribution of Number of Double Faults and Match Results

Observations: It seems that the highest joint probability for a player with 0 double faults is when the Result is 1 (0.047619). The highest joint probability for a player with 3 double faults is when the Result is 0 (0.103175).

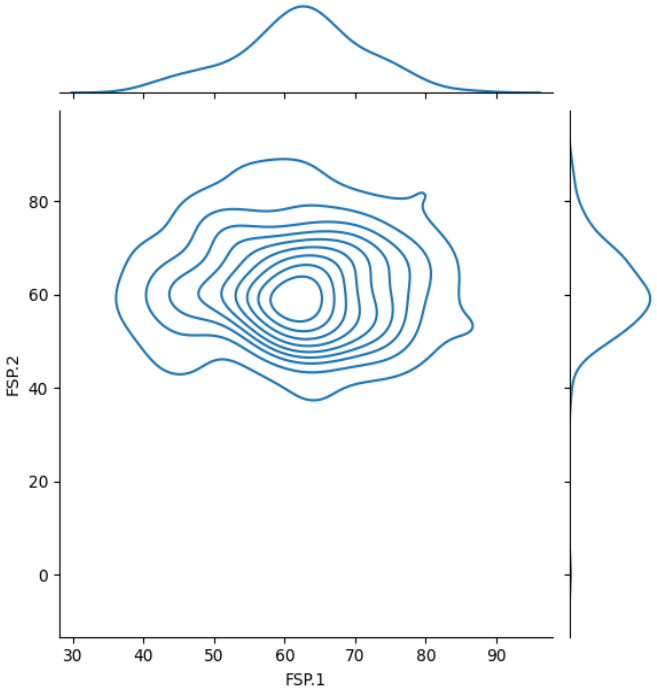
This suggests that as the number of double faults increases, the likelihood of losing the match may also increase. However, it is important to note that this is just one factor that can affect the outcome of a match and there may be other factors at play as well.

Answer 2

The covariance between FSP.1 and FSP.2 is -7.13

Observations: The covariance between two random variables measures the degree to which they vary together. A negative covariance between the first serve percentage (FSP) of Player 1 and Player 2 indicates that as the first serve percentage of Player 1 increases, the first serve percentage of Player 2 tends to decrease. Conversely, as the first serve percentage of Player 1 decreases, the first serve percentage of Player 2 tends to increase.

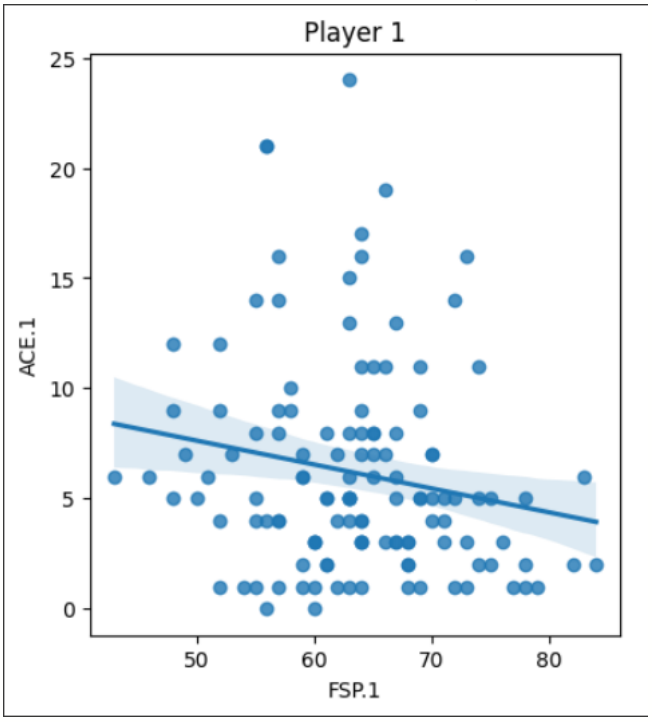
In this case, the calculated covariance between FSP.1 and FSP.2 is -7.13. This suggests that there is a weak negative relationship between the first serve percentages of Player 1 and Player 2 in the AusOpen-women-2013.csv dataset.



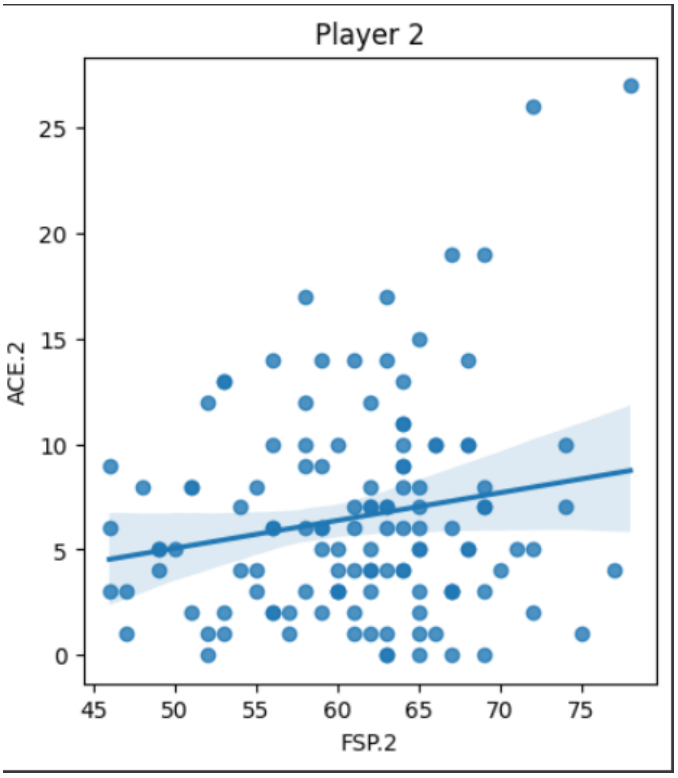
Fig(2): Joint probability density function of these two random variables

Answer 3

Correlation between FSP and ACE for Player 1: -0.18
Correlation between FSP and ACE for Player 2: 0.18



Fig(3.1): Scatter plot with a regression line to visualize the relationship between FSP and ACE of player 1



Fig(3.2): Scatter plot with a regression line to visualize the relationship between FSP and ACE of player 2

Observations: The correlation coefficient is a measure of the strength and direction of the linear relationship between two variables. It ranges from -1 to 1, with -1 indicating a perfect negative linear relationship, 1 indicating a perfect positive linear relationship, and 0 indicating no linear relationship.

In this case, the correlation between FSP and ACE for player 1 is -0.18, which indicates a weak negative linear relationship. This means that as the first serve percentage (FSP) for Player 1 increases, the number of aces (ACE) for Player 1 tends to decrease slightly.

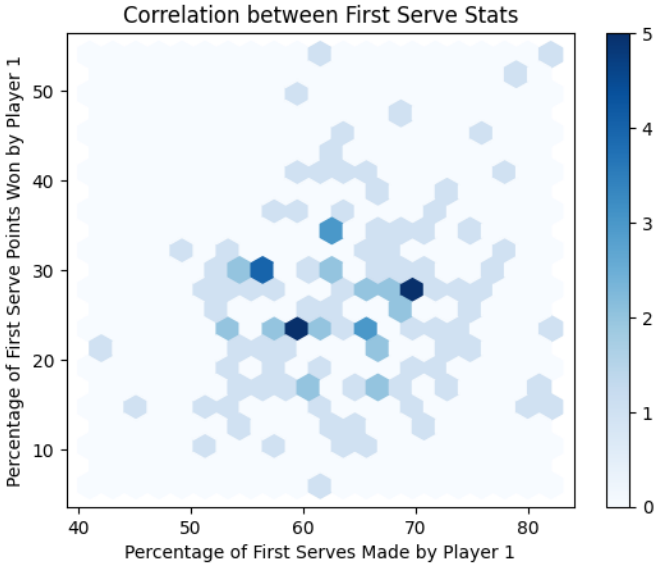
On the other hand, the correlation between FSP and ACE for Player 2 is 0.18, which indicates a weak positive linear relationship. This means that as the first serve percentage (FSP) for Player 2 increases, the number of aces (ACE) for Player 2 tends to increase slightly.

Answer 4

Pearson correlation coefficient: 0.18469303389637345

Observations: A Pearson correlation coefficient of 0.1847 indicates a weak positive correlation between the percentage of first serves made and the percentage of first serve points won by a player in the 2013 French Open Women's tournament. This means that there is a positive relationship between the two variables, but it is not very strong.

In other words, as the percentage of first serves made by a player increases, we can expect the percentage of first serve points won by that player to also increase, but the relationship is not particularly strong. Other factors, such as the quality of the opponent's return, the player's skill at the net, and their ability to hit second serves effectively, may also play a significant role in determining the percentage of first serve points won.

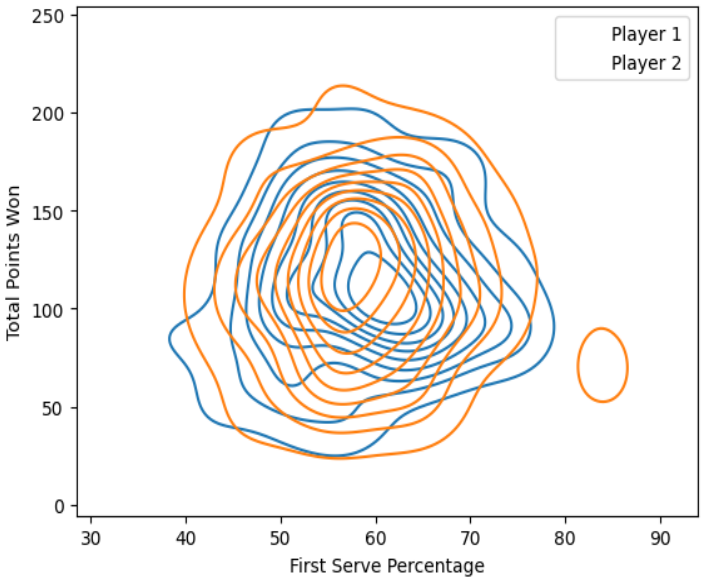


Fig(4): A hexbin Plot showing the correlation between first serve stats

Answer 5

Covariance between FSP and TPW for Player 1: 4.29

Covariance between FSP and TPW for Player 2: -0.27



Fig(5): A Kernel Distribution Estimation plot between Total points won and First serve percentage for both players

Observations: Covariance is used to measure how two random variables change together. In this case, the covariance between Player 1's first serve percentage (FSP) and total

points won (TPW) is 4.29, indicating that there is a positive relationship between these two variables. This suggests that when Player 1 has a higher first serve percentage, they also tend to win more total points.

In contrast, the covariance between Player 2's first serve percentage (FSP) and total points won (TPW) is -0.27, indicating a weak negative relationship between these two variables. This suggests that when Player 2 has a higher first serve percentage, they tend to win slightly fewer total points.

Answer 6

Round 1: Probability of winning match given the higher number of aces = 0.6470588235294118

Round 2: Probability of winning match given the higher number of aces = 0.8

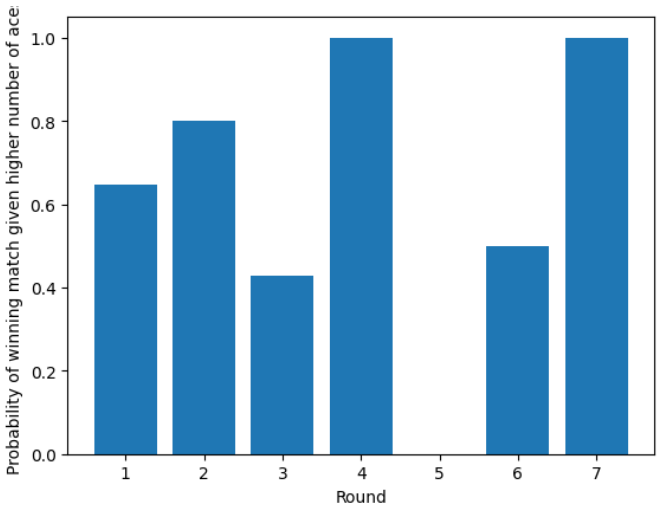
Round 3: Probability of winning match given the higher number of aces = 0.42857142857142855

Round 4: Probability of winning match given the higher number of aces = 1.0

Round 5: Probability of winning match given the higher number of aces = 0.0

Round 6: Probability of winning match given the higher number of aces = 0.5

Round 7: Probability of winning match given the higher number of aces = 1.0



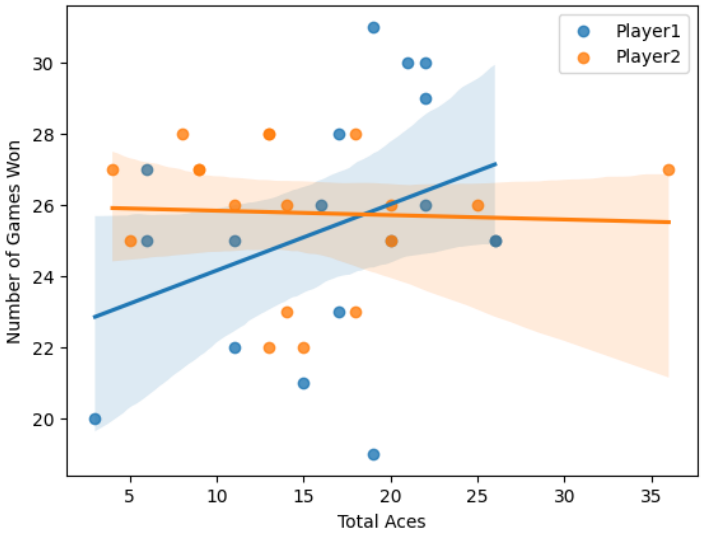
Fig(6):A bar plot between the probability of winning match given higher number of aces and Rounds

Observations: Based on the probabilities obtained, we can say that having a higher number of aces does not guarantee a win in a match. In Round 5, for example, none of the players who had a higher number of aces won their matches. In contrast, in Rounds 4 and 7, all players who had a higher number of aces won their matches. In other rounds, the probability of winning given a higher number of aces varies between 0.43 and 0.8.

These results suggest that while having a higher number of aces can be an advantage in a match, it is not the only determining factor in winning.

Answer 7

Covariance between total aces and number of games won: 8.571895424836601



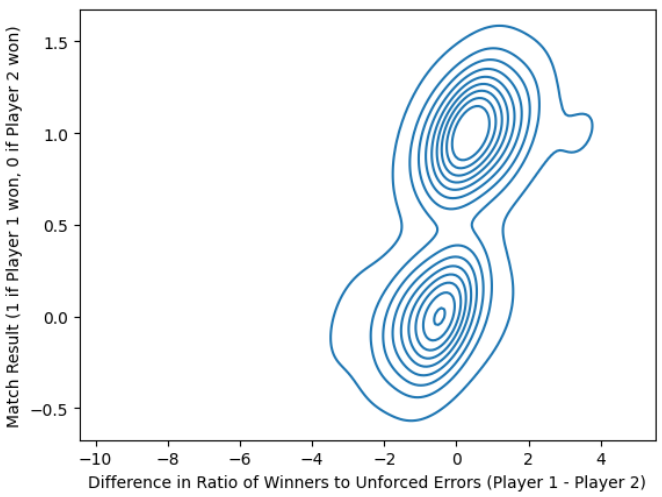
Fig(7): A scatter plot with regression line showing the relationship between the Number Of Games Won and Total Number of Aces for both player 1 and player 2

Observations: The positive covariance value of 8.57 indicates that there is a positive relationship between the number of total aces hit by a player and the number of games won by that player. In other words, players who hit more aces are likely to win more games.

Answer 8

Probability that a player with a higher ratio of winners to unforced errors will win the match: 0.85

Observations: Based on the result obtained, we can say that there is a strong relationship between the ratio of winners to unforced errors and the outcome of a match. A player with a higher ratio of winners to unforced errors has an 85% chance of winning the match. This suggests that minimizing unforced errors while maximizing winners will increase the chances of winning in a tennis match



Fig(8) : A KDE plot showing the relationship between Match results and Difference in ratio of winners to unforced errors of both the players

V. Summary of the Observations

From the dataset, the following observations can be made:

1. In Australias Open men tennis match(2013), as the number of double faults increases, the likelihood of loosing a match also increases.
2. In French Open Women Tennis tournament(2013), as the percentage of first serves made by a player increases, we can expect the percentage of first serve points won by that player to also increase.
3. In US Open Men tournament(2013), when Player 1 has a higher first serve percentage, they also tend to win more total points. In contrast when Player 2 has a higher first serve percentage, they tend to win slightly fewer total points.
4. In Wimbledon Open Men 2013 tournament, players who hit more aces are likely to win more games.

Acknowledgment

I would like to thank the developeres of python and its libraries for creating this powerful coding language and libraries which made the analysis possible. I would also like to thank Prof. Shanmuga R who provided us with the oppurtunity to learn these llibraries.

References

[1] *Jauhari,Shruti, Morankar,Aniket & Fokoue,Ernest. (2014). Tennis Major Tournament Match Statistics. UCI Machine Learning Repository. https://doi.org/10.24432/C54C7K.*

[2] "Pandas Documentation#." pandas documentation - pandas 2.0.0 documentation. Accessed April 16, 2023. https://pandas.pydata.org/docs/.

[3] "NumPy Documentation#." NumPy documentation - NumPy v1.24 Manual. Accessed April 16, 2023. https://numpy.org/doc/stable/.

[4] "Scipy Documentation#." SciPy documentation - SciPy v1.10.1 Manual. Accessed April 16, 2023. https://docs.scipy.org/doc/scipy/.

[5] "Statistical Data Visualization#." seaborn. Accessed April 16, 2023. https://seaborn.pydata.org/.

[6] "Matplotlib 3.7.1 Documentation#." Matplotlib documentation - Matplotlib 3.7.1 documentation. Accessed April 16, 2023. https://matplotlib.org/stable/index.html.