ES 114

Probability, Statistics and Data Visualization

Data Narrative III

HARSH KUMAR KESHRI
Roll Number: 22110094
Mechanical Engineering
IIT Gandhinagar

I. OVERVIEW OF THE DATASET

Data narrative III is based on a dataset on Tennis Major Tournaments which contains 8 Excel files that correspond to the Tennis matches conducted in Australia, French, the US, and Wimbledon both in the men's and women's leagues in 2013.

Therefore, the 8 files we have given are namely:

1. AusOpen-men-2013
2. AusOpen-women-2013
3. FrenchOpen-men-2013
4. FrenchOpen-women-2013
5. USOpen-men-2013
6. USOpen-women-2013
7. Wimbledon-women-2013
8. Wimbledon-men-2013

All these 8 files contain different attributes. Some of them are as follows:

1. Player1: Name of Player 1
2. Player2: Name of Player 2
3. ROUND: Number of Rounds
4. Result: Result of the match
5. FNL.1/FNL.2: Final number of games won by player1/player2
6. FSP.1/FSP.2: First-serve percentage for player 1/player2
7. FSW.1/FSW.2: First serve won by the player 1/player2
8. SSP.1/SSP.2: Second serve percentage for player1/player2
9. SSW.1/SSW.2: Second serve won by player1/player2
10. ACE.1/ACE.2: Aces won by player1/player2

II. SCIENTIFIC QUESTIONS/HYPOTHESES

According to the dataset on Tennis Major Tournaments with 8 files in it, I have made 8 different Scientific/Hypothesis questions which are as follows:

1. How does the first serve percentage of Player 1 correlate with their chances of winning the match?

2. What is the relationship between the number of breakpoints and the outcome of the match?

3. How do the serve statistics (FSW, SSW, DBF, UFE) vary from each other? Plot a correlation matrix for all the same.

4. Does the covariance between the percentage of second serve points won and the number of double faults committed differ significantly between the first and second players?

5. How does the breakdown of points won through forced errors compare between Player 1 and Player 2

6. How does the final result of the match depend on the average number of double faults in a match for player 1 and player 2?

7. Is there a relationship between the number of breakpoints faced and the number of break points saved for winning and losing players in the dataset?

8. Identify clusters of players based on their serving performance, including first serve percentage, first serve points won, second serve percentage and second serve points won.

III. DETAILS OF LIBRARIES AND FUNCTIONS

I have used various Python libraries and a lot of functions in my code to get it to run efficiently. Some of them are given below:

1. Pandas: Pandas: Pandas is a popular library for data manipulation and analysis in Python. It provides data structures like data frames and series and functions for reading, writing, and manipulating data [1].

Some of the functions of the Pandas are as given.

- pd.read_csv(): This function is used to read data from a CSV file and return it as a data frame.
- df.groupby(): This function is used to group data in a data frame by one or more columns.
- df.plot(): This function is used to create various types of plots using the data in a data frame.

2. Matplotlib: Matplotlib is a plotting library for Python. It provides a wide range of functions for creating different types of plots, along with extensive customization options [2].

Some of the functions used in the Matplotlib are:

- plt.plot(): This function is used to create a line plot using data.
- plt.bar(): This function is used to create a bar plot using data.
- plt.pie(): This function is used to create a pie chart using data.

3. Sklearn: Scikit-learn, commonly abbreviated as sklearn, is a popular machine learning library in Python that provides various tools and algorithms for data analysis, data pre-processing, feature extraction, feature selection, dimensionality reduction, model selection, and model evaluation [3].

Here are some of the major functions provided by sklearn which I have used:

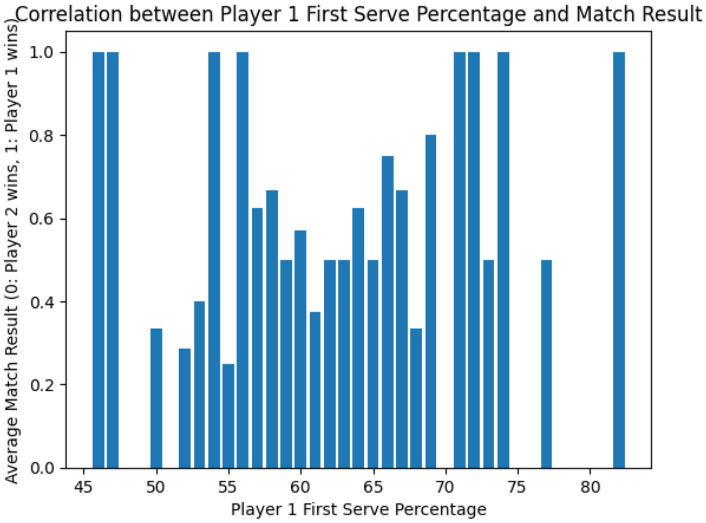
- Linear_model: The linear_model module of the scikit-learn library provides a range of linear models for various regression and classification tasks.
- LinearRegression: The linearRegression class is used for linear regression tasks, where the goal is to predict a continuous target variable based on one or more predictor variables. It fits a linear model to the data using the least squares method and can handle both simple and multiple linear regression.

4. Seaborn: Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures [4].

III. Answers to the Questions (with Appropriate Illustrations)

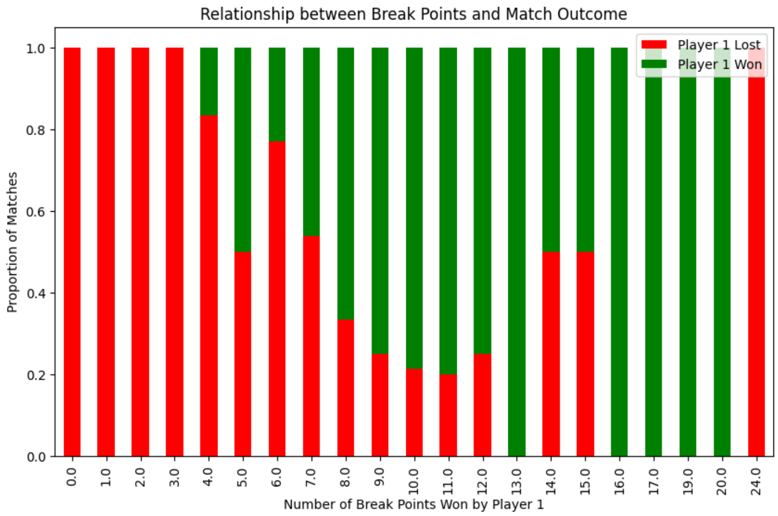
1. How does the first serve percentage of Player 1 correlate with their chances of winning the match?

- To determine the correlation between the first serve percentage of Player 1 with their chances of winning the match, we need to first group the data by first serve percentage and then calculate the mean result for each group. For example, we can group all matches where Player 1 had a first serve percentage between 60-70% and calculate the average match result for those matches.
- To illustrate this, we can make a bar graph with the x-axis and y-axis as the First serve percentage of player 1 and the Average match result respectively where the height of each bar represents the average match result for each group of matches based on the first-serve percentage.



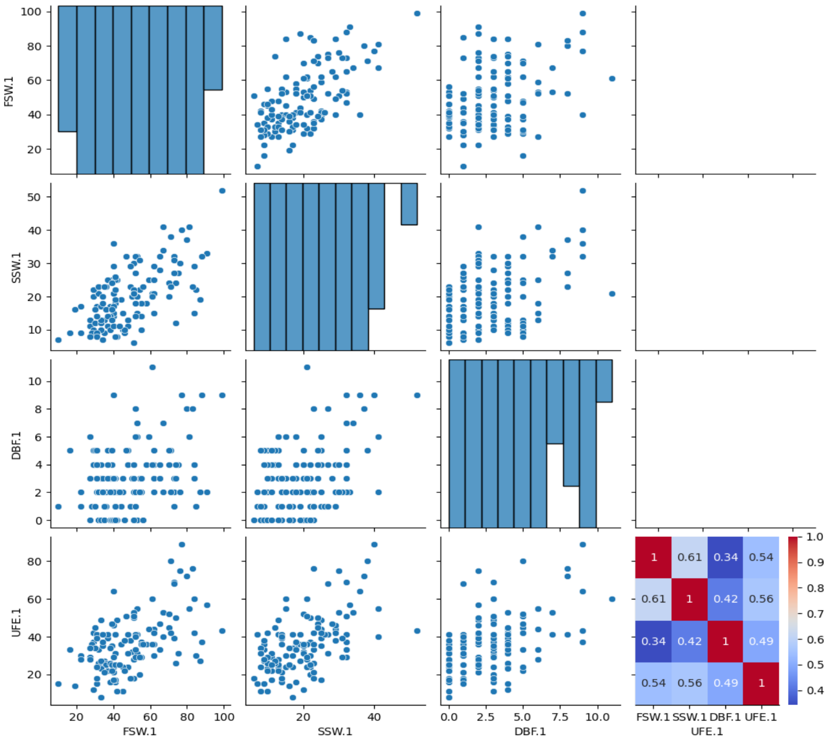
2. What is the relationship between the number of breakpoints and the outcome of the match?

- To determine the relationship between the number of breakpoints and the outcome of the match, we need to first load the data and then group it by the number of breakpoints and the match result. After this, we normalize the data we have grouped.
- To analyze the relationship visually, we can create a bar graph with the x-axis and y-axis as the Number of breakpoints by player 1 and the proportion of matches respectively.
- After this if we observe a trend where the number of breakpoints increases with the chances of winning the match, we can say that there is a positive relationship between the number of breakpoints and the outcome of the match. On the other hand, if we observe that the number of breakpoints does not seem to have a clear relationship with the outcome of the match, we can say that there is no significant relationship between the two variables.



3. How do the serve statistics (FSW, SSW, DBF, UFE) vary from each other? Plot a correlation matrix for all the same.

- To see how the serve statistics of FSW, SSW, DBF, and UFE vary from each other we have to first load the dataset and then create a scatter plot matrix and also to create a correlation matrix of all the desired attributes to make it more visually clear to observe the result.



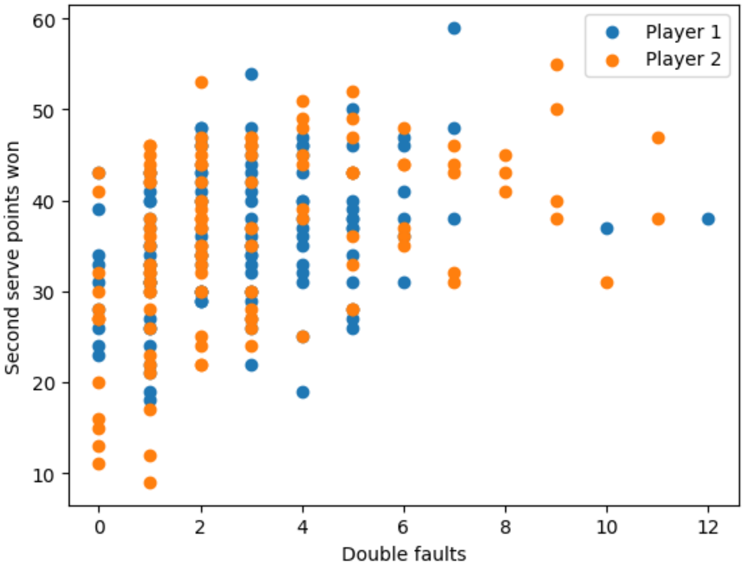
4. Does the covariances between the percentage of second serve points won and the number of double faults committed differ significantly between the first and second player?

- To check whether the covariances between the percentage of second serve points won and the number of double faults committed differ significantly between the first and second players, we need to first load the data and then subset the data for the two players and calculate the covariances.

- After that we perform the hypothesis test and then to make it more visual, we need to plot a scatter plot between SSP vs DBF for each player.

- After all these steps we get the following values

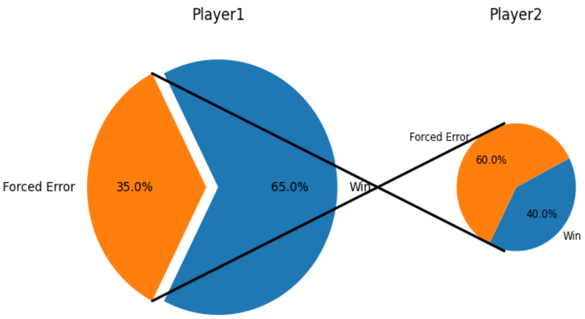
1. Covariance between SSP and DF for player 1: 4.610149488591664
2. Covariance between SSP and DF for player 2: 10.53829006031996
3. T-statistic: 0.4938827485166272
4. P-value: 0.6218371025559402



5. How does the breakdown of points won through forced errors compare between Player 1 and Player 2?

- To compare the breakdown of points won through forced errors between Player 1 and Player 2, we need to first load the data and then make figure and assign the axis object and then add large pie chart parameters for player 1 and then rotate so that first wedge is split by the x-axis.

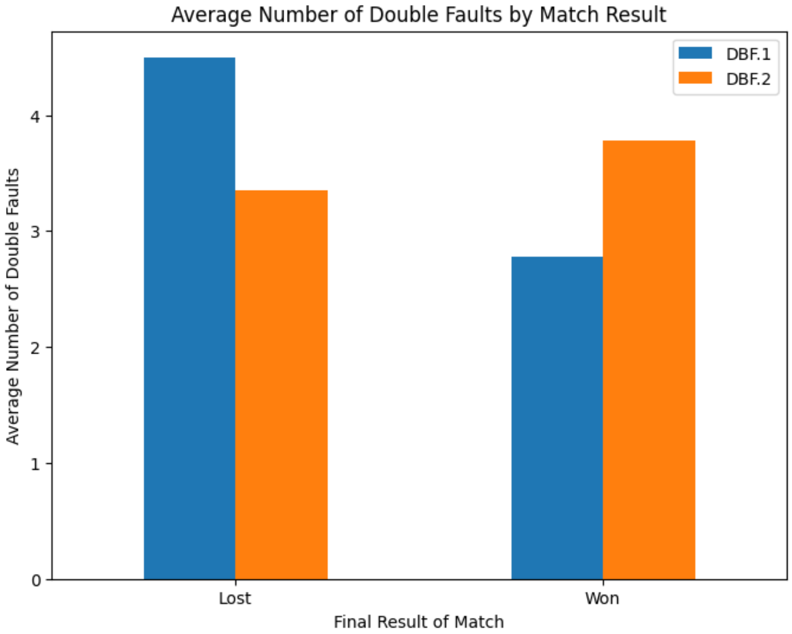
- After that make a small pie chart parameters for player 2 and then use ConnectionPatch to draw lines between the plots and then get the wedge data along with that draw top connecting line and also draw a bottom connecting line. This construction is basically for making a compound pie chart to make it more visual and clearer to understand.



6. How does the final result of the match depend on the average number of double faults in the match for player 1 and player 2?

- To see the dependencies of the final result of the match on the average number of double faults in the match for player 1 and player 2 we need to first load the dataset and then group the data by the final result of the match and then calculate the average number of double faults.

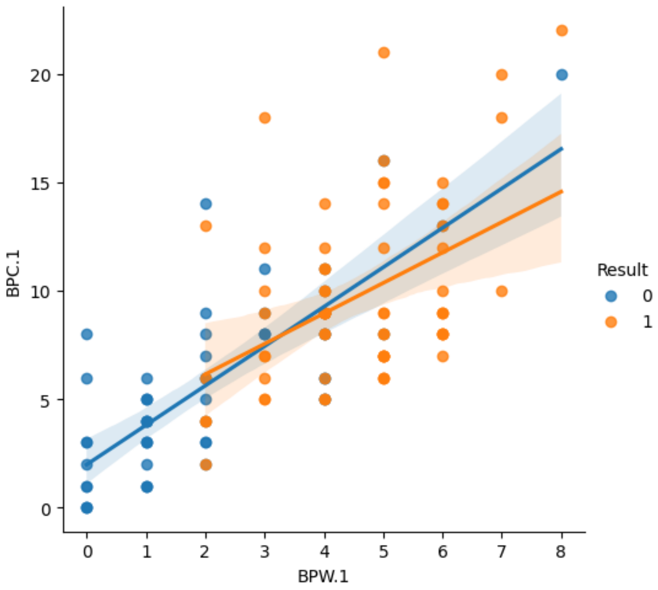
- After these steps then we have to make the plot to make it more clearly visible we need to draw a bar plot with an x-axis and y-axis set as the Final result of the match and the Average number of double faults which were made in match by player 1 and player 2

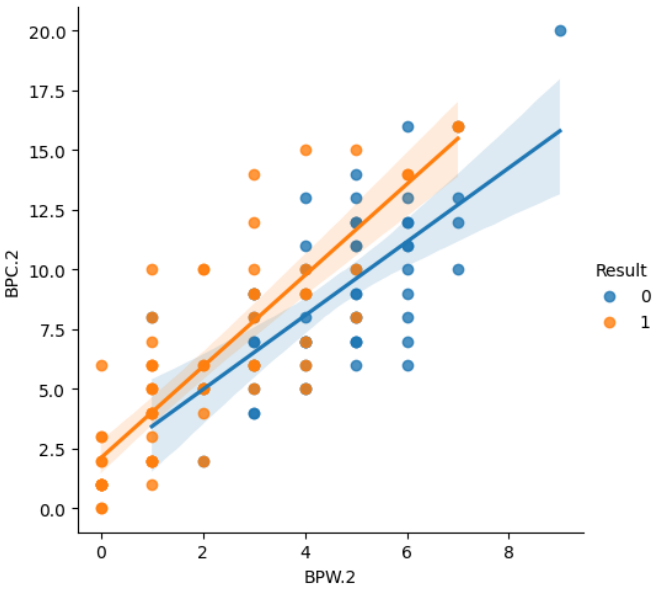


7. Is there a relationship between the number of break points faced and the number of break points saved for winning and losing players in the dataset?

- To check the relationship between the number of breakpoints faced and the number of breakpoints saved for winning and losing players in the dataset, we need to first load the dataset and then calculate the correlation between BPS and BPC for each player after which we printed the correlations.

- And then to make it more visually clear we plot the correlations separately for winning and losing players. After all these steps we get the value of Correlation between BPS and BPC for player 1 as 0.7060351082618115 and also the correlation between BPS and BPC for player 2 as 0.7960885889149741

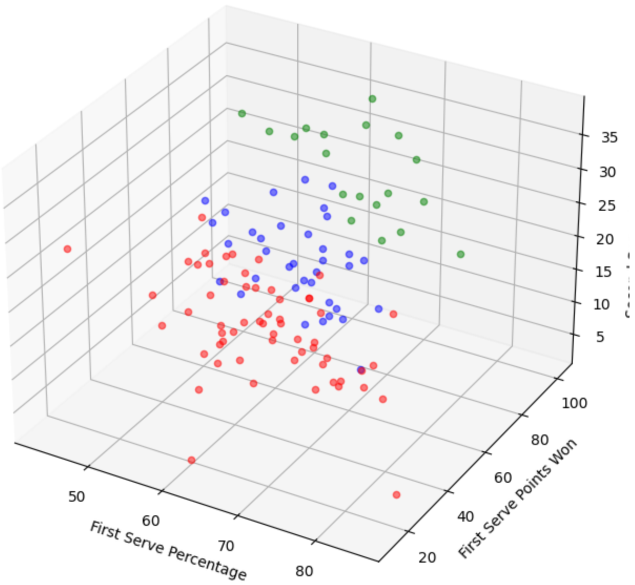




8. Identify clusters of players based on their serving performance, including first serve percentage, first serve points won, second serve percentage and second serve points won.

- To identify the clusters of players based on their serving performance, including first serve percentage, first serve points won, second serve percentage and second serve points won, we first need to load the dataset and then we have to select the appropriate column for clustering and then we have to perform clustering using K-means algorithm and add the cluster labels to the data.

- After all these steps we then plot the results in a 3D space with all the attributes with different colours to make it clearly visible to be defined as clusters.



IV. Summary of the Observations

1. According to the first question we came to an observation that there is a lot of relation between the first serve percentage of Player 1 and the final match result. According to the bar graph, the correlation between these two is maximum in the range of 50-75 per cent of Player 1 first serve and also from the graph we can see that the average match result is 0 for Player 2 wins and 1 for Player 1 wins.

2. Based on the analysis of the relationship between the number of breakpoints and the outcome of the match we can clearly observe that if the number of breakpoints won by player 1 is between 0-4 then player 1 lost the match and if the number of breakpoints won by player 1 is from 16-20 then the player 1 won the match and between these two ranges it could vary but mostly player 1 has won in the middle range also.

3. Based on the analysis, we came to the observation that the two serve statistics are compared to other attributes of FSW, SSW, DBF, and UFE and also there is a correlation matrix corresponding to this which shows the relation between two different attributes at a time.

4. Based on the analysis we can see that there is a good amount of difference between the covariance between Second Serve Points (SSP) and Double faults (DF) for player 1 and player 2 i.e., approximately 5.9. From the graph, we can observe that for player 2 double faults is maximum at the start of the match and also DF and SSP won is maximum for player 2.

5. According to this question, we need to compare the breakdown of points won through forced errors between Player 1 and Player 2 and we have observed from the compound pie chart is that the forced error was maximum for Player 2 as compared to Player 1 and win percentage was maximum for player 1 as compared to player 2.

6. Based on the analysis, we came to the observation that the average number of double faults for player 1 is more in cases when the final result of the match is lost and the average number of double faults for player 2 is more in cases when the final result of the match is won. Hence, the overall plot represents the average number of double faults by match result.

7. Based on the analysis, we came to the observation that there is a relationship between the number of breakpoints saved for winning and losing players in the dataset, and at the end, we came to the observation that 0.7060351082618115 and also the correlation between BPS and BPC for player 2 as 0.7960885889149741.

8. Based on the analysis, we came to the observation that there are three clusters formed after applying the K-means algorithm which is visible in red, blue and green colours associated with the separate clusters which are based on their serving performance, including first-serve percentage, first serve points won, second serve percentage and second serve points won.

Acknowledgment

I would like to express my gratitude to Prof. Shanmuga to give me a chance to work on this data narrative assignment and also for their guidance and support throughout this project.

I would also like to thank all the Teaching Assistants and my friends for their encouragement and assistance throughout this assignment.

Due to this assignment, I have learned a lot of new things which helped me to gain good knowledge in the course ES 114.

REFERENCES

[1] Pandas Documentation
https://pandas.pydata.org/docs/

[2] Matplotlib User Guide
https://matplotlib.org/stable/users/index.html

[3] Scikit-learn User Guide
https://scikit-learn.org/stable/user_guide.html

[4] Seaborn Tutorial
https://seaborn.pydata.org/tutorial/function_overview.html