ES114 Data Narrative 3

Pradeep Kumar Meena (22110196)
Chemical Engineering,
IIT Gandhinagar
Prof. Shanmuganathan Raman

I. Overview Of The Dataset

The tennis tournament dataset contains information on matches played during the year 2013 in Australia, France, USA and Wimbledon for both men and women, including the names of the players, the outcome of the match, and various statistics on the performance of each player. The dataset contains 45 columns and covers different aspects of the matches, including the number of games won, the percentage of successful serves, and the number of aces and double faults. The dataset also includes information on the number of break points saved, the number of net points attempted and won, and the total number of points won by each player. Overall, this dataset can be used to gain insights into the performance of individual players and the overall trends and patterns that emerged during the 2013 tennis tournament.

Overall, this dataset provides a comprehensive look at the matches played during the 2013 tennis tournament and can be used to gain insights into the performance of individual players and the overall trends and patterns that emerged during the tournament.

II. Scientific Questions or Hypotheses

A. *Is there a relationship between a player's unforced errors committed by player 1 (UFE.1) and their success in the tournament, and does this relationship vary depending on whether the player won or lost the match?*

B. *Do players who serve more aces (ACE) tend to have higher first serve percentages (FSP) compared to players who serve fewer aces?*

C. *Is there a relationship between a player's first serve speed (FSP.1) and their success in the tournament, and does this relationship vary depending on the round of the tournament?*

D. *Cluster the dataset according to the player's serve data and how it is helpful.*

E. *Is there a linear relationship between aces won by player 1 and the result? Design a model that can predict the possibility of winning the game by player 1 given the number of aces won. What is the probability of winning a round by player 1, given that the number of aces won by player 1 is 10?*

F. *What is the relationship between a player's unforced errors (UFE.1) and their success rate in winning breakpoints (BPW.1)?*

G. *What is the winning percentage (Result) of players who win more than 40% of their second serve points (SSP.1) and have less than 30 unforced errors (UFE.1) per match?*

H. *What is the average number of winners (WNR.1) hit by players who won the match in FNL.1 = 2 and had a first serve percentage (FSP.1) of at least 60% of the player 1? Visualize the distribution of these winner shots using a histogram.?*

II.

III. Details Of Libraries And Functions

I have used these functions and python libraries in my assignment.

A. *Libraries:*

- Pandas is an open-source library that is made mainly for working with relational or labelled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.[1]
- Matplotlib: Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.[2]
- Seaborn: Seaborn is a Python data visualization library based on Matplotlib It is built on top of Matplotlib and integrates with the data manipulation capabilities of Pandas. Seaborn offers a variety of visualizations, including scatterplots, line plots, bar plots, heatmaps, and more.[3]
- Sklearn: Scikit-learn, or sklearn, is a popular machine learning library for Python. It provides a range of tools for implementing various machine learning algorithms, such as classification, regression, clustering, and dimensionality reduction.[4]

B. *Functions:*

- pandas.read_csv(): To read and store data from csv file to a Pandas dataframe. [1]
- matplotlib.pyplot.scatter(): Makes a scatter plot of x versus y. [3]

- values_count(): It is used to get the occurrence of any list of strings for the top element of a series object.
- hist(): plots the histogram of a Series
- plot(): It is used to plot the graph this function is plt.xlabel() and plt.ylabel() functions are used to label available in both pandas and matplotlib.
- bar(): It is the function available in matplotlib to plot bar graph.

IV. Answers Of The Questions

A. Is there a relationship between a player's unforced errors committed by player 1 (UFE.1) and their success in the tournament, and does this relationship vary depending on whether the player won or lost the match?

Approach: By reading the CSV file "AusOpen-men-2013.csv", to achieve the target, first we have to separate the data into two groups matches that were won by player 1 and matches that were lost by player 1 And then create a scatter plot to look at the trend.

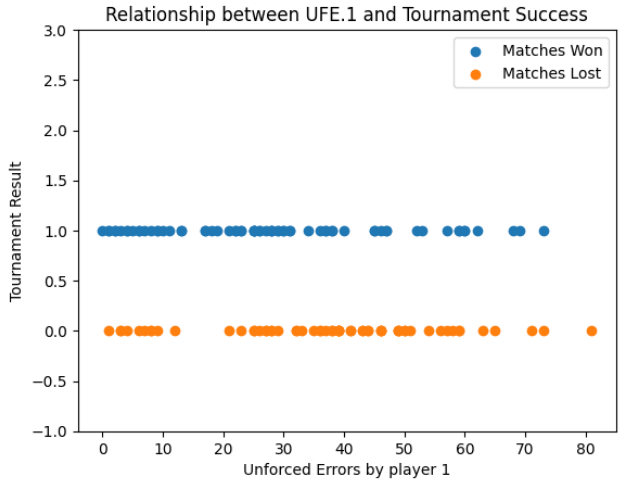

Fig.1 Tournament result vs Unforced error made by player 1

On the y-axis of Fig. 1, I have plotted the tournament results in which 1 represents that the player wins and 0 represents that the player loses the match. By observing the graph I saw that the blue points are clustered in the starting when the unforced error made by player 1 is less. This shows that the positive relationship between the tournament result and the unforced errors. So, we can conclude that when the unforced errors are fewer then the player has more chance to win and if the player has a high he/she has less chance to win.

B. Do players who serve more aces (ACE) tend to have higher first serve percentages (FSP) compared to players who serve fewer aces?

Approach: To get the result first divide the dataset into two groups matches where player one served more than their opponent and matches where player 1 served fewer aces than their opponent. Then, we compare the average first-serve percentage for both groups.

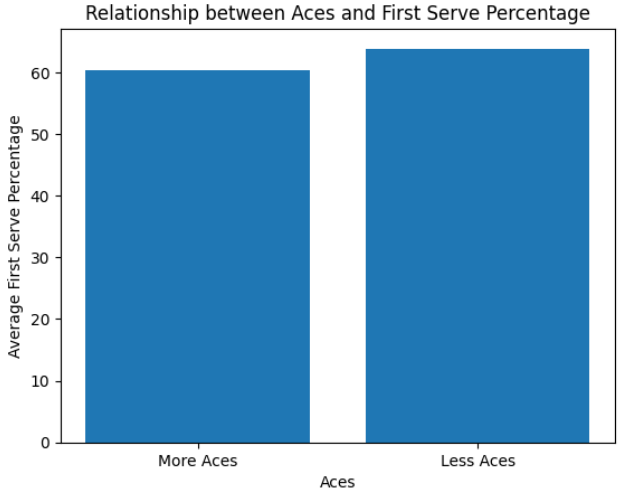

Fig.2 Average first serve vs Aces

By observing the Fig.2, we can see that those players which have more number of aces tends to have a low average first-serve percentage. Hence, there is negative relationship between the Average first serve and more Aces than their opponent.

C. Is there a relationship between a player's first serve speed (FSP.1) and their success in the tournament, and does this relationship vary depending on the round of the tournament?

Approach: First start by grouping the data by the round of the tournament and calculating the average FSP1 for matches won and lost in each round. Then, we can create a line plot to visualize the relationship between FSP1 and tournament success for each round separately.

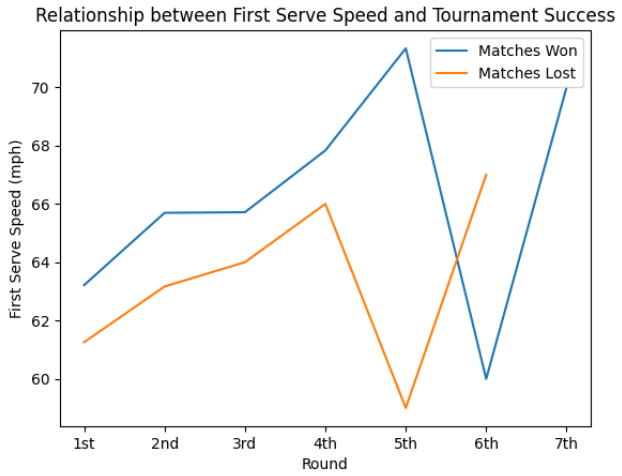

Fig.3. First serve speed(mph) vs the rounds

By looking at Fig.3, we can say that those rounds which start with first person with high speed serve tends to have more chance to go to the next tournaments or to the next round. The figure shows that those matches which was won have high-speed serve. And yes, this relationship depends on the round of the tournament.

D. Cluster the dataset according to the player's serve data and how it is helpful.

Approach: The approach is simple use the columns FSP1, and FSW1, which represent player 1's first serve points won percentage and first serve points won and fit the means and select the number of clusters is equal to 4 and assign points to the nearby clusters.

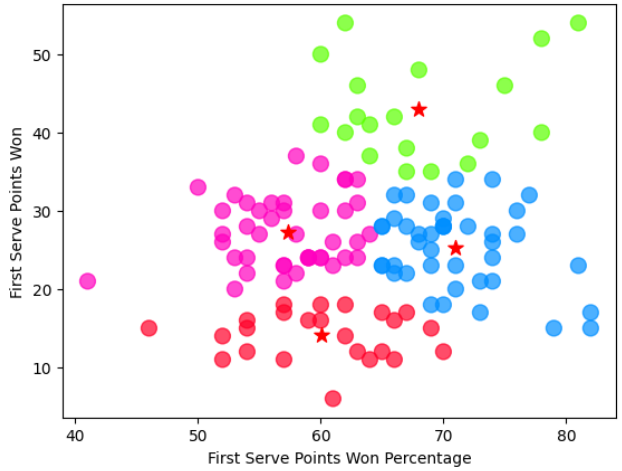

Fig.4 Clusters of the players and with their centroids

Fig. 4 shows four clusters with their centroids, By examining the clusters, we can classify the players on the basis of their serve and find that there are four types of players. From the figure, the green cluster is cluster consists of players with high first serve points won percentage and high first serve points won, indicating that these players have a strong first serve. The red cluster is the cluster with low first-serve points won percentage and low first-serve points won. This indicates that these players have weak first serve.

E. Is there a linear relationship between aces won by player 1 and the result? Design a model that can predict the possibility of winning the game by player 1 given the number of aces won. What is the probability of winning a round by player 1 given that the number of aces won by player 1 is 10?

Approach: Plot the scatter graph between aces won by player 1 and the result. And note the trend. By using the linear regression model, we can predict the possibility of winning the game by player 1.

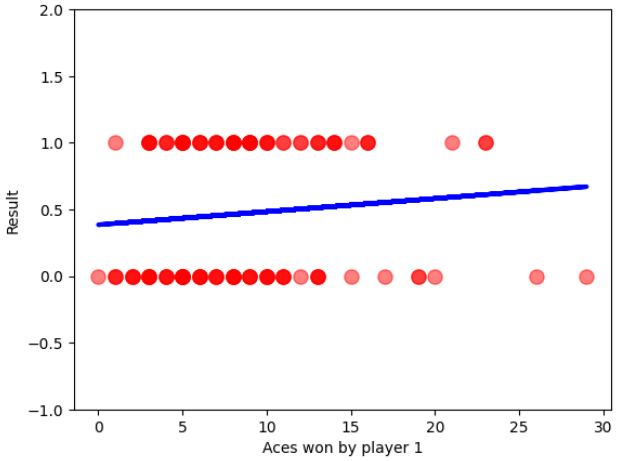

Fig.5 Relationship between aces won by player 1 and the result

By analysing the Fig.5 plot, we can say that as the number of aces won by player 1 increases, the probability of winning a round by player 1 also increases. Hence, there is a linear relation between the two parameters. We can also predict the possibility of winning the game by player 1 using a linear regression model. The slope and the intercept of the predicted line are 0.00985 and 0.3844, respectively.

The probability of winning the game by player 1, given that number of aces won by player 1 is 10, is 0.48289.

F. What is the relationship between a player's unforced errors (UFE.1) and their success rate in winning breakpoints (BPW.1)?

Approach: Plot a scatterplot to visualize the relationship between the two variables. Linear regression analysis is then performed using the sklearn library, with 'UFE.1' as the independent variable and 'BPW.1' as the dependent variable. The coefficients and R-squared value of the linear regression model are printed to assess the strength and direction of the relationship between the two variables.

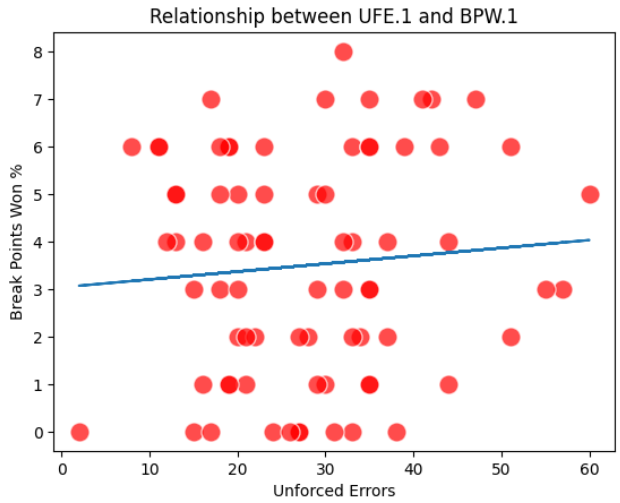

Fig.6 BPW.1 vs UFE.1

The scatterplot of Fig.6 clearly shows that there is no relationship between these two variables. We can also prove it by the Linear regression model. The R-squared value of the model indicates how well the model fits the data. A high R-squared value indicates a strong relationship between the variables. In this model, we found the R-squared value of the model is 0.00738, which is very low and indicates that there is no effective relationship between the percentage of breakpoints won an unforced error of player 1.

G. *What is the winning percentage (Result) of players who win more than 40% of their second serve points (SSP.1) and have less than 30 unforced errors (UFE.1) per match?*

Approach: To find the winning percentage of players, we need to first filter the dataset which has SSP.1>40% and UFE.1 < 30 from the given dataset. Then calculate the winning percentage for the new data.

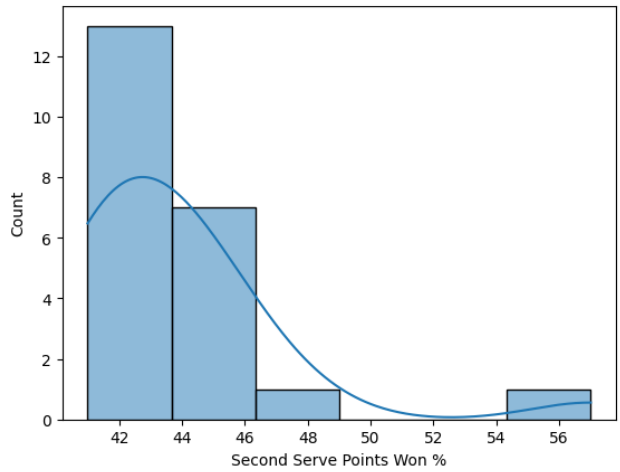

Fig.7 Distribution of SSP.1 values

Fig. 7 shows the histogram plot of SSP.1 values and the KDE plot. The winning percentage of players who win more than 40% of their second serve points and have less than 30 unforced errors is 36.36 percent.

H. *What is the average number of winners (WNR.1) hit by players who won the match in FNL.1 = 2 and had a first serve percentage (FSP.1) of at least 60% of the player 1? Visualize the distribution of these winner shots using a histogram.?*

Approach: First, we need to filter the dataset to include only those matches data where FNL.1 = 2, and FSP.1 is at least 60%. Then we can calculate the average number of winners hit by these players using the WNR.1 column. And finally, create a histogram to visualize the distribution of these winners.

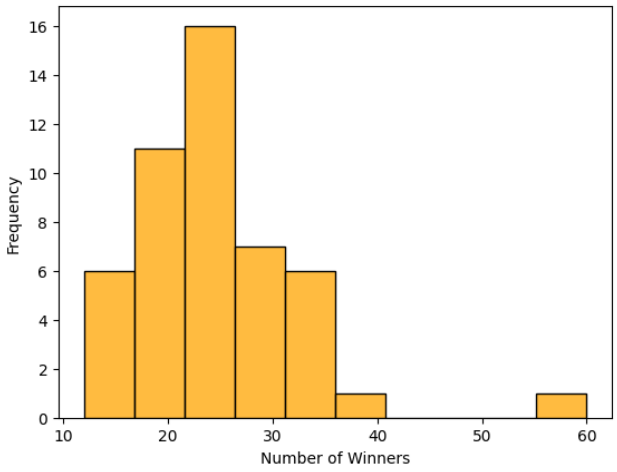

Fig.8 Distribution of winners

Fig.8 shows the distribution of the winners who won the match in FNL.1 = 2 and had a first-serve percentage (FSP.1) of at least 60% of player 1. From Fig. 8, the maximum number of occurrences of number of winners lies between 20 and 30 with the occurrence of 16, and the minimum number of occurrences of number of winners is between 40 and 50 with the occurrence of 0.

VI. References

[1] "Introduction to pandas in Python," *GeeksforGeeks*, 09-Feb-2023. [Online]. Available: https://www.geeksforgeeks.org/introduction-to-pandas-in-python/.

[2] *Matplotlib bars*. [Online]. Available: https://www.w3schools.com/python/matplotlib_bars.asp.

[3] "Statistical Data Visualization#," *seaborn*. [Online]. Available: https://seaborn.pydata.org/. [Accessed: 22-Apr-2023].

[4] “1. supervised learning,” *scikit*. [Online]. Available: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning.

VII. Acknowledgements

I am also grateful to Prof. Shanmuganathan Raman for giving the opportunity to work on this dataset. I am also thankful to Mrs Seema for helping me throughout this Assignment.