ES 114 - Data Narrative - III

Hrriday Ruparel
22110099
Freshman, Computer Science and Engineering
Indian Institute of Technology
Gandhinagar, India
hrriday.ruparel@iitgn.ac.in

***Abstract*—This is an inferential and statistical data analysis report on eight sets of datasets containing the math statistics of four major tennis tournaments held in 2013. The data analysis explores the data from a statistical standpoint and draws some conclusions based on the inferences and observations made using data analysis and data visualization tools supported by Python.**

I. Overview of Dataset

The entire dataset includes 8 files that contain information of math statistics for both men and women at four major tennis tournaments of the year 2013. Each of the dataset has around 127 instances/match records of tennis games held in the tournaments distributed over around 42 attributes related to tennis. Among these 42 attributes, the major ones include the result of the match, the round number, final number of points won by each player, records of set scores, number of breakpoints won, unforced errors and aces.

The dataset contains 8 important files relevant for our discussion.

1. **AusOpen-men-2013.csv** – Data related to Men's Australian Open 2013.
2. **AusOpen-women-2013.csv** – Data related to Women's Australian Open 2013.
3. **FrenchOpen-men-2013.csv** – Data related to Men's French Open 2013.
4. **FrenchOpen-women-2013.csv** – Data related to Women's French Open 2013.
5. **USOpen-men-2013.csv** – Data related to Men's US Open 2013.
6. **USOpen-women-2013.csv** – Data related to Women's US Open 2013.
7. **Wimbledon-men-2013.csv** – Data related to Men's Wimbledon 2013.
8. **Wimbledon-women-2013.csv** – Data related to Women's Wimbledon 2013

The datasets had many missing values already present in it. In order to draw some informative results out of the data, the datasets were cleaned at each step and hence, the number of data points used to answer each question varied from question to question.

II. Scientific Questions and Hypotheses

Some useful scientific questions and hypotheses made using U.S. News data for the data analyses are as follows:

1. *Design a performance index for each player that could be helpful for further analysis. Explain your approach.*
2. *Do the players play according to similar styles? Elaborate.*
3. *It is said that a good start means job half done. Does this idea apply to tennis? Does good first serves and aces imply winning the match?*
4. *Devise an estimator that would estimate the result of the match based on certain selected features of the match. Show the accuracy score of it. Explain in detail.*
5. *Is there any relationship between number of first serves won by player 1 and player 2 with the number of second serves won by them? Elaborate.*
6. *For the women's tournaments, is there any relationship between the games won by player 1 and player 2 for the 1st set with the number of games won by them in the 2nd set? If there is, explain the relationship.*
7. *What can you say about the performance of the finalists over the time as the tournament progressed? Was it evident who the winner of the tournament was going to be based their performances prior to the finals?*
8. *What conclusions can you draw out of the errors that finalists made as the tournament progressed? Explain the trends involved.*

III. Details of Libraries and Functions

The data narrative has used the various data analysis and data visualization tools supported by Python. These include:

1. **NumPy** – NumPy, also known as Numeric Python, is a Python library provides tools to deal with arrays and large datasets. This library is imported into Python as **np**.
2. **Pandas** – Pandas is Python library that allows the user to create table-like structures, organise data, perform operations on the data and compute various important statistical measures. Pandas has been imported as **pd**.
3. **Matplotlib** – Matplotlib is a Python library that provides the facility to visualize the data by providing tools for plotting graphs. The **pyplot** functionality of this library has been imported as **plt**.
4. **Sklearn** – Sklearn is a Python library that is an extension to the existing SciPy library with the added benefit of applying machine learning concepts in Python.
5. **Seaborn** – Seaborn is a Python library that is useful for data visualization, plotting various kinds of graphs and also for cutomising them. It has been imported as **sns**.

Each of these libraries have been extensively used for answering scientific questions stated in the previous section. Some of the notable functions of these libraries used for data analysis and visualization are:

1. **np.histogram2d()** – This functions helos to compute bi-dimensional histogram of two data variables.
2. **np**.**meshgrid() -** This functions returns coordiante matrices fomr coordinate vectors. It is used for 3D plotting.
3. **np.ones_like()** – This function returns an array of ones with the same shape and size as that of the given array.
4. **pd**.**DataFrame()** – This function helps to create a new DataFrame.
5. **pd.read_csv()** – This function helps convert .csv files to Pandas DataFrame.
6. **pd**.**DataFrame.drop()** – This funtion helps to remove rows or columns.
7. **plt.plot()** – This function helps to create plots of variousb types like bar, hist, scatter, pie, density, etc.
8. **plt**.**show()** – This function is used to view the plotted graph on the the figure window.
9. **plt.subplots()** – This function helps to work with subplots on a single figure. It provides tools for customisation of subplots and their layouts.
10. **plt.title()** – This function sets the title for the graphs.
11. **plt**.**xlabel()** & **plt.ylabel()** – These function help to label the x-axis and y-axis respectively.
12. **plt.xticks()** & **plt.yticks()** – These functions help to customise the ticks on the x-axis and y-axis respectively.
13. **plt.grid() –** This function displays the grid on the graphs.
14. **plt.legend()** – This function displays trh legend on the graph.
15. **sklearn.KMeans**.**fit –** This function helps to compute k-means clustering on the data provided.
16. **sklearn.KMeans**.**predict** – This function helps to compute the closest cluster each datapoint in a dataset belongs to.
17. **sklearn.neighbors**.**KNeighborsClassifier** –It is the classfier implementing the K nearest neighbot votes.
18. **sklearn.model_selection**.**train_test_split** – This function helps to obtain training and testing subset data by splitting the input arrays.
19. **sklearn.metrics**.**accuracy_score** – This function helps to compute the classification accuracy.
20. **sklearn.metrics**.**confusion_matrix** – This function helps to compute the confusion matrix to evaluate the accuracy of the classification.
21. **sklearn**.**decomposition.PCA** – It is a class that helps to perform Principal Component Analysis on a given dataset.
22. **sklearn**.**preprocessing.StandardScaler** – this function helpf to normalize/scale the data.
23. **sns**.**heatmap()** – This function helps to plot a heatmap. Its plots are more beautiful than what Matplotlib offers/

These are the major functions and methods used from the the imported Python libraries.

IV. ANSWERS TO SCIENTIFIC QUESTIONS AND HYPOTHESES

To answer the first question, the 'AusOpen-men-2013.csv' dataset was chosen. The main aim was to design a performance index or ranking mechanism for each player in the tournament. Initial approach was to implement a weighted average of certain selected features which would be assigned some weights Having the knowledge of neural networks would have simplified the problem of choosing appropriate weights. Since was a bit far-fetched topic, I came up with a way of assigning weights.

The correlation matrix is a matrix that calculates the correlation between the entries of an array pairwise. After dropping some irrelevant columns like 'Player1' and 'Player2' for which calculating correlation made no sense, a correlation matrix on 40 various features of the tennis match was formed as shown in Fig. 1a.

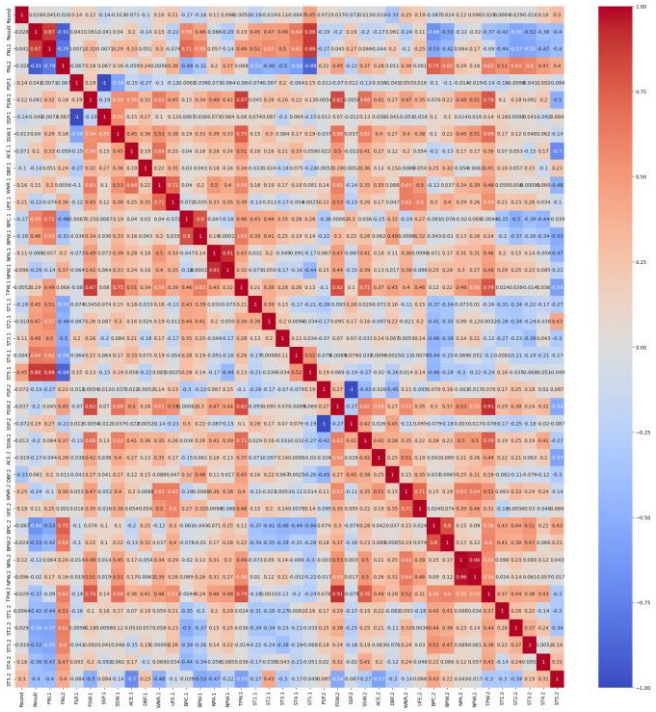

Fig. 1.a – Correlation matrix of 40 features of the tennis match

Out of the entire correlation matrix, the required row of correlation values was that of the 'Result' variable. This row would have the correlation of the each feature with the 'Result' variable and the correlation itself would act as a weight for that particular feature. The 'Result' row is depicted in Fig. 1b.

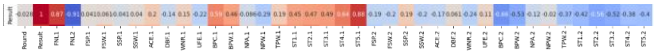

Fig. 1b – Correlation row of the 'Result' variable.

Taking these as weights, the performance index for each player as calculated according to the following equation:

$$score = \frac{w_1 * f_1 + w_2 * f_2 + \cdots + w_n * f_n}{w_1 + w_2 + \cdots + w_n},$$

where $w_1, w_2, \ldots, w_n$ are the weights (correlation values) of features and $f_1, f_2, \ldots, f_n$ are the feature values for a particular player.

The top 10 players in the performance index with respect to Men's Australian Open 2013 tournament are depicted in Fig. 1c.

```
Player rankings by performance index:
1. Daniel Brands: 158.61
2. Pablo Carreno Busta: 155.68
3. Marinko Matosevic: 143.71
4. Marcos Baghdatis: 130.12
5. Jordan Thompson: 129.42
6. Dudi Sela: 126.88
7. Di Wu: 126.20
8. Alexandr Dolgopolov: 125.06
9. David Ferrer: 123.95
10. Blaz Kavcic: 123.44
```

Fig. 1c - Performance index.

From Fig. 1c, one can clearly see the player name and the respective score of that player according to the algorithm. However, none among the semi-finalists and the finalists shown in Fig. 1d appear on the top of the performance index.

Fig. 1d - Fixture representing semi-finalists and finalists. '6' represents semi-final rounds and '7' represents final round.

It turns out that the algorithm is not right and to generate a performance index, one might need to bring in neural networks to assign appropriate weights.

However, one positive takeaway is feature modelling. From the correlation matrix, one can find the features that are strongly positively correlated or negatively correlated, thus providing one with reliable features to make any predictions in the future. The reliable features so obtained are showcased in Fig. 1e given below.

```
Important features Positive Correlation: ['Result', 'FNL1', 'BPC.1', 'ST4.1', 'ST5.1']
Important features Negative Correlation: ['FNL2', 'BPC.2', 'BPW.2', 'ST2.2', 'ST3.2']
```

Fig. 1e – Important features

Among these features, the final scores, 'FNL1' and 'FNL2', are irrelevant since the 'Result' variable is directly derived after comparing 'FNL1' and 'FNL2'. Hence, these can be dropped. 'Result' variable can also be dropped since it is the variable whose correlation is being found out with other features. The set scores, 'ST4.1', 'ST5.1', 'ST2.2' and 'ST3.2' are unique to a specific tournament and vary across different tournaments. Hence, the only reliable features include the data related to breakpoints, 'BPC.1', 'BPC.2' and 'BPW.2'. These would be used in the future to make predictions.

To answer the second question, the 'AusOpen-women-2013.csv' dataset was chosen. The main motive is to perform clustering on the dataset and obtain inferences from the clustered data. However, the dataset consisted of around 40 features. In order to visualize the data, principal component analysis was performed and the data was reduced to 2 and 3 principal components. KMeans clustering was performed on the reduced dimensional data and the following observations were made:

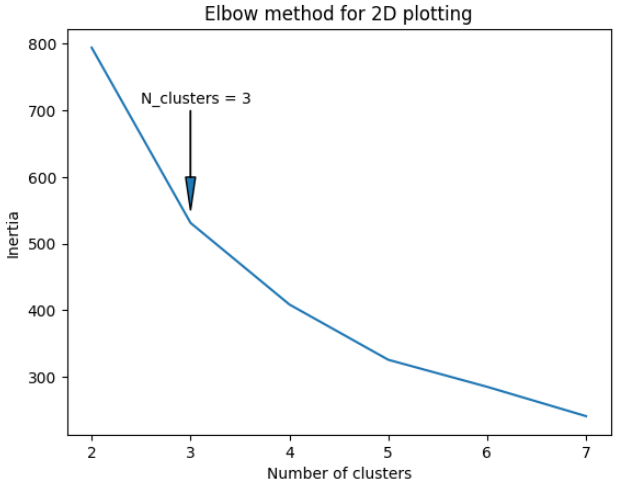

Fig. 2a – Elbow method analysis for 2D plotting.

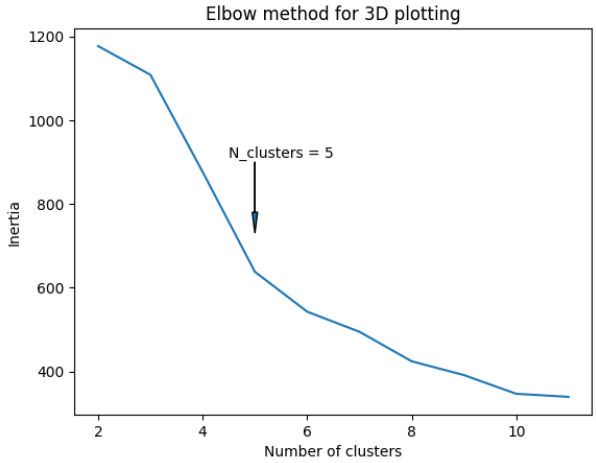

Fig. 2b – Elbow method analysis for 3D plotting.

The optimal number of clusters for 2-dimensionla reduced data was found out to be 3 and that of 3-dimensional reduced data was found out to be 5 from the elbow method.

As a result, the following scatter plots consisting of clustered data was obtained.

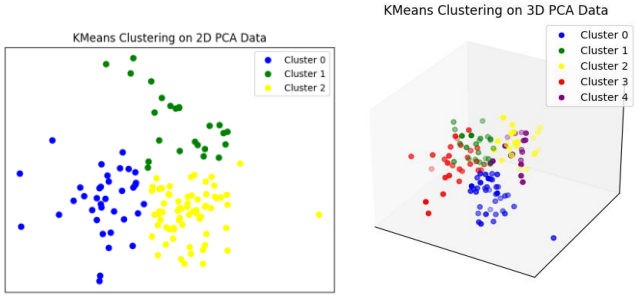

Fig. 2c (left) and Fig. 2d (right) – KMeans applied on reduced data.

From the figures Fig. 2c and Fig. 2d, it is clear that the clusters formed are quite close to each other, implying that the playing style of the players is quite similar, but there are some outliers which can be visually spotted.

On further analysis, the clusters seemed to be clearly separated when each clusters 'WNR.2' feature was plotted as a violin plot as follows:

Fig. 2e – Violinplot of WNR.2 feature for 2D data

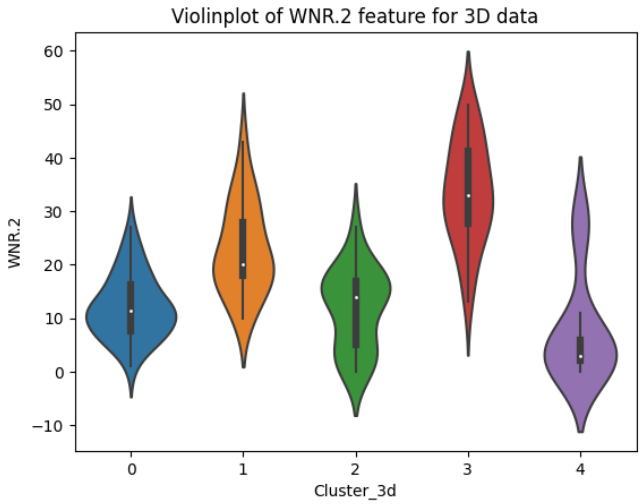

Fig. 2f - Fig. 2e – Violinplot of WNR.2 feature for 3D data

From the violinplots, it is clear that the 'WNR.2' (winners earned by player 2) feature clearly distinguish the clusters from each other. One can say that the clusters represent different classes of players based on the winners earned by them.

To answer the third question, 'FrenchOpen-men-2013.csv' dataset was chosen. The main objective was to investigate whether good first serves and aces implied winning the match, that is, a good start implies job half done. By virtue, the initial approach was to calculate the correlation values of the necessary feature against 'Result' variable. This would provide us with the basic knowledge of whether performing data analysis on these features is relevant to the question or not. Hypothesis says that the more the number of aces and good number of first serve wins ensure the victory of the player. In order to test this, a table of correlation values (as derived in answer to first question) was obtained. The features chosen to test the above hypothesis included the number of aces, the number of first serve wins, the percentage of first serves, number of second serve wins, percentage of second serves and number of double faults by each player in a match.

	Correlation Values
FSP.1	0.165377
FSW.1	0.140262
SSP.1	-0.165377
SSW.1	0.045168
ACE.1	0.237969
DBF.1	-0.142763
FSP.1	-0.082201
FSW.2	-0.196195
SSP.2	0.082201
SSW.2	-0.123538
ACE.2	-0.352246
DBF.2	0.166896

Fig. 3a – Correlation values of the selected features against 'Result' variable.

From Fig. 3a, it is clear that the correlation values for each of the feature are less than 40% in magnitude, implying that there is very little contribution of the these features to the overall result of the match. However, among these features, the number of aces scored by the players has a significant contribution (23.79% for player 1 and 35.22% for player 2) to the match result.

Hence, the hypothesis that a "good start implies job half done" does not apply to tennis as the match results depends on many other factors (like breakpoints won, breakpoints created, etc.) whose contribution to the match result is much more significant.

To answer the fourth question, 'FrenchOpen-women-2013.csv' dataset was chosen. The task was to design an estimator that would predict the results of a given match with the help of some reliable features. Importance must be laid on the quality and quantity of the features. The 'quality' aspect of the features was ensured from the previous results of feature modelling in the first question. It was derived that data related to breakpoints could be a good indicator of the match result. Moreover, choosing many features is definitely a good start but, it may immediately lead to data over-fitting issues.

Therefore, taking into consideration the quality and quantity of the features for estimation, the following features were selected to predict the match result of Men's French Open 2013 tournament:

- Breakpoints created by player 1 (BPC.1)
- Breakpoints won by player 1 (BPW.1)
- Breakpoints created by player 2 (BPC.2)
- Breakpoints won by player 2 (BPW.2)

The estimator model used here was K-Nearest Neighbours algorithm with the value of K set to 5 neighbours. The dataset was split with the help of train_test_split() function of the sklearn library such that 80% of the data was trained and tests were performed on the rest 20% of the data.

The accuracy of the results of estimation were computed with the help of sklearn's accuracy_score() and confusion_matrix() functions. The results obtained were as follows:

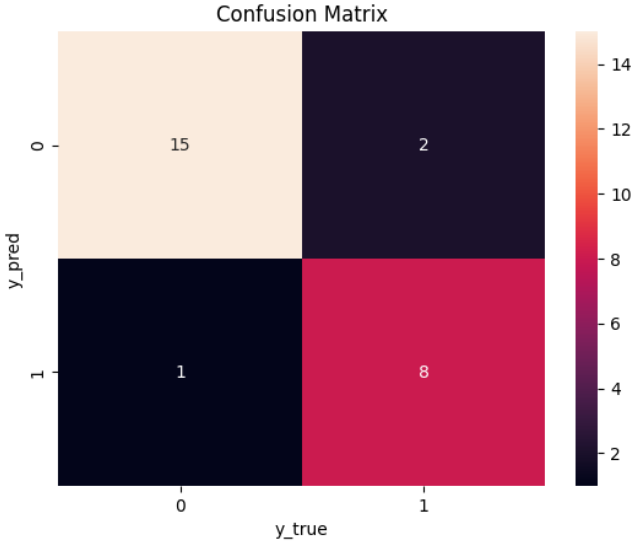

Fig. 4a – Confusion matrix derived to check accuracy of the estimation. Here, the axis labels (0 and 1) depict the value of the 'Result' variable ('1' implying player 1 won the match and '0' implying player 2 won the match)

The accuracy score so obtained for the classification was 0.8846153846153846, which is fairly high for estimation based on just 4 features.

This proves the importance of breakpoints as essential parameters in determining match results.

To answer the fifth question, 'USOpen-men-2013.csv' dataset was chosen. The main focus is to draw relationships between the number of 1st serves won by player 1 and player 2 against the number of 2nd serves won by them. In order to incorporate 4 variables (FSW.1, FSW.2, SSW.1 and SSW.2) into discussion, the difference of the FSW.1 and FSW.2 was taken and a new variable 'X' was created. Similarly, taking the difference between SSW.1 and SSW.2 would produce a new variable, say 'Y'. The sign of the new variables would indicate which player had more serve wins and the magnitudes of the new variables would indicate margin.

Plotting a joint distribution of the two new discrete variables in 2D was sufficient to draw relationships between the two.

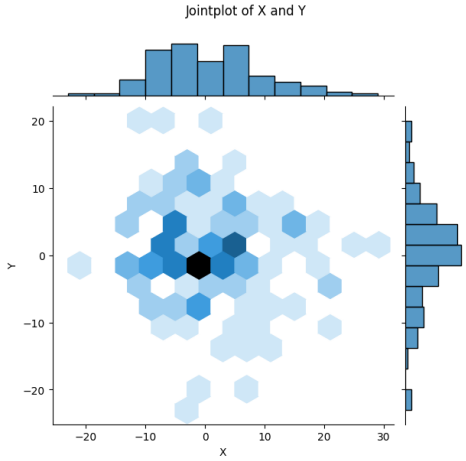

Fig. 5a – Joint distribution of X and Y.

From Fig. 5a, one can draw conclusion that there is high density in the region roughly lying between X = -10, X = 0, Y= -10 and Y= 10. In this region, the majority has the sign of X as negative implying that on an average player labelled 'Player 2' in the match had more first serve wins as compared to player labelled 'Player 1'. Similarly, the majority has the sign of Y as negative implying that on an average player labelled 'Player 2' in the match had more second serve wins as compared to player labelled 'Player 1'. Moreover, the marginal probability mass functions clearly show that the second serve wins follow the Normal/Gaussian distribution, but with some outliers.

To answer the sixth question, 'USOpen-women-2013.csv' dataset was chosen. Similar to the fifth question, the main focus is to draw relationships between the number of games won in 1st Set by player 1 and player 2 against the number of games won in 2nd Set by them. In order to incorporate 4 variables (ST1.1, ST1.2, ST2.1 and ST2.2) into discussion, the difference of the ST1.1 and ST1.2 was taken and a new variable 'X' was created. Similarly, taking the difference between ST2.1 and ST2.2 would produce a new variable, say 'Y'. The sign of the new variables would indicate which set dominated the other and the magnitudes of the new variables would indicate margin.

The 3D distribution plot of the two new discrete variables X and Y is depicted below.

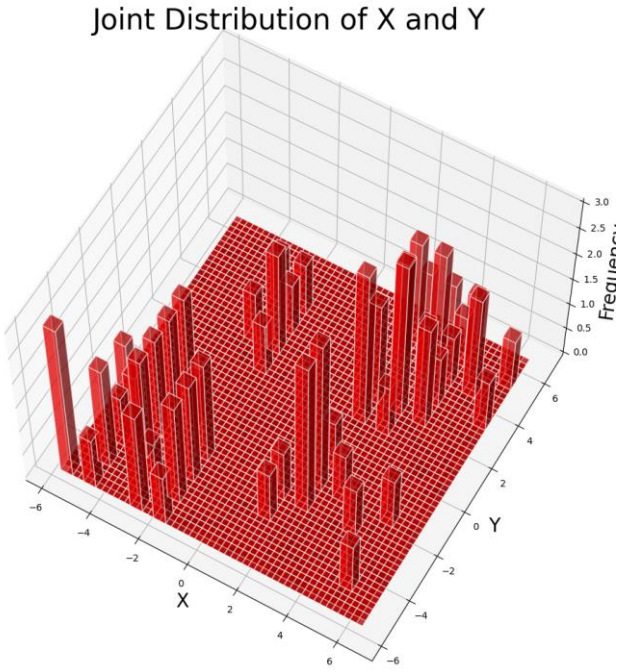

Fig. 6a – Joint distribution of X and Y.

From Fig. 6a, if a match lies in:

- 1st quadrant, then 1st player dominated the 2nd player in both the sets.
- 2nd quadrant, then 1st player dominated the 2nd player in Set 2 but not in Set 1.
- 3rd quadrant, then 2nd player dominated the 1st player in both the sets.
- 4th quadrant, then 2nd player dominated the 1st player in Set 2 but not in Set 1.

If a match is farther from origin, then it implies that the margin of dominance is higher.

From Fig. 6a, one can draw conclusion that there is high density in the region roughly lying in the 1st and 3rd quadrant of the X-Y. Among 1st and 3rd quadrant, 3rd quadrant has even higher density that 1st quadrant implying that on an average the player labelled 'Player 2' generally dominates player labelled 'Player 1' in both the sets. The high frequencies of the matches in the 3rd quadrant far from the origin implies that the margin of dominance was also very high. The graph also implies that usually if a player wins the 1st set, then the player is likely to win the next set as well.

To answer the seventh question, 'Wimbledon-men-2103.csv' dataset was chosen. The main motive of the question was to come up with trends showing the performance of the finalists of Men's Wimbledon 2013 tournament as the tournament progressed. The finalists of the tournament were Novak Djokovic and Andy Murray, out of which Andy Murray won the Grand Slam with a whopping set score 3-0.

In order to analyze the performance progress of the finalists, the trusty and reliable breakpoint related data was chosen and the data for each round was plotted for each player.

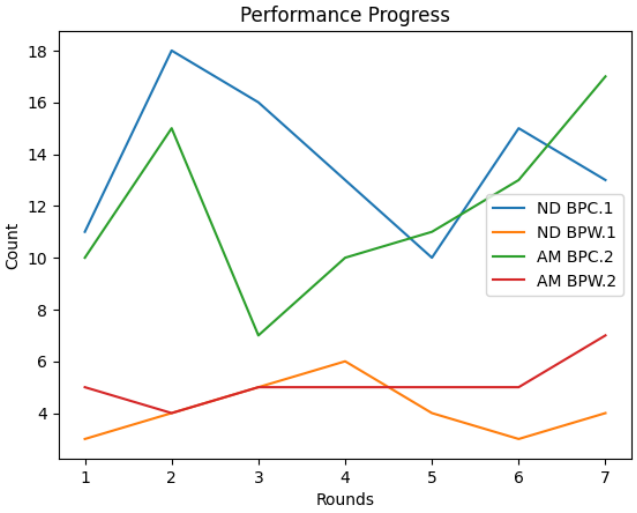

Fig. 7a – Performance progress of each player.

The line chart clearly shows that Andy Murray deserved to win the Grand Slam as compared to Novak Djokovic. Andy Murray's statistics of breakpoints created (in green) and breakpoints won (in red) had a non-decreasing trend midway through the tournament and continued to improve himself despite rising tensions of the later games. On the other hand, Novak Djokovic's performance declined midway through the tournament, but still managed to enter into the finals.

To answer the final eighth question, 'Wimbledon-women-2013.csv' dataset was chosen. The main objective of the question was to come up with trends showing the errors that the finalists made over the tournament and how they progressed. The finalists of the tournament were Sabine Lisicki and Marion Bartoli, out of which Marion Bartoli won the Grand Slam with a whitewashing set score 2-0.

In order to analyze the error progress of the finalists, the data related to errors namely double faults and unforced errors was chosen and the data for each round was plotted for each player.

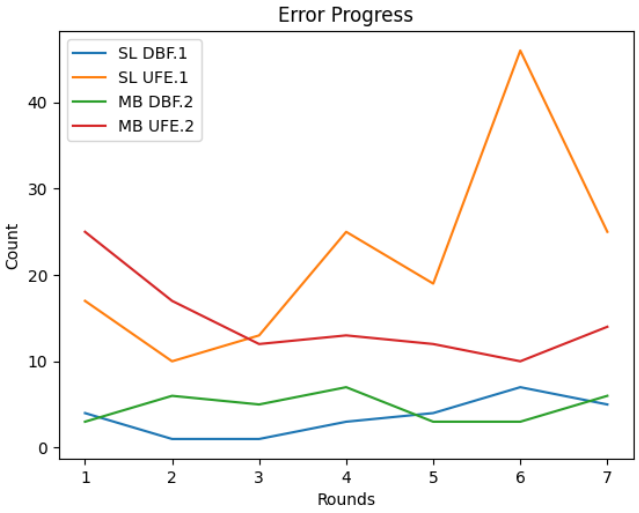

Fig. 8a – Error progress of each player

The line chart shows that Marion Bartoli deserved to win the Grand Slam because she controlled her errors, indicating improvement in her performance despite rising difficulty in the tournament. While in the case of Sabine Lisicki, the errors kept on increasing and touched all-time high in the semi-finals.

V. Summary of Observations and Learnings

The data sets provided an insight into the various variables that are associated with the tennis tournaments and helped develop an understanding how these variables are interconnected and correlated to each other. The players who consistently improve their performance over the tournament by improving the statistics related to breakpoints and reduce their errors are the victors. These players dominate their opponents by straight away winning the majority of the sets with huge game point margins.

Learning to plot graphs in 3D and incorporating machine learning techniques to predict outcomes was a new experience.

VI. Unanswerable Questions

1. Why the performance index created for 1st question failed to produce desirable results?
2. Why

VII. Acknowledgement

This is to thank Professor Shanmuganathan Raman for providing this opportunity to work on a dataset and perform data analysis and visualization in order to extract some useful information.

VIII. References

1. 'NumPy documentation,' accessed Apr 26, 2023, https://numpy.org/doc/1.24/
2. 'Pandas documentation,' accessed Apr 26, 2023, https://pandas.pydata.org/docs/
3. 'SciPy Documentation' accessed Apr 26, 2023, https://docs.scipy.org/doc//scipy/index.html
4. 'Sklearn Documentation' accessed Apr 26 2023, https://scikit-learn.org/stable/index.html
5. Seaborn Documentation' accessed Apr 26, 2023, https://seaborn.pydata.org/

6. ‘Matplotlib documentation,’ accessed Apr 26, 2023, https://matplotlib.org/stable/index.html

7. Siya Sivarajah, ’Most Important Pandas Functions-Full Tuorial,’ accessed Apr 26, 2023, https://towardsdatascience.com/pandas-full-tutorial-on-a-single-dataset-4aa43461e1e2