ES 114: Data Narrative 3

Jaidev Sanjay Khalane (22110103)

B.Tech 2022
Computer Science and Engineering
Indian Institute of Technology
Gandhinagar, India
jaidev.khalane@iitgn.ac.in

Abstract— **This report is about the analysis of the data given on the website https://archive-beta.ics.uci.edu/dataset/300/tennis+major+tournament+match+statistics about the various parameters of tennis matches in various tournaments, as given in the overview part. Various scientific questions and hypotheses were also created in order to effectively study the data. The software used is Python 3 with modules such as Pandas, Matplotlib, NumPy, Seaborn, Plotly, sklearn.**

I. Overview Of The Dataset

In this dataset, the data about various aspects of matches conducted in eight Tennis tournaments held in the year 2013 is given. These parameters contain data such as names of the players, the result of the match, first serve percentage, second serve percentage, number of aces won, number of double faults committed, number of winners won, number of unforced errors committed, number of breakpoints created, number of points attempted followed by the set results and many more parameters. On the whole, the dataset provides a very wide spectrum of parameters which were extensively analysed using the libraries of Python 3 such as Pandas, Matplotlib, NumPy, Seaborn, Plotly, sklearn, etc.

II. Scientific Questions about the Dataset

Analysis of Australia Open men's tournament 2013

Q1. Who are the top 10 players with the largest number of points? What is the **mean and variance** of the first serve percentage of player 1 and the first serve percentage of player 2? Who are the top 10 players with the largest number of first serves, that is, first serve percentage? What is the **correlation** between the first serve percentage of a player and the result of the game? What is the probability of a player winning a game given that the player has more first-serve percentage? What is the **conditional probability** of winning the game given that the first player has less first serve percentage compared to the second player? Provide a visualization for that. Considering the result of the game and the relative number of first serve percentage of the player compared to the other player **as two joint variables**. Plot these random variables in a two-dimensional space, and how can we visualise the **marginals** of these distributions with the help of scatter plots, histograms and smooth distribution curves? What can we conclude from these provided **visualisations** and what can we analyse from them?

Analysis of Australia Open women's tournament 2013

Q2. Who are the top 10 players with the largest number of points? What is the mean and variance of the number of aces of player 1 and the number of aces of player 2? Who are the top 10 players with the largest number of aces? What is the correlation between the number of aces of a player and the result of the game? What is the probability of winning a game given that the player has more first-serve percentage? What is the conditional probability of winning the game, given that the first player has a smaller number of aces compared to the second player? Provide a visualization for that. Considering the result of the game and the relative number of aces of the player compared to the other player as two joint variables, plot these joint variables in a two-dimensional space. How can we visualise the marginals of these distributions with the help of a scatter plot, histogram and smooth probability density curve? What can we conclude from this? Provide visualisations and what can we analyse from them?

Analysis of French Open Men's tournament 2013

Q3. Who are the top 10 players with the largest number of points? What is the mean and variance of the number of Double Faults committed by player 1 and the number of Double Faults committed by player 2? Who are the top 10 players with the least number of Double Faults committed? What is the correlation between the Double Faults committed by a player and the result of the game? What is the probability of winning a game given that the player has more Double Faults committed? What is the conditional probability of winning the game given that the first player has less number of Double Faults committed compared to the second player? Provide a visualization for that. Considering the result of the game and the relative number of Double Faults committed by the player compared to the other player as two variables, plot these variables in a two-

dimensional space. How can we visualise the marginals of these distributions with the help of scatter plot, histogram and smooth probability density curve? What can we conclude from this? Provide visualisations and what can we analyse from them?

Analysis of French Open Women's tournament 2013

Q4. Who are the top 10 players with the largest number of points? What is the mean and variance of the number of Break Points Created by player 1 and the number of Break Points Created by player 2? Who are the top 10 players with the greatest number of Break Points Created? What is the correlation between the Break Points Created by a player and the result of the game? What is the probability of winning a game given that the player has less Break Points Created? What is the conditional probability of winning the game given that the first player has more number of Break Points Created compared to the second player? Provide a visualization for that. Considering the result of the game and the relative number of Break Points Created by the player compared to the other player as two variables, plot these random variables in a two-dimensional space. How can we visualise the marginals of these distributions with the help of scatter plots, histograms and smooth distribution curve? What can we conclude from this? Provide visualisations. What can we analyse from them? Can we consider the ratio of breakpoints won divided by the number of breakpoints created by the player as the accuracy of the player? Who are the top 10 players with the largest accuracy in terms of breakpoints; that is, the greatest number of breakpoints won for every breakpoint created? What is the correlation between this ratio, that is, the accuracy of the player with the result of the match, and can we consider it as a valid variable to predict the result of the match?

Analysis of US Open Men's tournament 2013

Q5. Who are the top 10 players with the largest number of points? What is the mean and variance of the number of points attempted by player 1 and the number of points attempted by player 2? Who are the top 10 players with the least number of points attempted? What is the correlation between the Number of points attempted by a player and the result of the game? What is the probability of winning a game given that the player has a greater number of points attempted? What is the conditional probability of winning the game given that the first player has a smaller number of points attempted compared to the second player? Provide a visualization for that. Considering the result of the game and the relative number of points attempted by the player compared to the other player as two variables, plot these variables in a two-dimensional space. How can we visualise the marginals of these joint distributions? With the help of scatter plot, histogram and smooth probability density curve, what can we conclude from this? Provide visualisations. What can we analyse from them? Can we consider the ratio of Number of points won divided by the Number of points attempted by the player as the accuracy of the player? Who are the top 10 players with the largest accuracy in terms of Number of points; that is, most number of points won for every one of the points attempted? What is the correlation between this ratio, that is, the accuracy of the player with the result of the match, and can we consider it as a valid variable to predict the result of the match?

Analysis of US Open Women's tournament 2013

Q6. Who are the top 10 players in the tournament in terms of maximum number of points scored? If we consider player 1 and player 2 as two different entities, that is if we study these two parameters separately, what is the probability of player 1 to win the set one and the probability of player one to win the set 2? Speaking about the conditional probabilities, what is the conditional probability that the player one wins the set 2 given that the player has won set 1? Similarly, what is the probability that player 1 has won the set 2 given that he has lost the set one? What is the conditional probability that the player one loses the second set given that he has won the first set and given that he has lost the first set? Taking this further, what can we analyse for player 2 with the same parameters? What is the total probability of a player to win the second set given that he has won the first set, to win the second set given that he has lost the first set, to lose the second set given that he has one and lost the first set? How can we visualise this and what can we conclude from this visualisation?

Analysis of Wimbledon Open Men's tournament 2013

Q7. Who are the top 10 players in the tournament in terms of number of points scored and the matches won? Who are the top 10 players in terms of maximum number of aces scored, in terms of maximum number of first serve and second serve wins, largest number of breakpoints created, largest number of points attempted, number of unforced errors committed and number of winners won by the players? Which of these parameters has the largest amount of accuracy in terms of predicting the top 10 players of the game? Provide the value of accuracy in percentage and also the visualisation of the most accurate parameters that can be used to predict the top 10 players of a tournament?

Analysis of Wimbledon Open Women's tournament 2013

Q8. Who are the top 10 players in the tournament in terms of maximum number of matches won? In terms of the parameters that are given in the data frame, how can we visualise the correlation between all the parameters; that is, find out which parameter is most related to which another parameter? Using this, how can we find the parameters that are highly correlated in both the positive and negative sense to the result of the game? How can we visualise this? Can we use the top parameters in terms of correlation with the result to predict the result of the game; that is, can we use the parameters and take its first degree polynomial function and find the approximate value of the predicted result by minimising the main square error of the approximated value and the actual value of the result? What is the accuracy score of that and can we use it in the real-life situation to predict the outcomes of the match? Which parameters are used in this predictor that is which parameters have the maximum accuracy in terms of this prediction

III. Details of Libraries used

The libraries that are used include matplotlib, NumPy, Pandas, Seabor, Plotly, and sklearn .

- matplotlib.pyplot: This library was used for applications such as plotting in the form of a line graph, bar graph, histogram, pie chart etc. and other applications such as legend, the label of axes, title, grid, etc.

- pandas: pandas were used for reading the data from .csv files, creating data frames, iterating through them, and deleting the wrongly formatted data points using the function df.drop. The kernel density estimation KDE was also used to plot smooth probability distributions.

- NumPy: I used the NumPy library for creating NumPy arrays for faster computing and storing data, for calculating mean, for getting the unique elements out of the NumPy arrays, and by enabling the counts, I also got the number of times the unique element has occurred in the array and its index. It was also used to calculate the mean, standard deviation and variance of the data.

- Seaborn: This library was used to plot the various kind of plots such as heatmaps, jointplots, plots with kde, histograms to represent marginals on the jointplots, etc.

- Plotly: This was used to generate interactive Bar Graph plots for a more extensive study of the plots.

- SciKitLearn (sklearn): This library was used to provide the metrics of accuracy of the predicted values in dataset.

IV. Answers to the Questions and Summaries Drawn

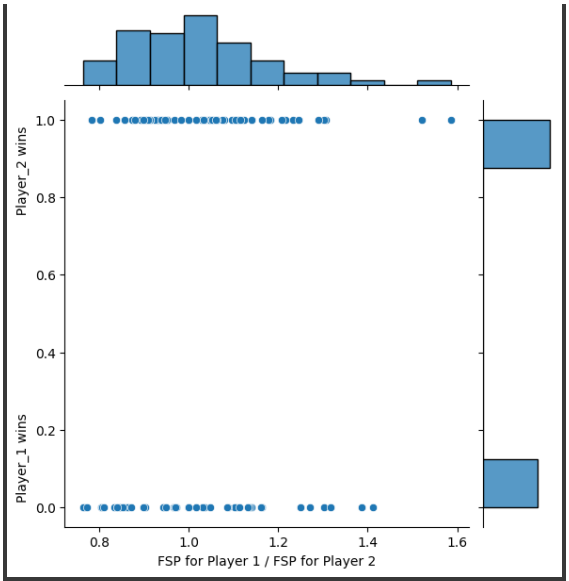
Analysis of Australia Open men's tournament 2013

Q1. Who are the top 10 players with the largest number of points? What is the **mean and variance** of the first serve percentage of player 1 and the first serve percentage of player 2? Who are the top 10 players with the largest number of first serves, that is, first serve percentage? What is the **correlation** between the first serve percentage of a player and the result of the game? What is the probability of a player winning a game given that the player has more first-serve percentage? What is the **conditional probability** of winning the game given that the first player has less first serve percentage compared to the second player? Provide a visualization for that. Considering the result of the game and the relative number of first serve percentage of the player compared to the other player **as two joint variables**. Plot these random variables in a two-dimensional space, and how can we visualise the **marginals** of these distributions with the help of scatter plots, histograms and smooth distribution curves? What can we conclude from these provided **visualisations** and what can we analyse from them?

Answer:
Procedure: First of all, the data set which was in the form of.csv file was converted into a Pandas data frame using the Pandas function to read csv files. Then, an empty dictionary was created in which the players' names were added against the number of points won by the players by iterating through the data frame, which was later converted into a list of tuples having the data of the players' names against the total points won. Then, the list was sorted, and the first 10 elements were taken, which represent the top 10 players with the largest number of points. Then, the data frame was iterated through the columns of FSP.1 and FSP.2 to find wrongly formatted rows and the wrongly formatted rows when dropped using the Pandas function of df.drop. The remaining rows were converted into integer format and then the function of Pandas named as

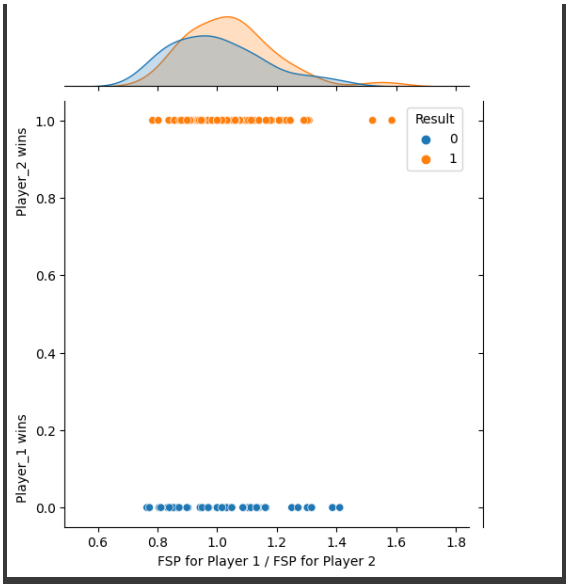
.corr was used to calculate the correlation between the result of the game and the first serve percentage. The functions of numpy library of mean and variance were also used to give the value of mean and variance of the total first serve percentage. A method of dictionary, similar to the first part of the question was used to find the players with the largest number of first serve percentage. Then I created variables to store the values of the number of situations in which the player won the game despite having a lower first serve percentage and the total number of situations in which the first player had more first serve percentage and less second serve percentage. Then these values were used to find the probability of winning the game given that the player has more first serve percentage than the other one and the probability of winning the game given that the player has less first serve percentage compared to the other player. The libraries such as matplotlib and seaborn were used to plot the visualisation. The pie chart was plotted using the matplotlib library and the visualisations of joint variable were plotted using seaborn library; using the function of joint plot and parameters such as regression, Kernel density estimation, etc. also to obtain the Marginals.

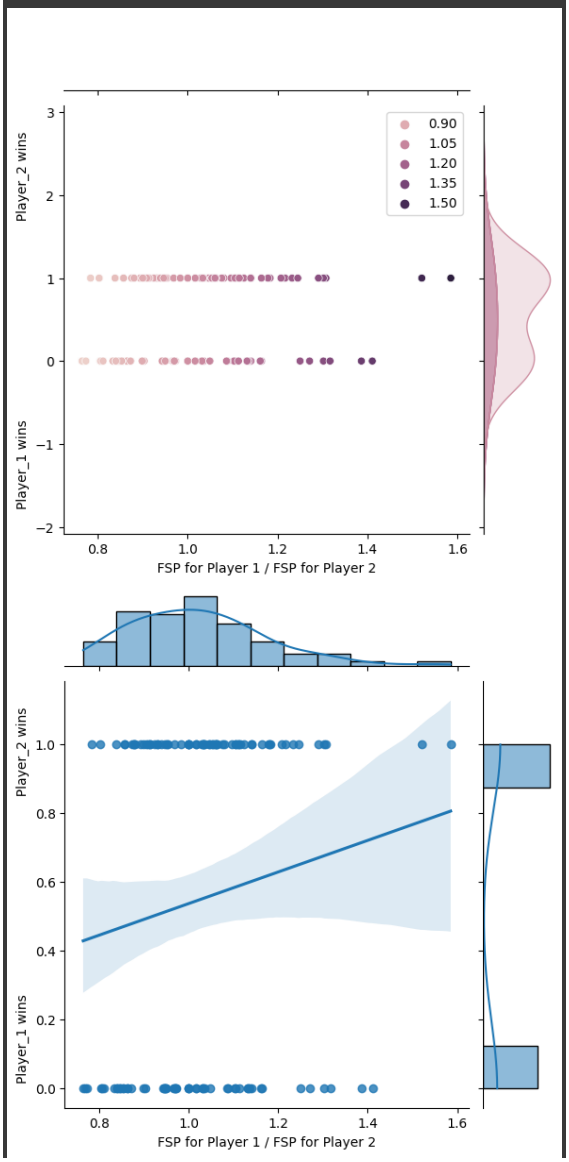
Results and observations:

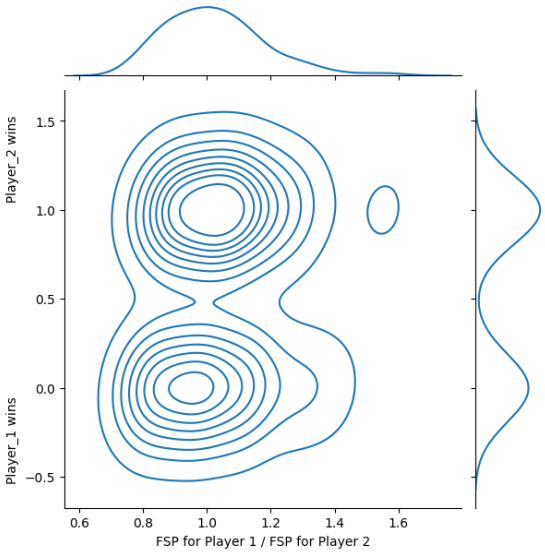


Joint plot of ratio of FSP of player 1/ player 2 with the result of the match.

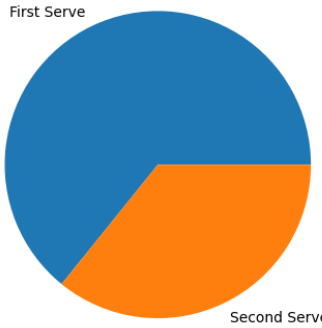
The histograms on the axes are the representation of the marginal distributions of these variables.







Probability of winning given the number of serve of the player




```
Top 10 players : [('Stanislas Wawrinka', 17), ('Rafael Nadal', 17), ('Tomas Berdych', 16), ('Roger Federer', 15), ('Novak Djokovic', 14), ('David Ferrer', 13), ('Andy Murray', 13), ('Grigor Dimitrov', 13), ('Florian Mayer', 10), ('Stephane Robert', 10)]

Probability of winning given there is a first serve: 0.6417910447761194

Probability of winning given there is a second serve: 0.3582089552238806

CORRELATION between the result of the match and the First Serve Percentage ratio: 0.04567445322956493

Top 10 players with the largest number of First serves:
[('S.Lisicki', 441), ('M.Bartoli', 431), ('A.Radwanska', 424), ('K.Flipkens', 358), ('S.Stephens', 342), ('N.Li', 342), ('P.Kvitova', 272), ('S.Williams', 270), ('C.Suarez Navarro', 259), ('M.Puig', 250)]

MEAN number of First serves: 63.85245901639344
VARIANCE in number of First serves:51.43724805159903
```

As we can observe from the data that the mean and the variance of the first serve percentage is as obtained. We can see that there is moderately low correlation between the first serve percentage and the result of the match. We can also see that the probability of winning the match given that the first serve and the second serve is not very much different which also implies that the correlation between these two parameters is low. This can also be visually confirmed using the plots that we have generated, we can see that the height of the histograms of the marginals of the result of the match from the joint distribution is almost equal.

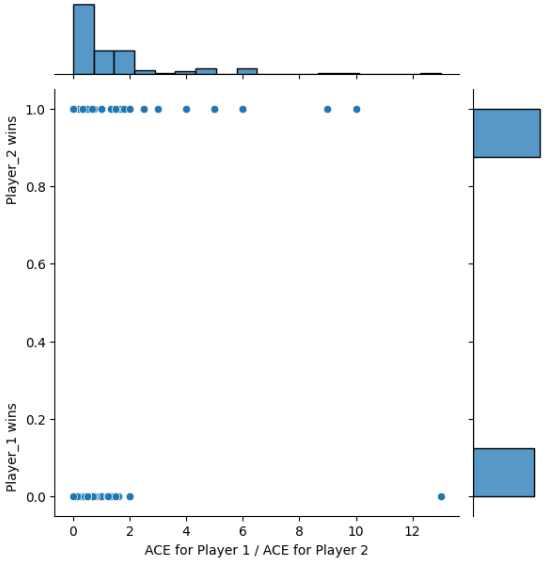
Analysis of Australia Open women's tournament 2013

Q2. Who are the top 10 players with the largest number of points? What is the mean and variance of the number of aces of player 1 and the number of aces of player 2? Who are the top 10 players with the largest number of aces? What is the correlation between the number of aces of a player and the result of the game? What is the probability of winning a game given that the player has more first-serve percentage? What is the conditional probability of winning the game, given that the first player has a smaller number of aces compared to the second player? Provide a visualization for that. Considering the result of the game and the relative number of aces of the player compared to the other player as two joint variables, plot these joint variables in a two-dimensional space. How can we visualise the marginals of these distributions with the help of a scatter plot, histogram and smooth probability density curve? What can we conclude from this? Provide visualisations and what can we analyse from them?

Answer:
Procedure: First of all, the data set, which was in the form of a .csv file, was converted into a Pandas data frame using the Pandas function to read csv files. Then an empty dictionary was created in which the players' names were added against the number of points won by the players by iterating through the data frame, which was later converted into a list of tuples having the values of the players' names against the total points won. Then the list was sorted and the first ten elements were taken, which represent the top 10 players with the largest number

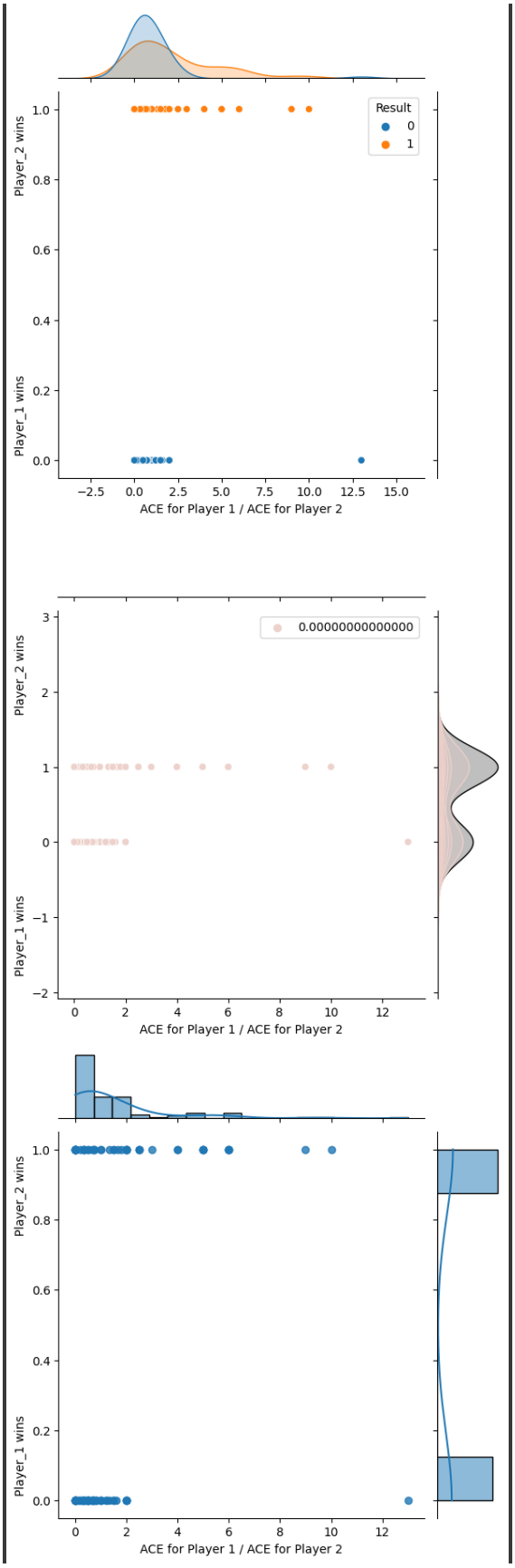
of points. Then the data frame was iterated through the columns of ACE.1 and ACE.2 to find wrongly formatted rows, and the wrongly formatted rows when dropped using the Pandas function of .drop(). The remaining rows were converted into integer format and then the function of Pandas named .corr was used to calculate the correlation between the result of the game and the number of aces. The functions of the numpy library of mean and variance were also used to give the value of mean and variance of the number of aces. A method using a dictionary, similar to the first part of the question, was used to find the players with the largest number of aces scored. Then I created variables to store the values of the number of situations in which the player won the game despite having a lower number of aces and the total number of situations in which the first player had more number of aces but still lost. Then these values were used to find the probability of winning the game given that the player has more number of aces than the other one and the probability of winning the game given that the player has less number of aces compared to the other player. The libraries such as matplotlib and Seaborn were used to plot the visualisation. The pie chart was plotted using the matplotlib library, and the visualisations of joint variables were plotted using the seaborn library using the function of joint plot and parameters such as regression, Kernel density estimation, etc. also to obtain the plot of marginals.

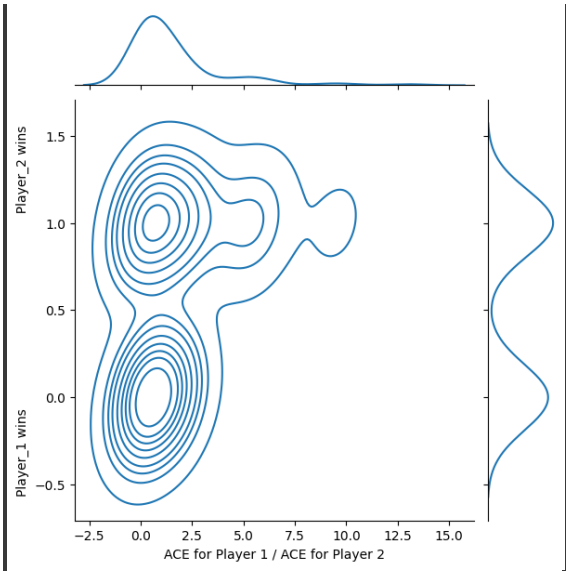
Results and Observations:



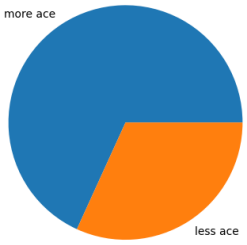
Joint plot of ratio of Number of Aces of player 1/ player 2 with the result of the match.

The histograms on the axes are the representation of the marginal distributions of these variables.





Probability of winning given the comparison between the aces of both the players




```
Top 10 players : [('Na Li', 14.0),
('Dominika Cibulkova', 12.0),
('Eugenie Bouchard', 10.0),
('Agnieszka Radwanska', 10.0),
('Ana Ivanovic', 9.0), ('Simona
Halep', 8.0), ('Jelena Jankovic',
7.0), ('Kurumi Nara', 4.0),
('Zarina Diyas', 4.0), ('Carla
Suarez Navarro', 4.0)]


Probability of winning given there
are more aces: 0.6818181818181818
probability of winning given there
are less aces: 0.3181818181818182

CORRELATION between the result of
the match and the ratio of the
number of Aces of player 1 and
player 2: 0.09073954090058929

Top 10 players with the largest
number of aces:
[('Serena Williams', 39.0),
('Eugenie Bouchard', 19.0),
('Daniela Hantuchova', 17.0),
('Karolina Pliskova', 16.0),
('Madison Keys', 14.0), ('Samantha
Stosur', 13.0), ('Ana Ivanovic',
13.0), ('Vesna Dolonc', 9.0),
('Virginie Razzano', 9.0),
('Heather Watson', 8.0)]

MEAN number of aces:
2.9016393442622954
VARIANCE in the number of aces:
7.7368986831496915
```

As we can observe from the data that the mean and the variance of the number of aces is as obtained, we can see that there is a moderately low correlation between the number of aces and the result of the match. We can also see that the probability of winning the match given that the more number of aces and the less number of aces slightly different which also implies that the correlation between these two parameters is moderate. This can also be visually confirmed using the plots that we have generated. We can see that the height of the histograms of the marginals of the result of the match from the joint distribution is almost equal.

Analysis of French Open Men's tournament 2013

Q3. Who are the top 10 players with the largest number of points? What is the mean and variance of the number of Double Faults committed by player 1 and the number of Double Faults committed by player 2? Who are the top 10 players with the least number of Double Faults committed? What is the correlation between the Double Faults committed by a player and the result of the game? What is the probability of winning a game given that the player has more Double Faults committed? What is the conditional probability of winning the game given that the first player has less number of Double Faults committed compared to the second player? Provide a visualization for that. Considering the result of the game and the relative number of Double Faults committed by the player compared to the other player as two variables, plot these variables in a two-dimensional space. How can we visualise the marginals of these distributions with the help of scatter plot, histogram and smooth probability density curve? What can we conclude from this? Provide visualisations and what can we analyse from them?

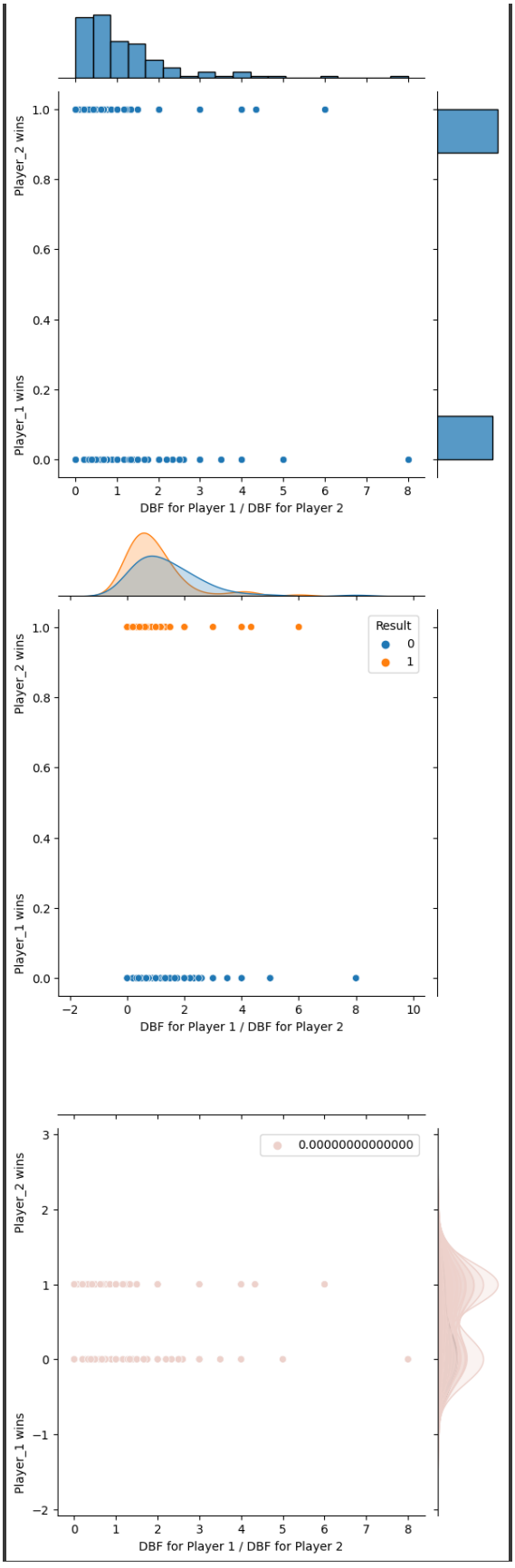
Answer:
Procedure: First of all, the data set, which was in the form of a .csv file, was converted into a Pandas data frame using the Pandas function to read csv files. Then an empty dictionary was created in which the players' names were added against the number of points won by the players by iterating through the data frame, which was later converted into a list of tuples having the values of the players' names against the total points won. Then the list was sorted

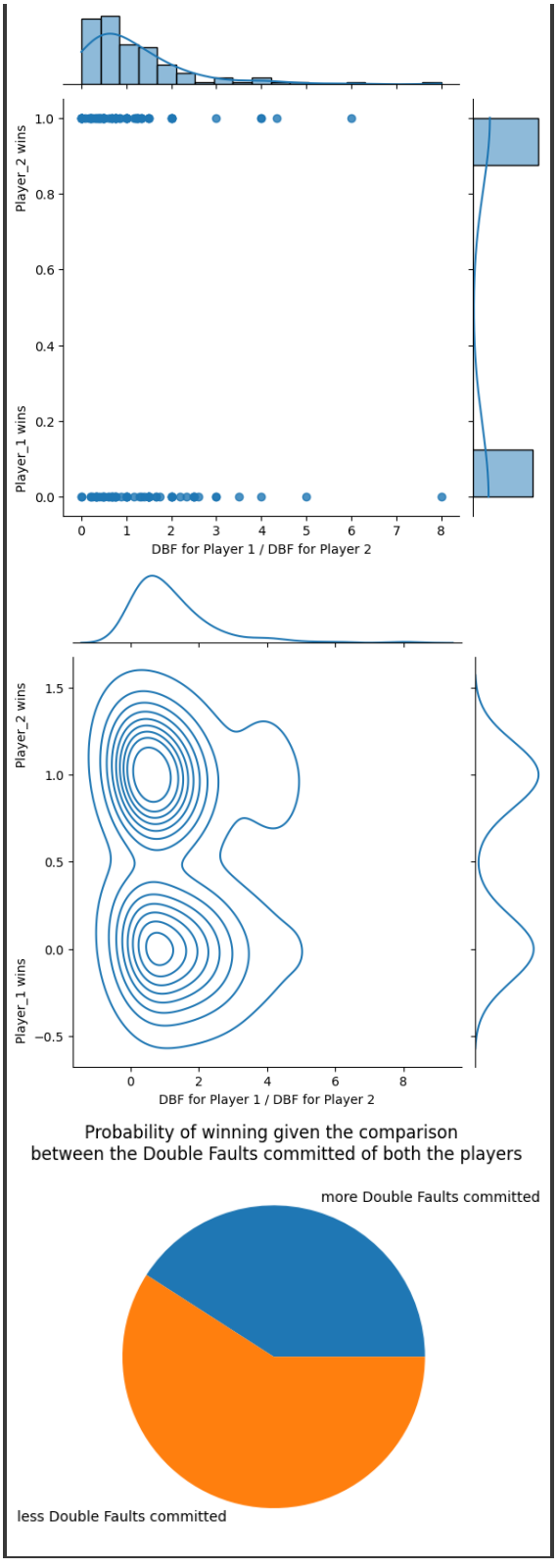
and the first ten elements were taken, which represent the top 10 players with the largest number of points. Then the data frame was iterated through the columns of DBF.1 and DBF.2 to find wrongly formatted rows and the wrongly formatted rows when dropped using the Pandas function of .drop(). The remaining rows were converted into integer format and then the function of Pandas named .corr was used to calculate the correlation between the result of the game and the number of Double Faults committed. The functions of the numpy library of mean and variance were also used to give the value of mean and variance of the number of Double Faults committed. A method using a dictionary, similar to the first part of the question, was used to find the players with the largest number of Double Faults committed. Then I created variables to store the values of the number of situations in which the player won the game despite having a higher number of Double Faults committed and the total number of situations in which the first player had a lesser number of Double Faults committed but still lost. Then these values were used to find the probability of winning the game given that the player has lesser number of Double Faults committed than the other one and the probability of winning the game given that the player has more Double Faults committed compared to the other player. The libraries such as matplotlib and Seaborn were used to plot the visualisation. The pie chart was plotted using the matplotlib library, and the visualisations of joint variables were plotted using the seaborn library using the function of joint plot and parameters such as regression, Kernel density estimation, etc. also to obtain the plot of marginals.

Results and observations:

Joint plot of ratio of Number of DBF of player 1/ player 2 with the result of the match.

The histograms on the axes are the representation of the marginal distributions of these variables.






```
Top 10 Players: [('Rafael Nadal', 21), ('David Ferrer', 18), ('Novak Djokovic', 17), ('Jo-Wilfried Tsonga', 15), ('Roger Federer', 12), ('Tommy Robredo', 12), ('Stanislas Wawrinka', 12), ('Gilles Simon', 11), ('Nicolas Almagro', 11), ('Richard Gasquet', 11)]

Probability of winning given there are more Double Faults committed: 0.4090909090909091
Probability of winning given there are less Double Faults committed: 0.5909090909090909

CORRELATION between the ratio of number of double faults committed and the result of the match: -0.14630339119189076


10 players with least Double Faults: [('Sara Errani', 0.0), ('Patricia Mayr-Achleitner', 1.0), ('Anabel Medina Garrigues', 1.0), ('Shahar Peer', 1.0), ('Shuai Zhang', 1.0), ('Sorana Cirstea', 1.0), ('Ying-Ying Duan', 1.0), ('Tsvetana Pironkova', 2.0), ('Silvia Soler-Espinosa', 2.0), ('Donna Vekic', 2.0)]

MEAN of the number of DBF: 3.9642857142857144
VARIANCE in the number of DBF: 7.458049886621315
```

As we can observe from the data that the mean and the variance of the number of Double Faults committed is as obtained, we can see that there is a moderate negative correlation between the number of Double Faults committed and the result of the match. We can also see that the probability of winning the match given that less number of Double Faults committed and the more number of Double Faults committed is slightly different which also implies that the correlation between these two parameters is moderate. This can also be visually confirmed using the plots that we have generated we can see that the height of the histograms of the marginals of the result of the match from the joint distribution is slightly different. Moreover, they are negatively correlated which can also be seen from the plots.

Analysis of French Open Women's tournament 2013

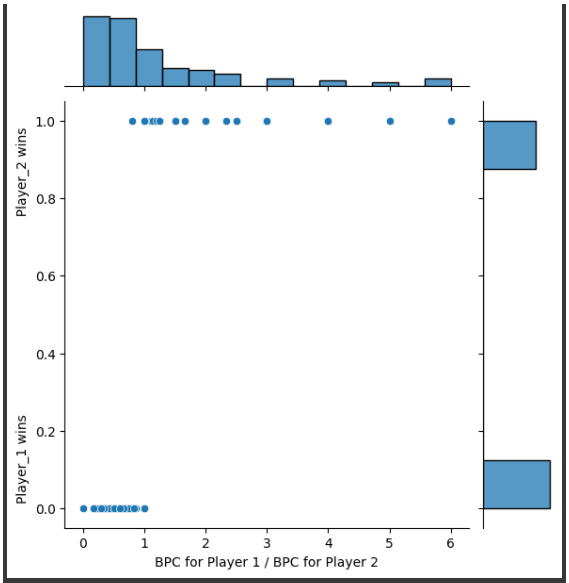
Q4. Who are the top 10 players with the largest number of points? What is the mean and variance of the number of Break Points Created by player 1 and the number of Break Points Created by player 2? Who are the top 10 players with the greatest number

of Break Points Created? What is the correlation between the Break Points Created by a player and the result of the game? What is the probability of winning a game given that the player has less Break Points Created? What is the conditional probability of winning the game given that the first player has more number of Break Points Created compared to the second player? Provide a visualization for that. Considering the result of the game and the relative number of Break Points Created by the player compared to the other player as two variables, plot these random variables in a two-dimensional space. How can we visualise the marginals of these distributions with the help of scatter plots, histograms and smooth distribution curve? What can we conclude from this? Provide visualisations. What can we analyse from them? Can we consider the ratio of breakpoints won divided by the number of breakpoints created by the player as the accuracy of the player? Who are the top 10 players with the largest accuracy in terms of breakpoints; that is, the greatest number of breakpoints won for every breakpoint created? What is the correlation between this ratio, that is, the accuracy of the player with the result of the match, and can we consider it as a valid variable to predict the result of the match?

Answer:
Procedure: First of all, the data set, which was in the form of a .csv file, was converted into a Pandas data frame using the Pandas function to read csv files. Then an empty dictionary was created in which the players' names were added against the number of points won by the players by iterating through the data frame, which was later converted into a list of tuples having the values of the players' names against the total points won. Then the list was sorted and the first ten elements were taken, which represent the top 10 players with the largest number of points. Then the data frame was iterated through the columns of DBF.1 and DBF.2 to find wrongly formatted rows and the wrongly formatted rows when dropped using the Pandas function of .drop(). The remaining rows were converted into integer format, and then the function of Pandas named .corr was used to calculate the correlation between the result of the game and the number of Break Points Created. The functions of the numpy library of mean and variance were also used to give the value of mean and variance of the number of Break Points Created. A method using a dictionary, similar to the first part of the question, was used to find the players with the largest number of Break Points Created. Then, I created variables to store the values of the number of situations in which the player won the game despite having a higher number of Break Points Created and the total number of situations in which the first player had a lesser number of Break Points Created but still lost. Then these values were used to find the probability of winning the game given that the player has lesser number of Break Points Created than the other one and the probability of winning the game given that the player has more Break Points Created compared to the other player. The libraries such as matplotlib and Seaborn were used to plot the visualisation. The pie chart was plotted using the matplotlib library, and the visualisations of joint variables were plotted using the seaborn library using the function of joint plot and parameters such as regression, Kernel density estimation, etc. also to obtain the plot of marginals.

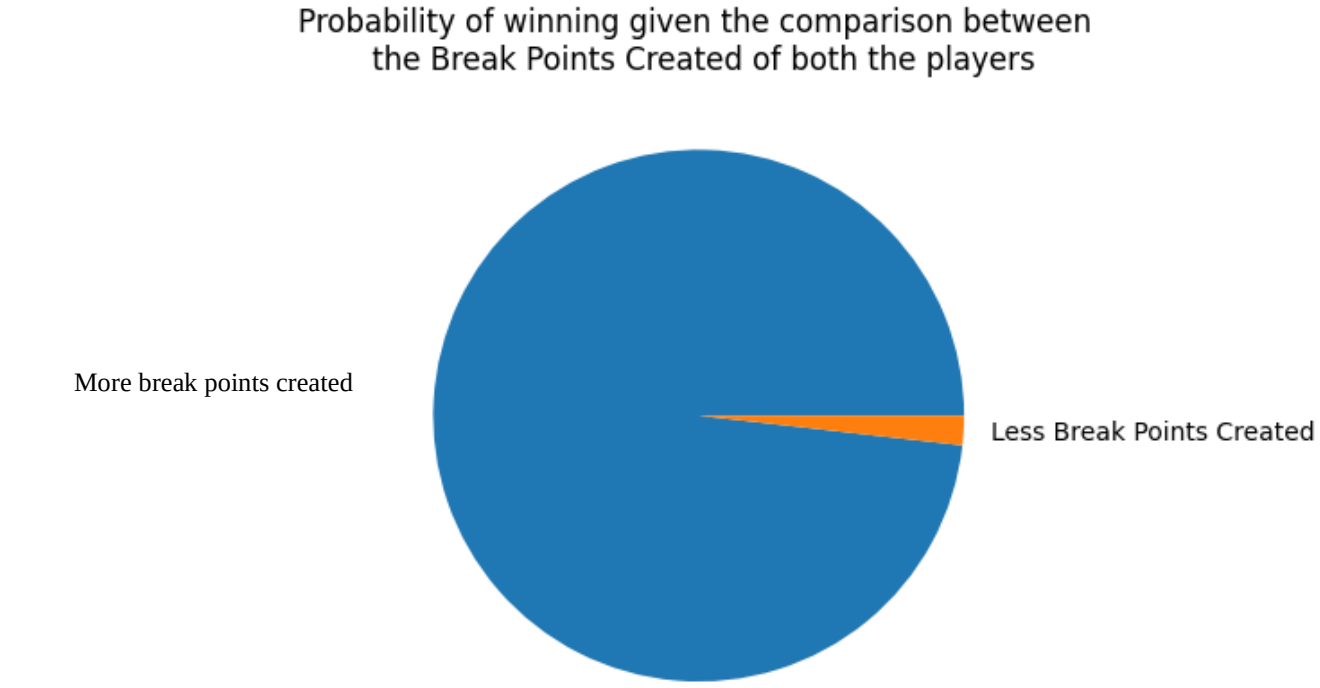
To analyse the accuracy of a player, if we consider the ratio of the number of break points won to the number of breakpoints created by the player, then, first of all, the rows in the data Frame were iterated and a dictionary was created in which the name of the player was added against the total ratio of the number of break-points won to number of breakpoints created. Then the list of dictionary items was sorted and the first 10 elements were printed. To get the correlation between the result of the match and the ratio of breakpoints won/ created the Pandas function of .corr was used.

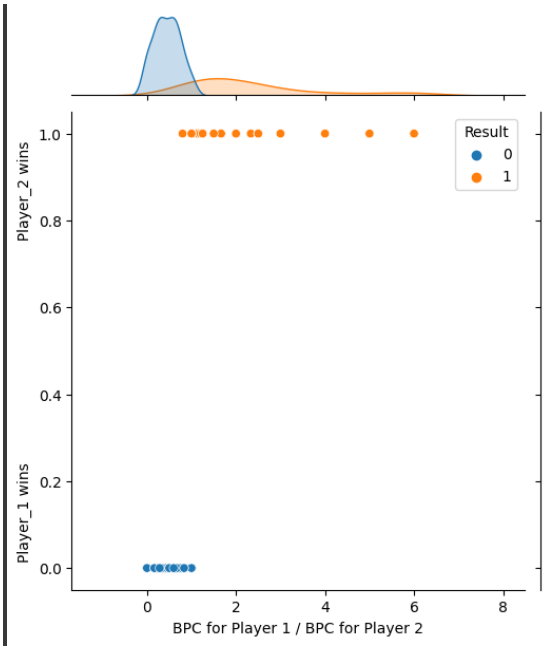
Results and observations:

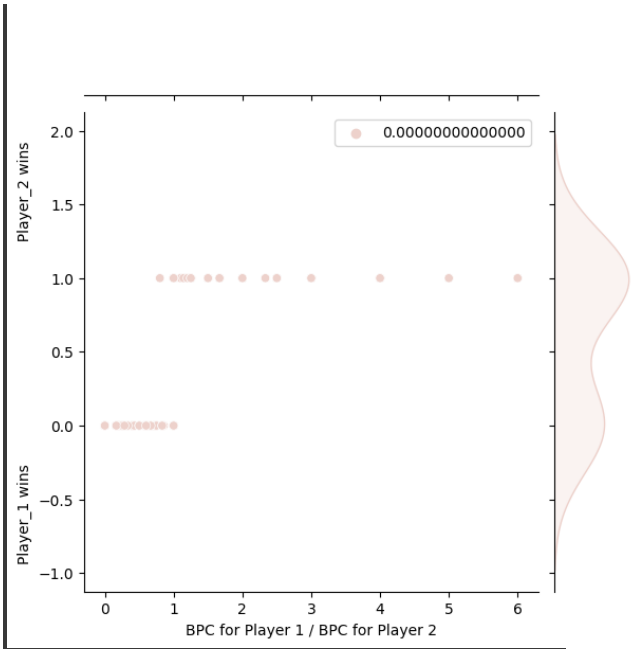


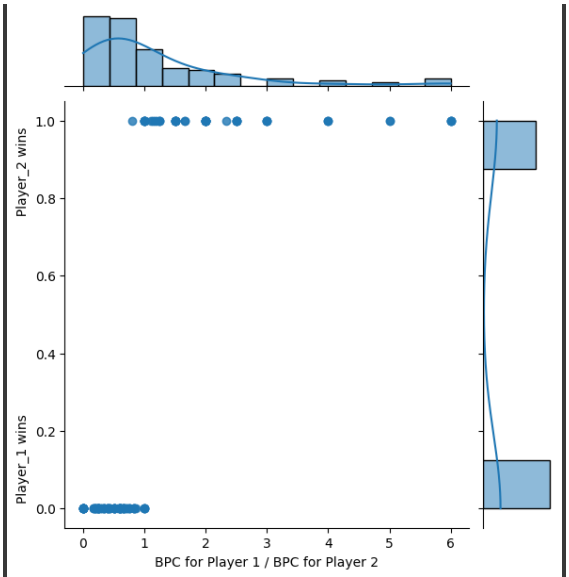
Joint plot of ratio of Number of BPC of player 1/ player 2 with the result of the match.

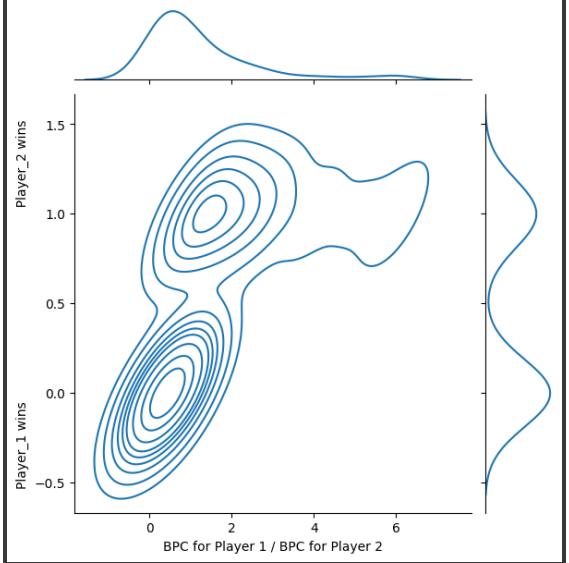
The histograms on the axes are the representation of the marginal distributions of these variables.


Probability of winning given the comparison between
the Break Points Created of both the players
More break points created
Less Break Points Created


Result
0
1
Player_2 wins
Player_1 wins
BPC for Player 1 / BPC for Player 2


0.000000000000000
Player_2 wins
Player_1 wins
BPC for Player 1 / BPC for Player 2






```
Top 10 players: [('Serena Williams', 14), ('Maria Sharapova', 12), ('Victoria Azarenka', 11), ('Sara Errani', 10), ('Jelena Jankovic', 9), ('Svetlana Kuznetsova', 9), ('Agnieszka Radwanska', 8), ('Maria Kirilenko', 7), ('Carla Suarez Navarro', 7), ('Angelique Kerber', 7)]

Probability of winning given there is a more Break Points Created: 0.9821428571428571
Probability of winning given there is a less Break Points Created: 0.017857142857142856

CORRELATION between the number of BPC and Result: 0.6465944606739645
Top 10 players with most BPC: [('Victoria Azarenka', 36), ('Serena Williams', 33), ('Maria Sharapova', 32), ('Sara Errani', 31), ('Jelena Jankovic', 24), ('Agnieszka Radwanska', 24), ('Svetlana Kuznetsova', 24), ('Carla Suarez Navarro', 23), ('Francesca Schiavone', 21), ('Bethanie Mattek-Sands', 20)]

Players with most number of BPW for every BPC: [('Karin Knapp', 1.0), ('Julia Glushko', 1.0), ('Pauline Parmentier', 1.0), ('Monica Niculescu', 1.0), ('Grace Min', 1.4285714285714286), ('Tatjana Maria', 1.5), ('Arantxa Rus', 1.5), ('Venus Williams', 1.5), ('Nadia Petrova', 1.6), ('Lucie Hradecka', 1.6666666666666667)]
```

As we can observe from the data that the mean and the variance of the number of Break Points Created is as obtained. We can see that there is a high correlation between the number of Break Points Created and the result of the match. We can also see that the probability of winning the match given that the less number of Break Points Created and the more number of Break Points Created is drastically different which also implies that the correlation between these two parameters is high. This can also be visually confirmed using the plots that we have generated. We can see that the height of the histograms of the marginals of the result of the match from the joint distribution is different.

The top 10 players with the highest accuracy are as given above. There are also players who have accuracy of more than 100% ; that is they have won the breakpoints created by the other player also.

Analysis of US Open Men's tournament 2013

Q5. Who are the top 10 players with the largest number of points? What is the mean and variance of the number of points attempted by player 1 and the number of points attempted by player 2? Who are the top 10 players with the least number of points attempted? What is the correlation between the Number of points attempted by a player and the result of the game? What is the probability of winning a game given that the player has a greater number of points attempted? What is the conditional probability of winning the game given that the first player has a smaller number of points attempted compared to the second player? Provide a visualization for that. Considering the result of the game and the relative number of points attempted by the player compared to the other player as two variables, plot these variables in a two-dimensional space. How can we visualise the marginals of these

joint distributions? With the help of scatter plot, histogram and smooth probability density curve, what can we conclude from this? Provide visualisations. What can we analyse from them? Can we consider the ratio of Number of points won divided by the Number of points attempted by the player as the accuracy of the player? Who are the top 10 players with the largest accuracy in terms of Number of points; that is, most number of points won for every one of the points attempted? What is the correlation between this ratio, that is, the accuracy of the player with the result of the match, and can we consider it as a valid variable to predict the result of the match?

Answer:
Procedure: First of all, the data set, which was in the form of a .csv file, was converted into a Pandas data frame using the Pandas function to read .csv files. Then an empty dictionary was created in which the players' names were added against the number of points won by the players by iterating through the data frame, which was later converted into a list of tuples having the values of the players' names against the total points won. Then the list was sorted and the first ten elements were taken, which represent the top 10 players with the largest number of points. Then the data frame was iterated through the columns of NPA.1 and NPA.2 to find wrongly formatted rows and the wrongly formatted rows when dropped using the Pandas function of .drop(). The remaining rows were converted into integer format and then the function of Pandas named .corr was used to calculate the correlation between the result of the game and the number of points attempted . The functions of the numpy library of mean and variance were also used to give the value of mean and variance of the number of points attempted. A method using a dictionary, similar to the first part of the question, was used to find the players with the largest number of points attempted. Then I created variables to store the values of the number of situations in which the player won the game despite having a HIGHER number of points attempted and the total number of situations in which the first player had a lesser number of points attempted but still lost. Then these values were used to find the probability of winning the game given that the player has a lesser number of points attempted than the other one and the probability of winning the game given that the player has more points attempted compared to the other player. The libraries such as matplotlib and Seaborn were used to plot the visualisation. The pie chart was plotted using the matplotlib library, and the visualisations of joint variables were plotted using the seaborn library using the function of joint plot and parameters such as regression, Kernel density estimation, etc. also to obtain the plot of marginals.

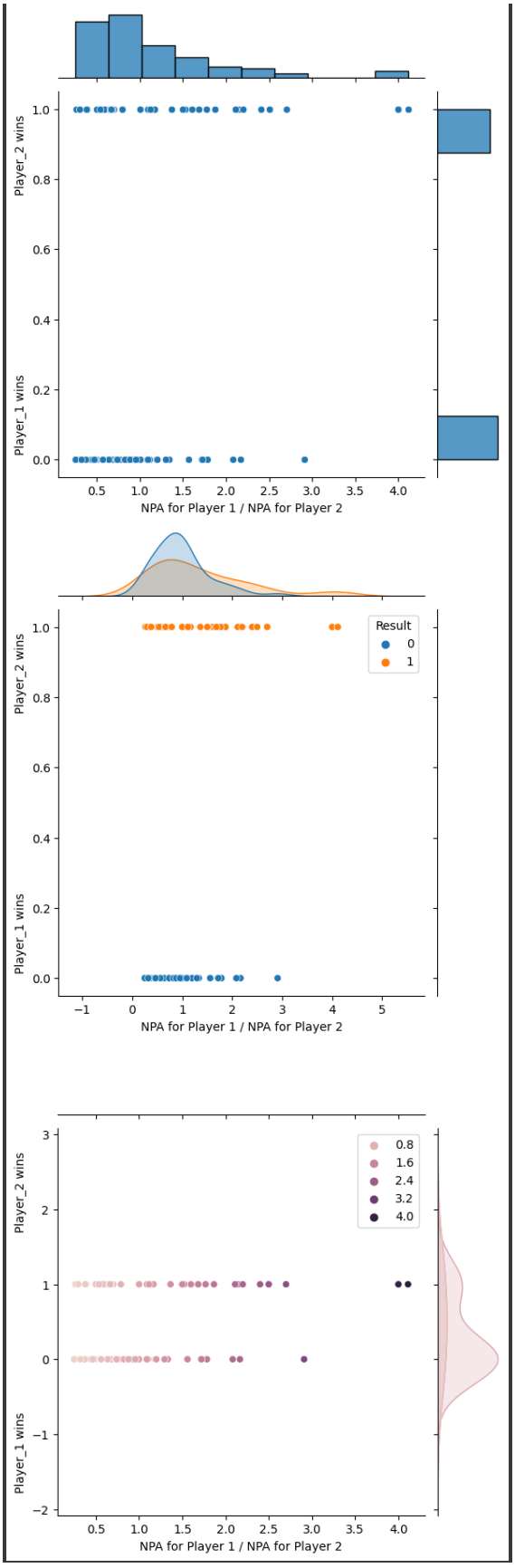
To analyse the accuracy of a player if we consider the ratio of the Number of points won to the Number of points attempted by the player, then, first of all, the rows in the data frame were iterated and a dictionary was created in which the name of the player was added against the total ratio of the Number of points won to the Number of points attempted. Then, the list of dictionary items was sorted, and the first 10 elements were printed. To get the correlation between the result of the match and the ratio of Number of points won to the Number of points attempted, the Pandas function of .corr was used.

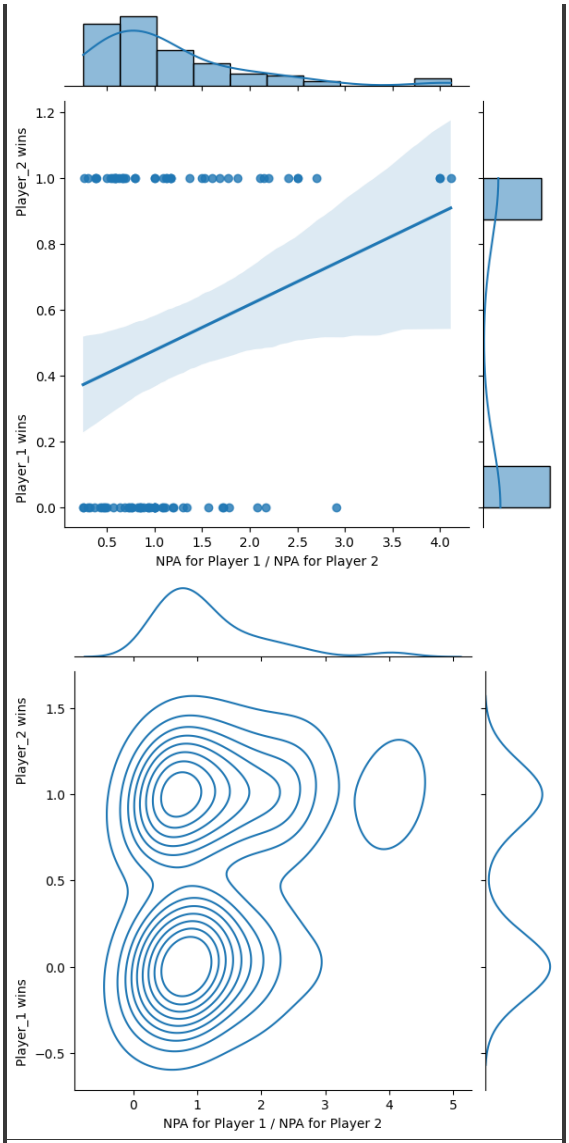
Results and observations:
As we can observe that the top 10 players in terms of points are as given, we can also see that there is not much difference in the probability of winning the game given that the player has more accuracy, but, the total number of points won highly correlated to the result of the match as the player with larger number of points won will definitely win the match. If we see the correlation between the result of the match and the accuracy of the player then these two parameters are moderately correlated from the value that we get.

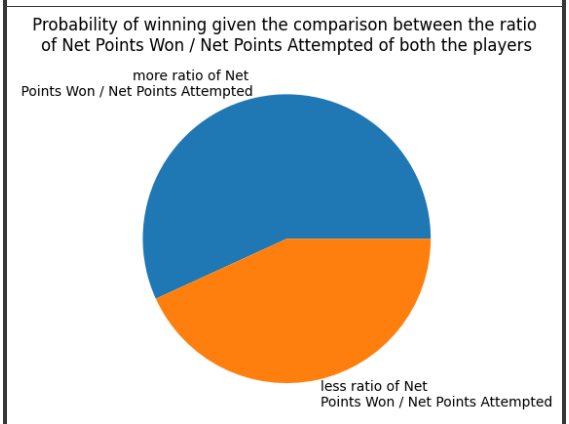
```
Top 10 players: [('Rafael Nadal', 21), ('Stanislas Wawrinka', 17), ('Novak Djokovic', 16), ('Richard Gasquet', 15), ('David Ferrer', 14), ('Mikhail Youzhny', 13), ('Tommy Robredo', 12), ('Andy Murray', 12), ('Milos Raonic', 11), ('Lleyton Hewitt', 11)]
```

Joint plot of the ratio of Number of Points attempted of player 1/ player 2 with the result of the match.

The histograms on the axes are the representation of the marginal distributions of these variables.








```
Probability of winning given there
is a more ratio of Net Points Won /
Net        Points        Attempted:
0.5681818181818182
Probability of winning given there
is a less ratio of Net Points Won /
Net        Points        Attempted:
0.4318181818181818
```

```
CORRELATION between the NPA/NPW by player 1 and result: -0.2886009804726583
CORRELATION between the NPA/NPW by player 2 and result: 0.545943294638453

Top players with most NPA/NPW:
[('Michael Russell', 1.3333333333333333), ('Richard Gasquet', 9.195148489503328)]
MEANS of the number of points attempted: 19.0625
VARIANCE in the number of points attempted: 72.24883780991738
```

Analysis of US Open Women's tournament 2013

Q6. Who are the top 10 players in the tournament in terms of maximum number of points scored? If we consider player 1 and player 2 as two different entities, that is if we study these two parameters separately, what is the probability of player 1 to win the set one and the probability of player one to win the set 2? Speaking about the conditional probabilities, what is the conditional probability that the player one wins the set 2 given that the player has won set 1? Similarly, what is the probability that player 1 has won the set 2 given that he has lost the set one? What is the conditional probability that the player one loses the second set given that he has won the first set and given that he has lost the first set? Taking this further, what can we analyse for player 2 with the same parameters? What is the total probability of a player to win the second set given that he has won the first set, to win the second set given that he has lost the first set, to lose the second set given that he has one and lost the first set? How can we visualise this and what can we conclude from this visualisation?

Answer:

Procedure: First of all, the data set was opened as a data frame using Pandas library. Then, an empty dictionary was created in which the names of the players were added against the final number of points won by the players. Then, the list containing the dictionary items was sorted in descending order and the first 10 elements were picked up. These 10 players are the top 10 players with the highest points. Then, various Pandas series were created which contain the number of points won by the players 1 and 2 after set one and set two. Then, some variables were created which would store the number of instances in which the probability and conditional probability of the events as mentioned in the question was satisfied. Then all the series were iterated and the condition for a player to win was that in a particular set the player should score more points in that particular set compared to the other player. With this condition in mind, the variables were incremented if they satisfied the conditions. Then the probability of player 1 to win the game was found out by some all the events in which the player 1 had more points than the player 2 divided by the total number of events. Then, coming to the conditional probability of player one winning the game 2 given that he had one the game 1 was found out by dividing the total number of instances in which the player 1 had won the game 1 and the game 2 divided by the total number of instances in which the player 1 had won the game 1. Similarly, all the conditional probabilities were calculated for player 1 and using the similar steps, that was calculated for player 2 as well. Then, the mean of player 1 and player 2 was calculated to get the total probability; that is, the weighted sum of the probability of player 1 and player 2 to win game 2 given that they have won game one was calculated with equal weights. Then this value for all the other parameters also was formed into a data frame which was then plotted as a heat map using seaborn library.

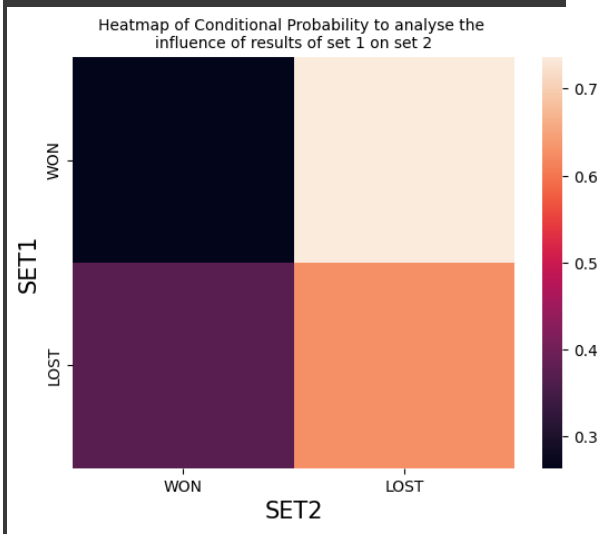
Results and observations:

```
Top 10 players: [('V Azarenka', 13), ('S Williams', 12), ('N Li', 10), ('F Pennetta', 8), ('C Suarez Navarro', 8), ('A Ivanovic', 7), ('A Kerber', 7), ('D Hantuchova', 6), ('S Stephens', 6), ('A Riske', 5)]
Prob of player 1 to win game 1: 0.35526315789473684
Prob of player 1 to win game 2: 0.4605263157894737
Prob of player 1 to win game 2 given that game 1 won: 0.1111111111111111
Prob of player 1 to lose game 2 given that game 1 won: 0.8888888888888888
Prob of player 1 to win game 2 given that game 1 was lost: 0.2857142857142857
Prob of player 1 to lose game 2 given that game 1 was lost: 0.7142857142857143
```

```
Prob of player 2 to win game 1: 0.35526315789473684
Prob of player 2 to win game 2: 0.4605263157894737
Prob of player 2 to win game 2 given that game 1 won: 0.4166666666666667
```

```
Prob of player 2 to lose game 2 given that game 1 won:
0.5833333333333334
Prob of player 2 to win game 2 given that game 1 was lost:
0.46153846153846156
Prob of player 2 to lose game 2 given that game 1 was lost:
0.5384615384615384
```

```
Probability of a player to win game 2 given that they have won game 1:
0.2638888888888889
```



	WON	**LOST**
WON	0.263889	0.736111
LOST	0.373626	0.626374

<u>Analysis of Wimbledon Open Men's tournament 2013</u>

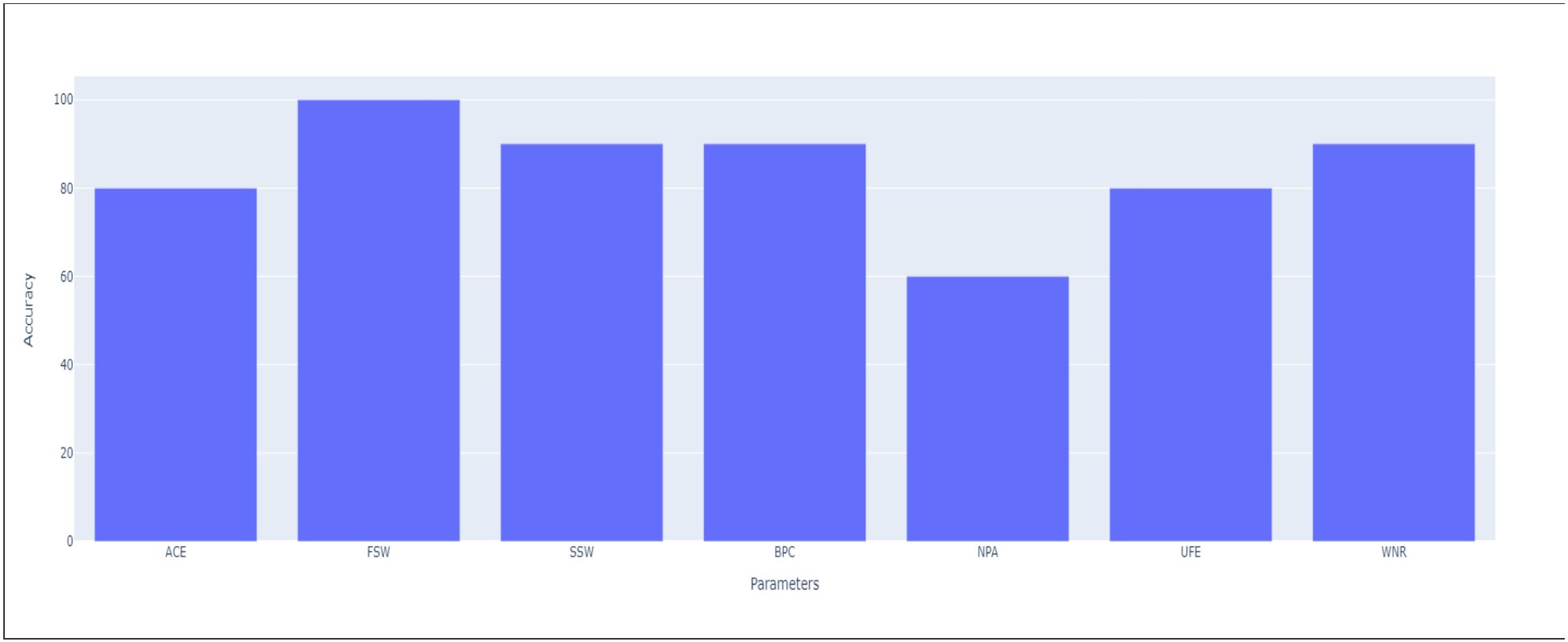
Q7. Who are the top 10 players in the tournament in terms of number of points scored and the matches won? Who are the top 10 players in terms of maximum number of aces scored, in terms of maximum number of first serve and second serve wins, largest number of breakpoints created, largest number of points attempted, number of unforced errors committed and number of winners won by the players? Which of these parameters has the largest amount of accuracy in terms of predicting the top 10 players of the game? Provide the value of accuracy in percentage and also the visualisation of the most accurate parameters that can be used to predict the top 10 players of a tournament?

Answer:
Procedure: First of all, the data set was opened as a data frame using the Pandas library. Then, all the rows were iterated and the dictionary containing the name of the players against the number of points won and the number of matches won was created. Then, the dictionary items were converted into a list, which was then sorted with respect to the number of games won and then the top 10 elements were taken and formed into a set. Then, the parameters like the number of Aces won was taken and all the rows were iterated in the data frame in the respective columns and a dictionary was created in which the names of the players and the number of Aces won was written. Then, the list of the dictionary items was sorted in the reverse order and the first and elements were taken which corresponded to the names of the players with the largest number of Aces which was later formed into a set and after taking an intersection with the set of top 10 players in terms of number of points I took the length of that intersection set which provided the number of correct guesses out of 10. A similar technique was used with other parameters like first serve wins denoted by fsw, second serve wins denoted by ssw, break points created denoted by BPC, number of points attempted denoted by npa, unforce committed denoted by ufe, number of winners won denoted by wnr, then these values of correct predictions was divided by 10 to give the ratio of correctly predicted values to the total number of values which was later multiplied by 100 to get the percentage accuracy. Then, this percentage accuracy of every parameter was plotted as a bar graph using the plotly library.

Results and observations:

Top 10 players: `{'A.Murray', 'B.Tomic', 'D.Ferrer', 'F.Verdasco', 'I.Dodig', 'J.Del Potro', 'J.Janowicz', 'J.Melzer', 'N.Djokovic', 'T.Berdych'}`

```
Number of correctly predicted top 10 players by the parameter:
ACE=8
FSW=10
SSW=9
BPA=9
NPA=6
UFE=8
WNR=9
```


100
80
60
40
20
0
Accuracy
ACE
FSW
SSW
BPC
NPA
UFE
WNR
Parameters

As we can see from the results that the parameter with the largest number of correctly predicted guesses was the number of first serve wins that is 100%. So, we can say that we can safely consider the first serve wins as a parameter that has a very high influence on the final result of the game in terms of predicting the top 10 players of the tournament. The parameter that has the least amount of accuracy that we have considered is the number of points attempted. Therefore, we can say that this parameter in itself may be incomplete in predicting the top 10 players of the tournament in this case.

Analysis of Wimbledon Open Women's tournament 2013

Q8. Who are the top 10 players in the tournament in terms of maximum number of matches won? In terms of the parameters that are given in the data frame, how can we visualise the correlation between all the parameters; that is, find out which parameter is most related to which another parameter? Using this, how can we find the parameters that are highly correlated in both the positive and negative sense to the result of the game? How can we visualise this? Can we use the top parameters in terms of correlation with the result to predict the result of the game; that is, can we use the parameters and take its first degree polynomial function and find the approximate value of the predicted result by minimising the main square error of the approximated value and the actual value of the result? What is the accuracy score of that and can we use it in the real-life situation to predict the outcomes of the match? Which parameters are used in this predictor that is which parameters have the maximum accuracy in terms of this prediction.

Answer:
Procedure: First of all, the data set was opened as a data frame using Pandas library. Then, using the Pandas function of df.corr, the data frame was formed out of that correlation matrix which was then plotted as heat map using the Seaborn library. Then, the column of results of the data frame was taken to get the correlation with other columns of the data frame and formed into another data frame through which the NaN values were dropped using the Pandas function of .dropna() and then value of the correlation and the parameter was plotted using the plotly library. Then, this matrix with the values of correlation between the parameters and the result column was sorted in the ascending order in order to get the most positively correlated parameters and then they were approximately used to predict the values of the result. For this, first of all, the values of result were made into another numpy array known as actual value and in order to get the predicted value first of all, the expectation of the variable x was calculated as E(X) using the numpy library. Then, the expectation of variable y, that is the result, variance of X and covariance between X and y was calculated. Then, in order to reduce the minimum mean square error, new variable which will be the coefficient of linear relationship that is a and b were calculated as a equal to covariance by variance of x and b is equal to expectation of y - covariance of x and y divided by the variance of x. Then, this value of predicted y was a rounded to the nearest integer because the values of the actual result were binary integers, so the predicted value of Y was obtained in the form of a float number. The rounded value was either 0 or 1. Then from sklearn library, accuracy score parameter function was imported in order to obtain the accuracy score of the predicted y against the actual y.

Results and Observations:

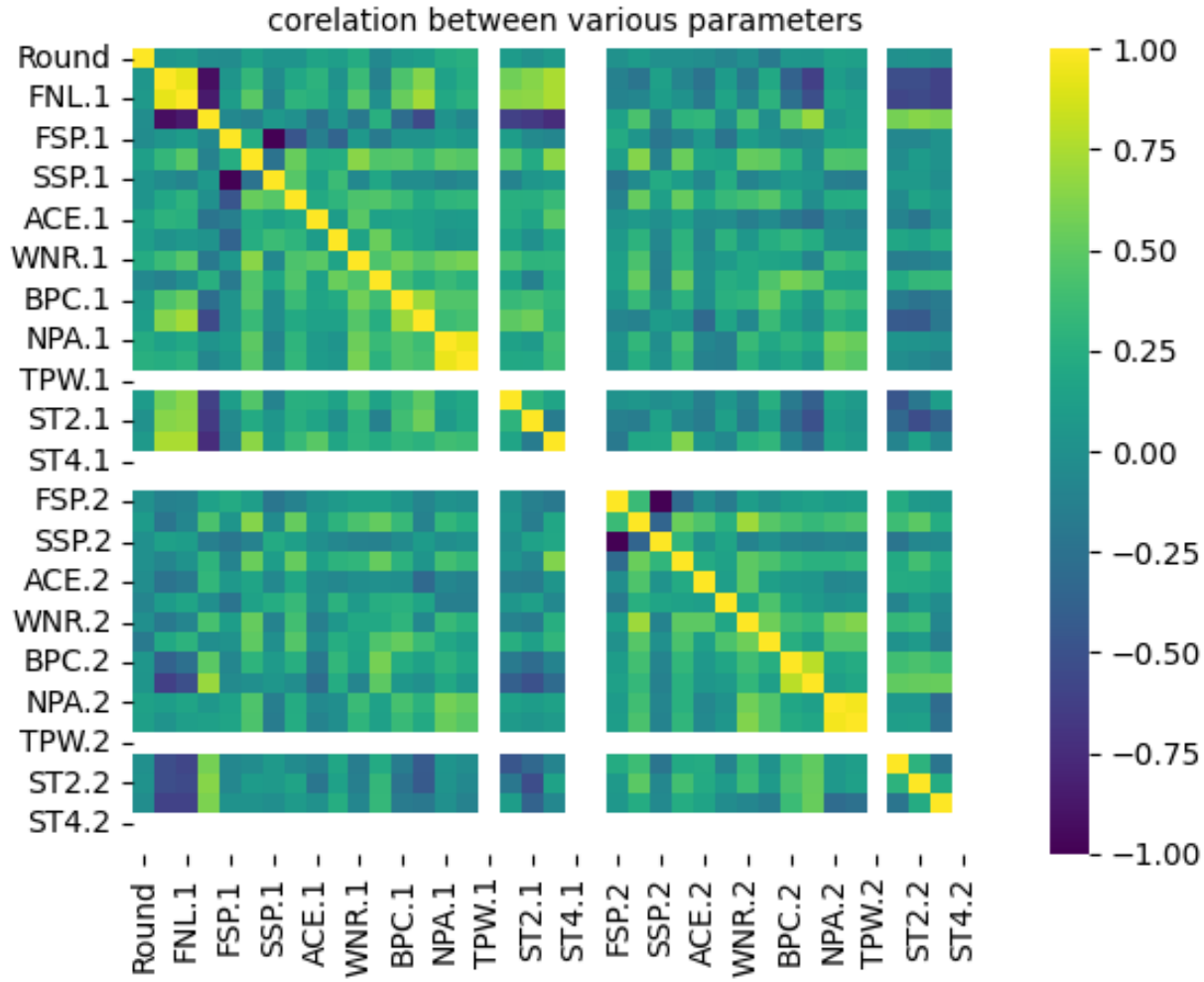
As we can observe from the obtained plot of the correlation between various parameters, we can see that every parameter is having the correlation as one with itself. It may not have the correlation one with other parameters. We can also observe that there are a few parameters which are highly correlated and are represented by yellow colour and there are some parameters which are negative correlated with each other and are denoted by dark blue colour. We can also observe the values of correlation between the result sub matrix and all the other parameters in the form of a plot. Also, we can see that parameter such as final score breakpoints won, breakpoints created, number of aces won, number of winners won, etc. have moderately high correlation.

As we can see from the accuracy score, we can say that the final score of the match is a very precise predictor of the result of the match as it has the accuracy of 100% ; where as other parameter such as breakpoints won, first serves, winners won, number of points attempted, etc. have low accuracy of about 77% to 53%.

```
FNL.2    -0.913940
ST5.2    -0.805609
BPW.2    -0.715185
BPC.2    -0.596371
ST3.2    -0.577760
ST2.2    -0.575809
ST4.2    -0.504317
ST1.2    -0.462658
ACE.2    -0.334428
WNR.2    -0.302885
SSP.1    -0.250266
UFE.1    -0.248728
NPW.2    -0.148848
FSW.2    -0.116951
DBF.1    -0.090449
SSW.2    -0.083403
```

Round	-0.062208
FSP.2	-0.054471
NPA.2	-0.000222
SSP.2	0.054471
NPA.1	0.068167
SSW.1	0.075514
NPW.1	0.128324
FSW.1	0.142591
UFE.2	0.154742
DBF.2	0.219954
FSP.1	0.250266
WNR.1	0.301725
ACE.1	0.305676
ST5.1	0.444500
ST1.1	0.483366
ST3.1	0.516165
BPC.1	0.518418
ST2.1	0.537916
ST4.1	0.557596
BPW.1	0.675129
FNL.1	0.913438
Result	1.000000

TPW.1	NaN
TPW.2	NaN



```
accuracy score of FNL: 1.0
Result=round(0.5323848258094395*F
NL+-0.12284641946437436)

accuracy score of BPW:
0.7704918032786885
Result=round(0.15837048716833407*
BPW+-0.00122202097847246)

accuracy score of FSW:
0.6065573770491803
Result=round(0.016456160797965576
*FSW+0.06817484093815374)

accuracy score of WNR:
0.6721311475409836
Result=round(0.020589986600000437*
WNR+0.09906971916219964)

accuracy score of NPA:
0.5327868852459017
Result=round(0.007013279152192498
*NPA+0.436680431960322)
```

V. Uniqueness of Approaches Used

The approaches used are unique in the way that the libraries of Pandas, NumPy, Plotly, Seaborn, sklearn and matplotlib, along with the prior knowledge of Python, were used in a complementary manner along with the concepts of this course to analyze the various parameters effectively. Various concepts of the course, such as basics of probability, probability distribution, cumulative distribution functions, smoothening the curve using kernel density estimation, mean, variance, standard deviation, correlations, joint variables, conditional probabilities, prediction of one variable based on other by minimizing the minimum mean square error, expectation, covariance, correlation matrix, etc., along with visualizations were also used by invoking various functions of the modules.

VI. Summary

Therefore, in this data narrative, various parameters of the tournaments' matches, such as, the first serve percentage, number of double faults committed, number of aces scored, number of break points created, total number of points attempted, total number of points won, total number of breakpoints won, etc. were extensively analysed in terms of the top 10 players, mean, variance, correlation with the result of the game, conditional probability of winning the game considering various instances of these parameters, providing the visualisation of the parameters and the result as joint variables and also visualising their marginal, finding the top 10 players of the game, analysing the accuracy of the players in terms of number of breakpoints won for every breakpoint created and the accuracy of the players in terms of the number of points won for every point attempted. We also analyse this using the visualisation of pie charts, joint plots with marginals plotted as histograms, smooth curves using kernel density estimation, regression, etc. Then, I also analysed the conditional probability of any player to win the second set given that he has won the first set, lost first set and the conditional probability of any player to lose the second set given that he has won and lost the first set. We also analysed the probability of a player to win the first and second set. We also created the visualisation for visualising this in terms of heat maps. We also found the accurate parameter that can predict the top 10 players satisfying that parameter and found out the accuracy of our predictor of this type by taking the length of the set of the intersection of the top 10 players predicted and actual top 10 players. We also analysed a more precise predictor by taking the first degree polynomial of a parameter. Then, we minimized its minimum mean square error and optimised the coefficients. By this, we analysed the relation between these parameters' function and the result of the game. We also found out the accuracy of these parameters by the sklearn function and found the most and least accurate parameters to predict the outcome of a match.

REFERENCES

[1] "NumPy User Guide." NumPy user guide - NumPy v1.24 Manual. Accessed April 17, 2023. https://numpy.org/doc/stable/user/index.html

[2] "Matplotlib - Visualization with Python v3.7.0." Matplotlib. Accessed April 17, 2023. https://matplotlib.org/

[3] "Pandas User Guide." User Guide - pandas 1.5.3 documentation. Accessed April 17, 2023. https://pandas.pydata.org/docs/user_guide/index.html#user-guide

[4] "Sklearn." scikit. Accessed April 20, 2023.https://scikit-learn.org/stable/.

[5] "Statistical Data Visualization." seaborn. Accessed April 23, 2023. https://seaborn.pydata.org/.

[6] "Plotly in Python." Getting started with plotly in Python. Accessed April 23, 2023. https://plotly.com/python/getting-started/.