DATA NARRATIVE 3 REPORT

[1]

Pankaj
First year btech. Chemical eng.,
IIT Gandhinagar,India

Abstract - This report finds answers to various questions from 8 datasets based on 4 Grand slam tennis tournaments, for men and women. This report gives insights of players performances and their dependencies on various factors such as unforced errors, number of aces, break point etc. .This report gives us a decent idea about the factors that matter most in a player's overall performances.

I. OVERVIEW OF DATA SET

These datasets are a compilation of match statistics for men's and women's singles matches played at significant tennis tournaments, such as the Australian Open, French Open, Wimbledon, and US Open in 2013.

For each player in each match, the dataset includes data on numerous match statistics, such as the number of aces, double faults, first serve %, total points gained, and more. The dataset also contains details about the break points, round number, and total points.The dataset can be used for a variety of machine learning and data analysis activities in the area of sports analytics, including analyzing player performance and making predictions about the results of games, among other things.

II. SCIENTIFIC QUESTIONS / HYPOTHESES

1) First (AusOpen-men-2013.csv) : Hypotheses - If a player wins the first set then he will most probably win the whole match but if he won the first set by a very low margin then the probability of winning the match is not that much high.

2)Second(AusOpen-women-2013.csv): Using linear regression, predict player's performance by the number of the number of unforced error she had does in a match.

3)Third(FrenchOpen-men-2013.csv) : Plot the probability density functions of number of Aces won by player 1 when he won and also plot pdf when he lost. What are your observations?

4)Fourth(FrenchOpen-women-2013.csv) : What are average unforced errors by the players in a match in each round ? Will the pattern remain the same in big matches like semifinals and finals?

5)Fifth(USOpen-men-2013.csv) : What is the correlation between number of breakpoints created by player1 and total points won by him? Visualize using plot.

6) Sixth (USOpen-women-2013.csv) : What is the average winners earned by a player if she wins the match? Plot bar graph between 20 players that won the match and their winners earned.

7) Seventh(Wimbledon-men-2013.csv): What are the total first serve points won by each of the finalists in the tournament and does that reflect on the final result.

8) Eight(Wimbledon-women-2013.csv): What is the probability that if a player has more than 8 breakpoints then he will win the match and what is the correlation between breakpoints and unforced errors?

III. DETAILS OF LIBRARIES AND FUNCTIONS

Pandas- Pandas is a library in python that is used to work, manipulate and analyze dataframe and series. It is very good tool to perform different operations on a dataset. Here it is used excessively to answer questions and in analyzing both dataset

.

Matplotlib- It is a library in python which is used to plot different kinds of plots such as bar graph, line graph, pie chart, histograms. Here I used it to better visualize the answers in graphical forms.

NumPy - NumPy is a library in python that helps us to do a wide range of operations on multidimensional arrays and matrices . Here it is used to get unique elements of an array in a question.

Seaborn - Seaborn provides a variety of features, including pre-built themes, distribution visualization, linear modeling, categorical data visualization, time series data visualization, and complicated multi-plot visualizations. Here I have used it to visualize the probability distribution function.

sklearn - sklearn library in python provides a wide range of operations in especially machine learning. Here we have used it for linear regression.

IV. ANSWERS TO THE QUESTIONS

1. Answer 1

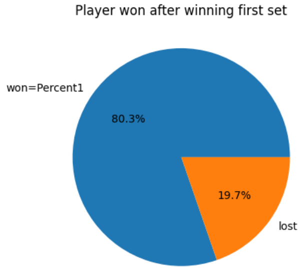
I first made a dataset out of original in which player1 won the first set and then out of it I further made a dataset in which player1 won the match , then I found the probability by dividing the number of rows of both datasets. For a very low margin win I took a dataset in which player 1 wins set 1 by only 1 point and then found its probability of winning in a

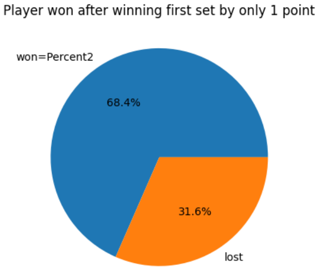
[1]

similar way .And I found out that probability of winning reduces if win margin is low in set1.

Code output -

```
Probability of a player winning the match if he won the first set : 0.803030303030303
Probability of a player winning the match if he won the first set by only 1 point: 0.6842105263157895
```



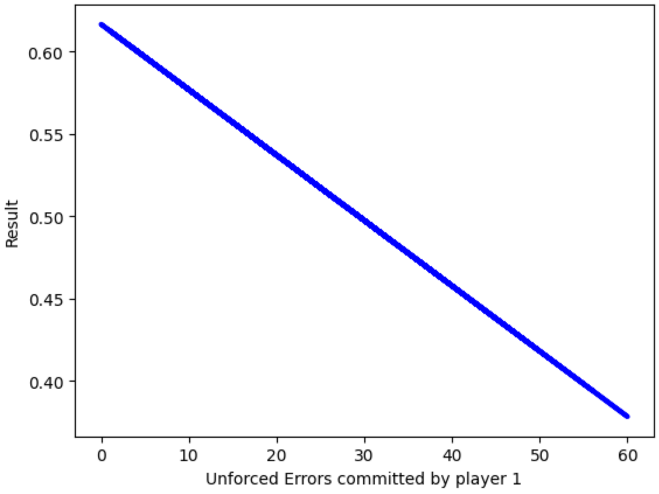


2. Answer 2

First I imported linear regression from sklearn library in python using -”from sklearn.linear_model import LinearRegression”

Then i trained the data in the model in which x was unforced error and y was match winning probability. Code output -

```
Coefficients:  [-0.00397142]
Intercept:  0.6165000505453048
```



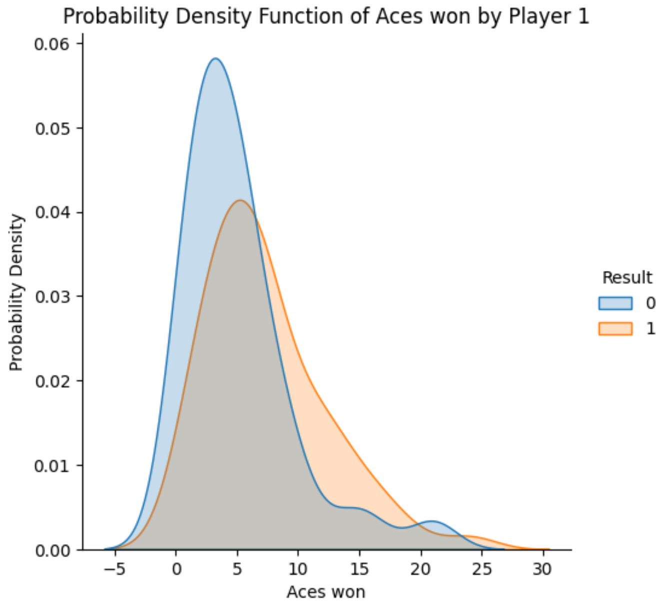
3. Answer 3

I imported seaborn library in python as sns and then using -

```
“import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns


df = pd.read_csv('FrenchOpen-men-2013.csv')
X = df[['ACE.1']]
y = df['Result']
sns.displot(df, x="ACE.1", hue="Result", kind="kde", fill=True)
plt.title("Probability Density Function of Aces won by Player 1")
plt.xlabel("Aces won")”
plt.ylabel("Probability Density")”
```

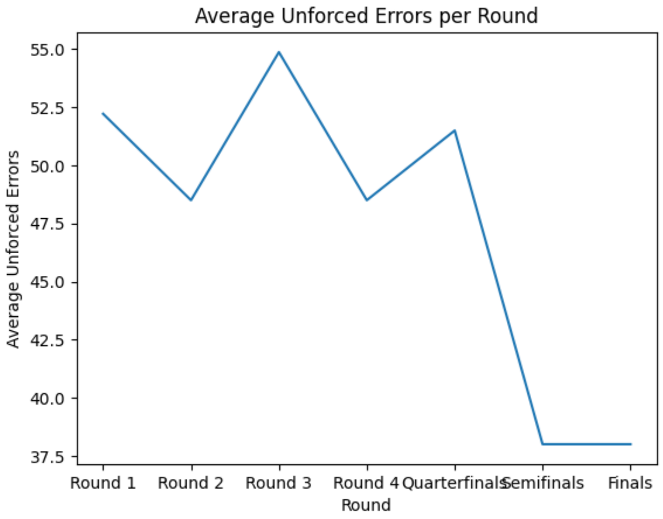
and in plot i observed that the pdf by of winning matches is slightly shifted towards right of losing matches pdf, which tells that if a player wins more aces points his chances of winning increases.



4. Answer 4

First I made 7 datasets from the original for all 7 rounds and then i calculated average unforced errors per match in each round .And finally i observed that the number of unforced errors significantly went down in semifinals and finals, which tells us about the quality of games in semis and finals in this tournament. Code output -

```
Average unforced errors roundwise: [52.21875, 48.5, 54.875, 48.5, 51.5, 38.0, 38.0]
```



5 Answer 5

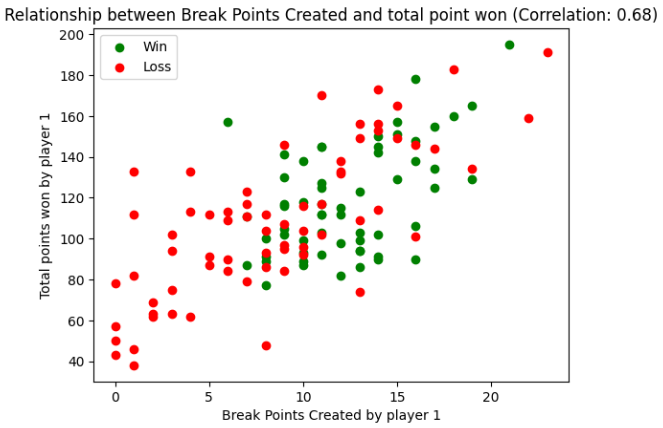
For correlation between break points created and total points won, i used -”corr = data['BPC.1'].corr(data['TPW.1'])

For plotting we will use

```
-”plt.scatter(data['BPC.1'][data['Result']==1],
data['TPW.1'][data['Result']==1], color='green', label='Win')
plt.scatter(data['BPC.1'][data['Result']==0],
data['TPW.1'][data['Result']==0], color='red', label='Loss')


plt.xlabel('Break Points Created by player 1')
plt.ylabel('Total points won by player 1')
plt.title('Relationship between Break Points Created and total
point won (Correlation: {})'.format(round(corr, 2)))
plt.legend()
plt.show()”
```

We can observe in the plot that as break points increase chances of winning too increases.

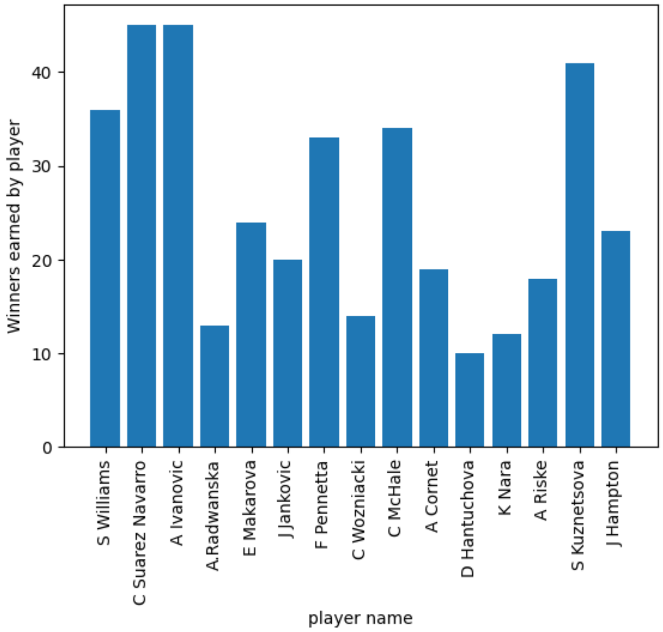


6. Answer 6

I first made a dataset of all winners and for average winners earned by a player i found out mean of the “WNR.1” column of this dataset, and then plotted the bar graph of 20 winner players with their winners earned.

Code output -

```
Winners earned by player 1: 23.36111111111111
```



7. Answer 7

I first found out the finalists players by getting round 7 players name, and then i found out the sum of all the first serve won by each finalist throughout the tournament and then i finally found their results in final, and found out that the player who won more first serve throughout the tournament actually lost the game in final. My code - “import pandas as pd

```
import matplotlib.pyplot as plt
data = pd.read_csv("Wimbledon-men-2013.csv")


df = data[data['Round']==7]
p1 = df['Player1'].iloc[0]
p2 = df['Player2'].iloc[0]
df2 = data[data['Player1']==p1]
df3 = data[data['Player2']==p1]


Serve_point1 = df2["FSW.1"].sum()
Serve_point2 = df2["FSW.2"].sum()


if df.iloc[0]['Result'] ==1:
   w = p1
else:
   w = p2
```

```
print(f"First serve won  by {p1} in the tournament:
{Serve_point1}")
print(f"First serve won by {p2} in the tournament:
{Serve_point2}")
print("Final won by:",w) ″
```

Code output -

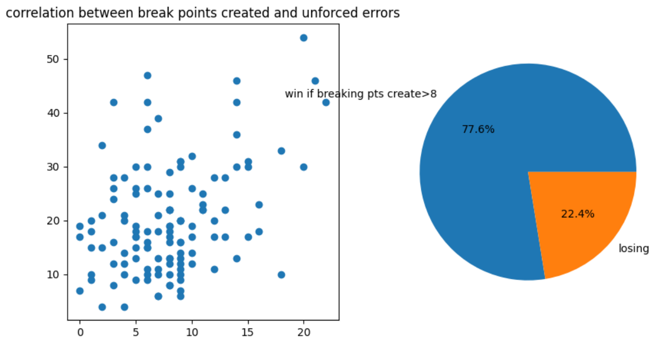
```
"First serve won  by N.Djokovic in the tournament: 348
First serve won by A.Murray in the tournament: 337
Final won by: A.Murray”
```

8. Answer 8

First i found out that players with more than 8 breakpoints , then i found how many of them won, and found probability and then for correlation between breakpoints and unforced errors i used - “corr = data['BPC.1'].corr(data['UFE.1'])”

Code output -

Probability that if a player has more than 8 breakpoints then he will win the match : 0.7755102040816326



V. SUMMARY OF THE OBSERVATIONS

So, summarizing my observations i would like to say in Probability of winning decrease if the winning margin in the first set is low, as number of aces won by player increases its chance of winning increases. We also found out that number of unforced errors reduced significantly in semifinals and finals.As break points created increases chances of winning increase and if break points created is more than 8 than their is more than 70 percent chances that the player will win the game .Also we observed that the player who won more first serve throughout the tournament lost the final game to a player who won less first serve.

VI. REFERENCES

“Pandas - Python Data Analysis Library,” n.d. https://pandas.pydata.org/.

“Matplotlib — Visualization with Python,” n.d. https://matplotlib.org/.

“NumPy,” n.d. https://numpy.org/.

“Seaborn: Statistical Data Visualization — Seaborn 0.12.2 Documentation,” n.d. https://seaborn.pydata.org/.

“Scikit-Learn: Machine Learning in Python — Scikit-Learn 1.2.2 Documentation,” n.d. https://scikit-learn.org/stable/.

VII ACKNOWLEDGEMENT

I am really thankful to the professor and teaching assistant for this learning opportunity. I learned so many things while making this report and this could not have been done without the help of my teaching assistant and professor. Their guidance was instrumental in helping me to complete this report Once again thank you.