ES114 Probability,Statistics and Data Visualization Report: Data Narrative 3

Keshav Sobania, B TECH 22,Civil Engineering, ROLL NO. 22110118

A. Overview of The Dataset:-

The dataset includes data on tennis matches played in 2013 at the Australian Open, French Open, US Open, and Wimbledon spanning four events and each file has 42 columns and a number of rows for each match.
Each tournament has separate files for the men's and women's matches.The information includes some stats about each player.The dataset can be utilized for various things, such as analyzing player performance individually and spotting trends.

B. Scientific Questions/Hypothesis:-

Question 1: Total no. of unique players in all tournaments and plot the pie chart of distribution of individual players in each tournament in the men category.

Question 2: Top 10 women players who played the highest number of matches in the Australia tournament and plot the bar graph.

Question 3: If Player 1 has won 50% or more of their net points attempted in a match in the French men Tournament, what is the probability that player 1 will win the match?

Question 4: Plot a pie chart for the distribution of wins and losses of player 2 has a first serve win percentage of less than 50%.

Question 5:Name the top 3 players who won most no. of matches in the US tournament in the men category.

Question 6: Plot a scatter plot of first serve percentage vs. second serve percentage for player1 in the US women's tournament.

Question 7: What is the probability of player1 has a higher percentage of break points than player 2 and given that player 1 won the match in wimbledon men tournament?

Question 8: Plot a scatter plot between first serve percentage of player1 and result of that match.

C) Details of libraries and functions:-

a) Libaries:-

● **Numpy-**A Python library for numerical computing that offers substantial array and matrix manipulation capabilities and various array-operating mathematical functions. Like Python lists, NumPy arrays are far more effective for numerical computations.

● **Pandas-**The Pandas module for Python is a popular open-source tool for handling and analyzing data. For reading and writing data into and out of some file formats, such as CSV, Excel and others, Pandas provides various data structures besides the DataFrame, such as Series, a one-dimensional labeled array that may hold any data type.

● **Matplotlib**-A well-liked open-source Python data visualization package that offers extensive programmable 2D and 3D plotting functions for data visualizations. It provides various graphs with Matplotlib, including line plots, scatter plots, bar plots, histograms, etc. And including the ability to modify labels, axis limitations, colors, styles, and more.

b) Functions:-

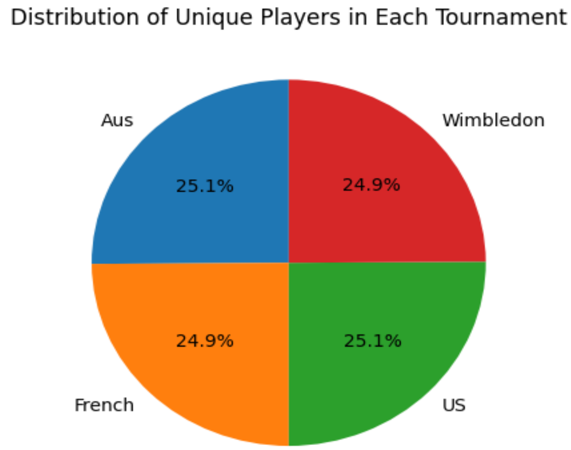
● Loops, indexes, and if else functions have been used by me for deducing the answers.

D) Answers to the Questions:-

1.

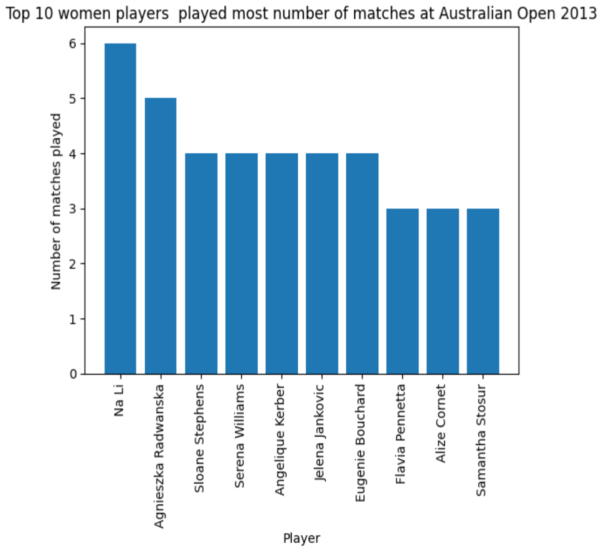
By code we found total number of unique men players in all 4 tournaments are 300

And at each tournament unique players are
Aus: 128, french: 127, Us: 128, Wimb: 127



2.

By code and with the help of matplotlib we found the top 10 women who played the most number of matches in the Australia tournament.

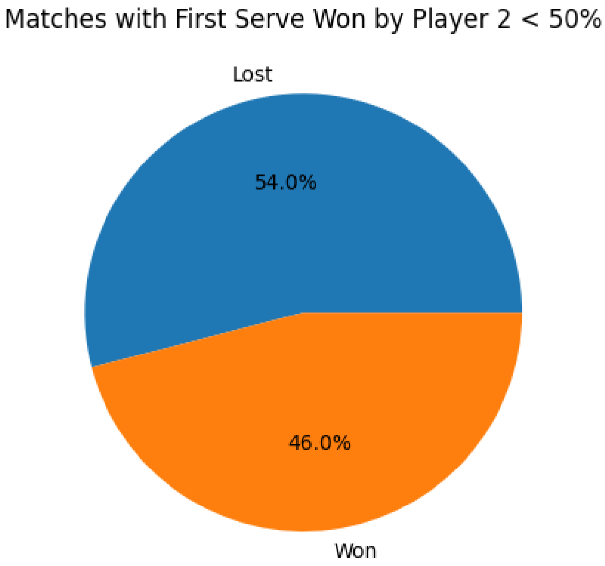


3. Matches where Player 1 has won 50% or more of their net points attempted found by help code=109

Number of matches in which player 1 won= 53

probability=53/109=0.4862385321100917

4.



Matches of player 2 with FSW is less than 0.5 and then we divided these matches on the basis of won and lost.

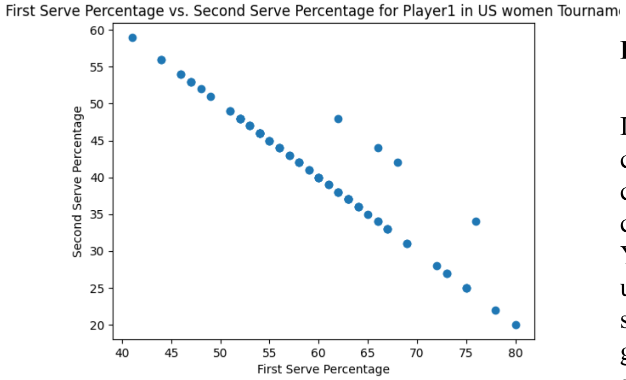
5.

By code first we count the win of each player with the help of a for loop and insert it in the dictionary and then sort it.

Rafael Nadal : 7
Richard Gasquet : 5
Novak Djokovic : 5

6.

With the help of matplotlib we compare the first and second serve percentage of player1 in US women 's tournament.



7.
This question is based on conditional probability.

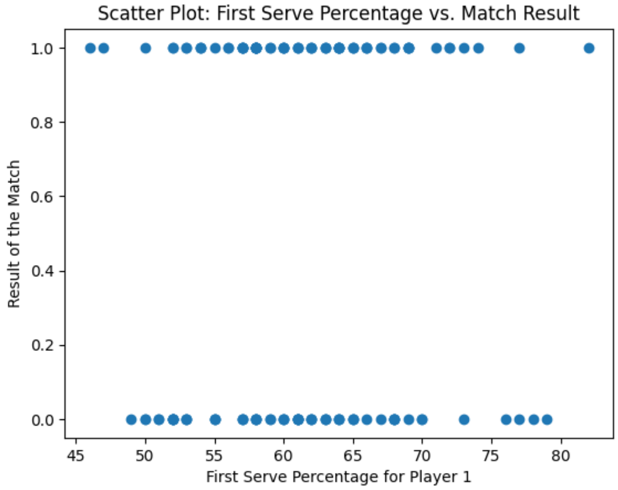
P(A/B)=Probability of(A intersection B) / Probability of B

B=Number of matches won by player 1.

A=number of matches where player 1 has a higher percentage of break points than player 2.

P=52/55=0.9454545454545454

8. We plotted the scatter plot between the first serve percentage and result of that match.



From this graph we can conclude that having a higher first serve percentage is generally beneficial and increases the chance of winning.
As “1” and “0” shows that player 1 won the match and lost the match resspectively.

E)Summary to the observation:-

I've come to the conclusion that a Python library can be used to read CSV files like pandas containing very large amounts of data. You can also create a data frame and draw a chart using this data. You can also perform many operations on this data using pandas functions. Also, there is an open source library called matplotlib that is used to plot graphs such as histograms, bar charts, pie charts, scatter plots. Numpy is also an open-source library for the Python programming language that allows working with matrices and one-dimensional arrays. Seaborn is also a very popular open source library in Python. Used for data visualization. Used to create informative statistical charts.

F) References:-

- https://towardsdatascience.com/python-pandas-tricks-3-best-methods-4a909843f5bc
- https://favtutor.com/blogs/pandas-unique-values-in-column
- https://www.w3schools.com/python/pandas/pandas_dataframes.asp
- https://www.w3schools.com/python/pandas/pandas_csv.asp
- https://www.w3schools.com/python/pandas/pandas_plotting.asp
- https://www.w3schools.com/python/pandas/pandas_ref_dataframe.asp
- https://www.geeksforgeeks.org/how-to-get-column-names-in-pandas-dataframe/
- https://www.geeksforgeeks.org/how-to-get-column-names-in-pandas-dataframe/
- https://www.geeksforgeeks.org/python-pandas-dataframe-series-head-method/?ref=lbp

G) Acknowledgements:-

I want to thank Professor Shanmuga for all his help and guidance. His knowledge has played a significant role in forming my perspective on the topic. I am also thankful to my TAs for their guidance in complementing my assignment.