Module 6: Data Manipulation Using pandas

Assignment - I

1. Extract data from the given SalaryGender CSV file and store the data from each column in a separate NumPy array.

2. Find:
   1. The number of men with a Ph.D.
   2. The number of women with a Ph.D.

3. Store the "Age" and "Ph.D." columns in one DataFrame and delete the data of all people who don't have a Ph.D. from the SalaryGender CSV file.

4. Calculate the total number of people who have a Ph.D. degree from the SalaryGender CSV file.

5. How do you count the frequency of each value appearing in the given array of integers?

   [0, 5, 4, 0, 4, 4, 3, 0, 0, 5, 2, 1, 1, 9]

   Answer: array([4, 2, 1, 1, 3, 2, 0, 0, 0, 1]) which means 0 comes 4 times, 1 comes 2 times, 2 comes 1 time, 3 comes 1 time and so on.

6. Create a NumPy array [[0, 1, 2], [3, 4, 5], [6, 7, 8],[9, 10, 11]]) and filter the elements greater than 5.

7. Create a NumPy array having NaN (Not a Number) and print it.

   array([nan, 1., 2., NaN, 3., 4., 5.])

   Print the same array omitting all elements, which are NaN.

8. Create a 10x10 array with random values and find the minimum and maximum values.

9. Create a random vector of size 30 and find the mean value.

10. Create a NumPy array having elements 0 to 10 And negate all the elements between 3 and 9.

11. Create a random array of 3 rows and 3 columns and sort it according to the first, second, or third column.

12. Create a four dimensions array to get the sum over the last two axes at once.

13. Create a random array and swap two rows of an array.

14. Create a random matrix and compute the matrix rank.

15. Assuming you are a public-school administrator, analyze various school outcomes in Tennessee using pandas. Some schools in your state of Tennessee are performing below average academically. Your superintendent, under pressure from frustrated parents and voters, approached you with the task of understanding why these schools are underperforming. To improve school performance, you need to learn more about these schools and their students, just as a business needs to understand its strengths and weaknesses, and its customers. Though you are eager to build an impressive explanatory model, you know the importance of conducting preliminary research to prevent possible pitfalls or blind spots. Thus, you engage in thorough exploratory analysis, which includes: a lit review, data collection, descriptive and inferential statistics, and data visualization.

Phase 1: Data collection

Here is the data of every public school in middle Tennessee. The data also includes various demographic, school faculty, and income variables. You need to convert the data into useful information.

- Read the data in the pandas DataFrame
- Describe the data to find more details

Phase 2: Group data by school ratings

Selects indicators, which describe school administration (**stu_teach_ratio**) or the student body (for example, **reduced_lunch**). **reduced_lunch** is a variable that measures the average percentage of students (per school) enrolled in a federal program providing lunch to students from impoverished households. In short, **reduced_lunch** is a good proxy for household income.

Separates **reduced_lunch** and groups the data by **school_rating** using the pandas **groupby()** function.

Phase 3: Correlation analysis

Find the correlation between **school_rating** and **reduced_lunch**. The values in the correlation matrix table will be between -1 and 1. A value of -1 indicates the strongest possible negative correlation, meaning as one variable decreases, the other increases. And a value of 1 indicates the opposite.

Phase 4: Scatter plot

Find the relationship between **school_rating** and **reduced_lunch**. Plot a graph with the two variables on a scatter plot. Each dot represents a school. The placement of the dot represents the school's rating (Y-axis) and the percentage of its students on reduced lunch (X-axis). The downward trend line shows the negative correlation between **school_rating** and **reduced_lunch** (as one increases, the other decreases). The slope of the trend line indicates how much **school_rating** decreases as **reduced_lunch** increases. A steeper slope would indicate that a small change in **reduced_lunch** has a big impact on school_rating, while a more horizontal slope would indicate that the same small change in **reduced_lunch** has a smaller impact on **school_rating**.

Phase 5: Correlation matrix

Red cells indicate a positive correlation; blue cells indicate a negative correlation; white cells indicate no correlation. The darker the colors, the stronger the correlation (positive or negative) between those two variables. Draw a graph of the correlation matrix having all the important fields of the DataFrame.