Module 4: Introduction to NumPy & Pandas

Case Study I

Case Study

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 PhD
   2. The number of women with a PhD

3. Store the “Age” and “PhD” columns in one DataFrame and delete the data of all people who don’t have a PhD from SalaryGender CSV file.

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

5. How do you Count The Number Of Times Each Value Appears In An Array Of Integers?

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

   Answer should be 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 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 1st column, 2nd column or 3rd column.

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

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

14. Create a random matrix and Compute a matrix rank.

15. Analyse various school outcomes in Tennessee using pandas. Suppose you are a public school administrator. 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 under-performing. To improve school performance, you need to learn more about these schools and their students, just as a business needs to understand its own strengths and weaknesses and its customers. Though you is eager to build an impressive explanatory model, you know the importance of conducting preliminary research to prevent possible pitfalls or blind spots. Thus, you engages in a thorough exploratory analysis, which includes: a lit review, data collection, descriptive and inferential statistics, and data visualization.

Phase 1 - Data Collection

Here is a 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 pandas data frame
- Describe the data to find more details

Phase 2 - Group data by school ratings

Chooses indicators that describe the student body (for example, reduced_lunch) or school administration (stu_teach_ratio) hoping they will explain school_rating. reduced_lunch is a variable measuring the average percentage of students per school enrolled in a federal program that provides lunches for students from lower-income households. In short, reduced_lunch is a good proxy for household income.

Isolates 'reduced_lunch' and groups the data by 'school_rating' using pandas groupby method and then uses describe on the re-shaped data

Phase 3 – Correlation analysis

Find the correlation between 'reduced_lunch' and 'school_rating'. 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 that 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

An efficient graph for assessing relationships is the correlation matrix, as seen below; its color-coded cells make it easier to interpret than the tabular correlation matrix above. Red cells indicate positive correlation; blue cells indicate 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 correlation matrix having all important fields of data frame.