DATA ANALYSIS ON ADULT INCOME – DATA PROJECT 9

Part 1

1. Import required libraries.
2. Read csv file.
3. Display top 10 rows of the dataset.
4. Check last 10 rows of the dataset.
5. Find shape of dataset (number of rows and number of columns).
6. Getting information about dataset like total number rows, total number of columns, datatypes of each column and memory requirement.
7. Fetch random sample from the dataset (50%).
8. Check null values in the dataset
   a. Check column and row wise
   b. Please show heatmap as well
9. Perform data cleaning [replace '?' with NaN]
   a. Find out how many '?' are there
   b. And after finding replace them with nan
      i. Try to show all columns to get to know on which column you must work

        ii. After finding columns kindly start process of replacing

        iii. After having replace show if there is any column left for replacing or not?

    c. Visualize it with heatmap too.

10. Drop all the missing values

    a. First check missing values in percentage

    b. After dropping values check shape as well

11. Check for duplicate data and drop them

    a. Check for duplication in boolean

    b. Now drop duplicates

Part 2

1. Get overall statistics about the dataframe
    a. Try to describe the data
    b. Try to describe categorical as well as numerical data
    c. Show unique values of ‘education’ column
    d. Show unique values of ‘educational-num’ column
2. Drop the columns education-num, capital-gain, and capital-loss
    a. After dropping kindly check whether do we have those columns left or not

Univariate analysis -in this analysis we take one variable at a time and perform analysis on it.

1. What is the distribution of age column?
    a. Check min, max and average values
    b. Use histogram
2. Find total number of persons having age between 17 to 48 (inclusive) using between method
    a. First try to do it without using between method
3. What is the distribution of workclass column?
    a. Use histogram for that
4. How many persons having bachelors and master’s degree?
    a. Try to use two different methods for solving this

Bivariate analysis – it is used to find relationship between two different variables

1. Check relationship between ‘income ‘and ‘age’ column using boxplot.
2. Replace income values [‘<=50k’, ’>50k’] with 0 and 1
    a. Show how many have less than 50 k and greater than 50 k income.
    b. Use count plot to visualize it.
    c. Create a function which will return 0 and 1 according to condition.
    d. Assign returning result into newly added column named as ‘econded income’

    e. Do the same with using ‘replace’ method.

3. Which workclass getting the highest income?
4. How has better chance to get income greater than 50k male or female?
5. Convert workclass columns datatype to category datatype
    a. Converting datatype can be useful in case of saving memory, you need to optimize it.

Bonus – Can go ahead as much as you can.