

About data:

The insurance.csv dataset contains 1338 observations and 7 attributes.

Context: The data contains medical costs of people characterized by certain attributes. Let's see if we can dive deep into this data to find some valuable insights.

Attributes:-

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m^2) using the ratio of height to weight, ideally 18.5 to 24.9
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance.

Tasks to perform:

1. Import the necessary libraries
2. Read the data as a data frame
3. Perform basic EDA which should include the following and print out your insights at every step.
 - a. Shape of the data
 - b. Data type of each attribute
 - c. Checking the presence of missing values
 - d. 5 point summary of numerical attributes
 - e. Distribution of 'bmi', 'age' and 'charges' columns.
 - f. Measure of skewness of 'bmi', 'age' and 'charges' columns - **optional**
 - g. Checking the presence of outliers in 'bmi', 'age' and 'charges' columns - **optional**
 - h. Distribution of categorical columns (including children)
 - i. Pair plot that includes all the columns of the data frame
4. Answer the following questions with statistical evidence
 - a. Do charges of people who smoke differ significantly from the people who don't?
 - b. Does bmi of males differ significantly from that of females?
 - c. Is the proportion of smokers significantly different in different genders?
 - d. Is the distribution of bmi across women with no children, one child and two children, the same ?