Linear Regression - Insurance

Data:

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

Domain: Healthcare

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.
   - Shape of the data
   - Data type of each attribute
   - Checking the presence of missing values
   - 5 point summary of numerical attributes
   - Distribution of 'bmi', 'age' and 'charges' columns.
   - Measure of skewness of 'bmi', 'age' and 'charges' columns
   - Checking the presence of outliers in 'bmi', 'age' and 'charges columns
   - Distribution of categorical columns (include children)
   - Pair plot that includes all the columns of the data frame

Key asks:

- Do charges of people who smoke differ significantly from the people who don't?
- Does BMI of males differ significantly from that of females?
- Is the proportion of smokers significantly different in different genders?
- Is the distribution of BMI across women with no children, one child and two children the same?