greatlearning

Statistical Learning Project

Data Description:

The data at hand contains medical costs of people characterized by certain attributes.

Domain:

Healthcare

Context:

Leveraging customer information is paramount for most businesses. In the case of an insurance company, attributes of customers like the ones mentioned below can be crucial in making business decisions. Hence, knowing to explore and generate value out of such data can be an invaluable skill to have.

Attribute Information:

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.

Learning Outcomes:

- Exploratory Data Analysis
- Practicing statistics using Python
- Hypothesis testing

Objective:

We want to see if we can dive deep into this data to find some valuable insights.

Steps and tasks:

1. Import the necessary libraries (2 marks)
2. Read the data as a data frame (2 marks)
3. Perform basic EDA which should include the following and print out your insights at every step. (28 marks)
    a. Shape of the data (2 marks)
    b. Data type of each attribute (2 marks)
    c. Checking the presence of missing values (3 marks)
    d. 5 point summary of numerical attributes (3 marks)
    e. Distribution of 'bmi', 'age' and 'charges' columns. (4 marks)
    f. Measure of skewness of 'bmi', 'age' and 'charges' columns (2 marks)
    g. Checking the presence of outliers in 'bmi', 'age' and 'charges columns (4 marks)
    h. Distribution of categorical columns (include children) (4 marks)
    i. Pair plot that includes all the columns of the data frame (4 marks)
4. Answer the following questions with statistical evidence (28 marks)
    a. Do charges of people who smoke differ significantly from the people who don't? (7 marks)
    b. Does bmi of males differ significantly from that of females? (7 marks)
    c. Is the proportion of smokers significantly different in different genders? (7 marks)
    d. Is the distribution of bmi across women with no children, one child and two children, the same? (7 marks)

References:

- Applications of Data science in insurance domain
- Data science in Insurance