**APPLIED STATISTICS**

**DATA DESCRIPTION:** The data at hand contains medical costs of people characterized by certain attributes. The insurance.csv dataset contains 1338 observations and 7 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

**Tasks to performed**:

1. Import the necessary libraries

2. Read the data as a data frame

3. Performed basic EDA which should include the following and print out the insights gathered 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

g. Checking the presence of outliers in ‘bmi’, ‘age’ and ‘charges columns

h. Distribution of categorical columns (include children)

i. Pair plot that includes all the columns of the data frame

4.Now different questions with statistical evidence that arises and needs insight upon.

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 ?

Answer to the questions are provided using **Hypothesis test( p-value, Chi-square test and Anova test)**