**INSURANCE DATASET**

The dataset includes the following variables: Age, Sex, BMI, Children, Smoke, Region, and Charges. These variables were analyzed to understand the demographic and behavioural characteristics of the subjects, as well as their medical charges.

A **five-point summary** (minimum, first quartile, median, third quartile, maximum) was calculated for numerical variables to summarize the central tendency and dispersion.

**Box plots** for Age, BMI, and Charges were constructed to visualize the data distribution and identify potential outliers.

Bar graphs were created to depict the number of people with different numbers of children, distribution of subjects by Sex, Smoking Status (Yes/No), and Region. These visualizations provided insights into the demographic composition of the dataset.

**TYPES OF TESTINGS**

1. **Hypothesis test** was conducted to **compare the BMI of male and female subjects**. The null hypothesis stated that BMI of male and female customers are same. We get a p value greater than 5% and fail to reject the null hypothesis.
2. A **chi-squared test** was performed to evaluate the **independence between Smoking Status and Region.** As per the null hypothesis, there is no relationship between the region and the smoking status. The p-value obtained failed to reject the null hypothesis.
3. **Hypothesis test** was done to know about the proportion of smokers in each gender. Null hypothesis saying, ‘Proportion of smokers in each gender’ was rejected.
4. Studied the **Pearson correlation coefficients** to understand the strength of relationships between Age, Smoking Status, vs Charges. Influence of smoking on charges was more than the influence of age.