# Problem Statement

**Objective:**

To analyze the dataset that will help to create a model that will predict the cost of medical insurance based on various input features.

**Problem Statement:**

An insurance agency, ABC Insurance, has a large dataset containing information about their policyholders and claims. They want to perform exploratory data analysis (EDA) on this dataset to gain insights that can help them make better business decisions and improve their operations.

The agency wants to analyze the different body types and the environment that affect the premium. The disease's effect or the cost of treatment differs depending on the circumstances. For example, a smoker's medical insurance premium may be higher than that of a healthy person, because smokers are more likely to develop chronic diseases. The agency wants to analyze the data to research healthcare premium costs.

**Domain:** Healthcare

**Dataset:** insurance dataset (insurance.csv)

**Dataset Description:**

|  |  |
| --- | --- |
| **age** | Age of the person |
| **sex** | Female or Male |
| **BMI** | BMI value to estimate an individual's health and fitness condition |
| **children** | number of children (1,2,3,4, or 5) |
| **smoker** | The person is a smoker or not |
| **region** | Specifies the region (northeast, northwest, southeast, southwest) |
| **charges** | the amount of insurance |

# Task

1. Import libraries such as Pandas, matplotlib, NumPy, and seaborn and load the insurance dataset

A screenshot of a computer

Description automatically generated

1. Check the shape of the data along with the data types of the column

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated

1. Check missing values in the dataset and find the appropriate measures to fill in the missing values

A screenshot of a computer

Description automatically generated

1. Explore the relationship between the feature and target column using a count plot of categorical columns and a scatter plot of numerical columns

A screenshot of a data analysis

Description automatically generated

A screenshot of a computer program

Description automatically generated

A computer code with many text

Description automatically generated with medium confidence

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A group of graphs showing different types of graphs

Description automatically generated with medium confidence

A white box with red text

Description automatically generated

A graph of a rectangle with numbers

Description automatically generated

A black numbers on a white background

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A group of graphs with different colored lines

Description automatically generated

A graph with numbers and lines

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

A screenshot of a number

Description automatically generated

A group of diagrams with different colors

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

A black text on a white background

Description automatically generated

A graph showing a number of people

Description automatically generated with medium confidence

A pie chart with numbers and a number of people

Description automatically generated

### Observation:

* The dataset is almost evenly distributed among genders, with 676 Males (50.5%) and 662 Fenales (49.5%)

A graph of a number of people

Description automatically generated

A pie chart with numbers and a percentage

Description automatically generated

A black text on a white background

Description automatically generated

A graph of different colored bars

Description automatically generated

A pie chart with numbers and a few different colored circles

Description automatically generated

### Observation:

* All four regions are represented approximately evenly in the dataset.

A graph with different colored bars

Description automatically generated

A pie chart with numbers and a few children

Description automatically generated

### Observation:

* In the dataset, approximately 85% (1138 / 1338) of the insured have less than 3 children.

1. Perform data visualization using plots of feature vs feature

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A graph of different sizes and shapes

Description automatically generated with medium confidence

A partcularly interesting relationship between Insurance Premium Charges, BMI and Smoking status (Smoker / Non - smoker) can be seen in this graph.

A computer code with text

Description automatically generated with medium confidence

A graph of colored dots

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a chart

Description automatically generated

### Observation:

* From the correlation heatmap, we can conclude that the premium charges show a weak positive correlation with Age and BMI of the insured, and a strong positive correlation with smoking habit.

1. Check if the number of premium charges for smokers or non-smokers is increasing as they are aging.

A screenshot of a computer

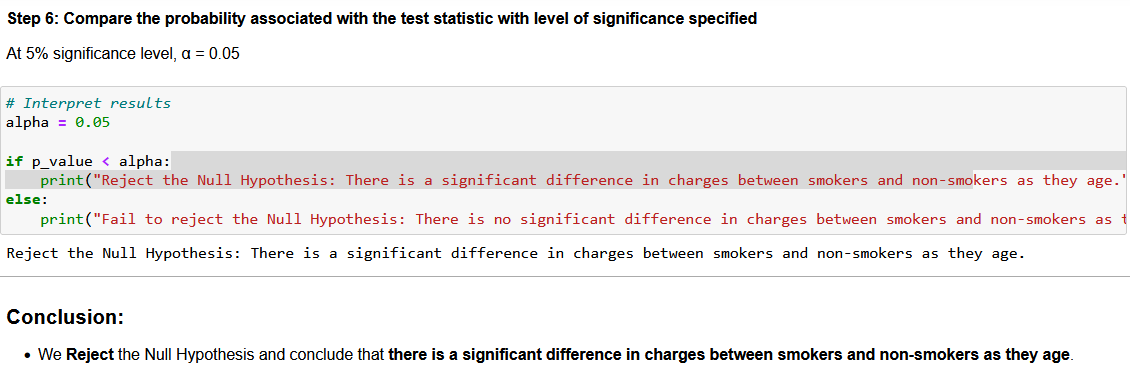
Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated



1. After each step, specify the observations