**Insurance Data Analysis**

**HETAL SANGANI**

Data Analyst

Python

**Table of Contents:**

* Understanding Business Problem
* Variable Description
* Data Understanding
* Data Cleaning
* Data Visualization

**Understanding Business Problem:**

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.

**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

**Domain:** Healthcare

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

**Variable 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 |

**Data Understanding:**

importing necessary library required and understanding the dataset

A screenshot of a computer code

Description automatically generated

A screenshot of a graph

Description automatically generated

df.head() returns the first 5 rows of the dataset and df.tail() to get last 5 rows.

A screenshot of a computer

Description automatically generated

df.info() method returns information about the whole DataFrame including the index data type and columns, non-null values, and memory usage.

A screenshot of a data

Description automatically generated

df.describe() method generates descriptive statistics for us. For numeric data, the result’s index will include count, mean, std, min, max as well as lower, 50(median) and upper percentiles.

We can easily notice that the minimum age is 18 and the maximum age is 64. We can also see mean and median values of age are almost the same.

I also noticed that the maximum charge value is 63770.42 which might be an unusual value. We can investigate this in the data visualization part.

A screenshot of a computer

Description automatically generated

We see that the most frequent value for sex is male and shown in the dataset 676 times.

There are 4 unique regions in our dataset and the most frequent value is southeast and 364 times counted in the data.

Most people are not smokers with a number of 1064 observations.

**Data Cleaning:**

In this part of the EDA. We will check :

* Missing Values
* Duplicated Values

The purpose of data cleaning is to get our data ready to analyze and visualize.

A screenshot of a computer code

Description automatically generated

When combining .isnull() method with .sum() we can sum up all the missing values for each column.

There are no missing values in this dataset. We will now proceed to analyze the data, observe patterns, and identify outliers with the help of visualization methods.

Now we will check for duplicated values.

A screenshot of a computer

Description automatically generated

Dataset has only one duplicated row. We can simply drop this row using the drop\_duplicates() method.

A screenshot of a computer

Description automatically generated

**Data Visualization:**

Now, We understood our dataset in general and checked the missing values. We also deleted duplicated values from the data frame.

The next part is data visualization! We have to perform univariate, bivariate and multivariate analysis to see the distribution and relationship between variables.

We will use the seaborn library for statistical data visualization. Seaborn is a data visualization library based on matplotlib.

**Univariate Analysis**

The purpose of the univariate analysis is to understand the distribution of values for a single variable.

We will perform univariate analysis by using visualization techniques.

Univariate Analysis for Numerical Features

A screenshot of a computer

Description automatically generated

A graph with lines and text

Description automatically generated with medium confidence

Displot stands for distribution plot and shows us distribution and kernel density estimation by default.

The distribution plot shows us how our variable is distributed.

We see that our data looks like the right(positive) skewed. Most of the charges are between 0 – 10000 dollars.

A graph with blue bars

Description automatically generated

We see that most of the customers are 18–19 years old.

A graph of a diagram

Description automatically generated with medium confidence

BMI of people seems like a normal distribution. Most people have BMI between 27 -34 years old.

A screen shot of a graph

Description automatically generated

By using a boxplot, We can easily interpret our variable if it has outliers.

A screenshot of a computer

Description automatically generated

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

***Bivariate Analysis***

Bivariate analysis is the analysis of exactly two variables. We will use bivariate analysis to find relationships between two variables.

For bivariate analysis, we usually use boxplot(categorical vs numerical), scatterplot(numerical vs numerical).

A screen shot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

The boxplot shows us the cost of insurance for smokers is higher than for non-smokers.

Pair plot is another method that shows us the relationship between two numerical values as well as the distribution of each variable.

A screenshot of a graph

Description automatically generated

**Correlation**

Correlation is used to test relationships between quantitative variables or categorical variables. It’s a measure of how things are related. The heatmap() method shows us the relationship between numeric variables.

We will combine the .corr() method with heatmap so that we will be able to see the relationship in the graph.

A screenshot of a computer

Description automatically generated

The heatmap shows us there is a correlation between age and charges. As the age increases insurance charges also increase.

We can also see that there is a weak correlation between BMI and charges.

**Conclusion**

The main purpose of EDA is to help understand data before making any assumptions. EDA helps us to see distribution, summary statistics, relationships between variables and outliers.