Insurance Data Analysis Project

1. Introduction

ABC Insurance seeks to better understand the factors influencing medical insurance charges among its policyholders. Using a dataset containing demographic and health-related information, this analysis aims to uncover key drivers of premium costs and provide actionable business insights.

---

2. Objective

The primary objective is to analyse the insurance dataset to identify patterns and relationships between customer attributes (such as age, sex, BMI, children, smoker status, and region) and insurance charges. The findings will help inform risk assessment and pricing strategies.

---

3. Dataset Description

The dataset (insurance.csv) includes the following columns:

- **age:** Age of the insured individual
- **sex:** Gender (male/female)
- **bmi:** Body Mass Index
- **children:** Number of dependent children covered by insurance
- **smoker:** Smoking status (yes/no)
- **region:** Residential region in the US (northeast, northwest, southeast, southwest)
- **charges:** Annual medical insurance charges (target variable)

---

4. Data Preparation & Cleaning

- Loaded the dataset using pandas.
- Inspected data types and checked for missing or undefined values.
- Found some rows with undefined or missing values; these were removed to ensure data quality.
- Converted relevant columns to appropriate numeric types.

- Checked for and removed duplicate entries.
- Confirmed that categorical variables (sex, smoker, region) were properly formatted for analysis.

---

5. Exploratory Data Analysis (EDA)

Univariate Analysis

- **Demographics:** The dataset is balanced across sexes and regions. Most policyholders are non-smokers, and the majority have 0–2 children.
- **Distributions:**
- **Age** is fairly uniform among adults, with a slight peak at age 18.
- **BMI** follows a near-normal distribution cantered around 30.
- **Charges** are right-skewed, with most values below $15,000 but several outliers above $40,000.

Bivariate Analysis

- **Smoker Status:** Smokers pay dramatically higher premiums than non-smokers.
- **Sex:** No significant difference in charges between males and females.
- **Region:** Charges are similar across regions, with the Southeast showing slightly higher median charges.
- **Children:** Number of children does not significantly affect charges.
- **Age & Charges:** Charges increase with age, especially for smokers.
- **BMI & Charges:** Higher BMI is associated with higher charges, particularly for smokers.

Multivariate Analysis

- Pair plots and grouped boxplots show that smoking status dominates other features in determining charges.
- Age and BMI further increase charges, especially for smokers.
- Correlation heatmap confirms that smoker status has the strongest positive correlation with charges, followed by age and BMI.

6. Visualizations

A variety of plots were used to support the analysis:

- **Histograms** for age, BMI, children, and charges distributions.
- **Count plots** for categorical variables (sex, smoker, region).
- **Boxplots** comparing charges by sex, smoker status, region, and age group.
- **Scatter plots** for relationships between numerical features and charges.
- **Line plot** showing how charges change with age for smokers vs. non-smokers.
- **Pairplot** and **correlation heatmap** for multivariate exploration.

7. Observations and Insights

- **Smoking status is the single most important factor affecting insurance charges.** Smokers pay significantly more than non-smokers, regardless of age, BMI, or region.
- **Age and BMI also influence charges**, with older and higher-BMI individuals incurring higher costs—especially if they smoke.
- **Sex and number of children have negligible impact** on insurance charges.
- **Regional differences are minor**; while the Southeast shows slightly higher median charges, the effect is small compared to smoking, age, and BMI.
- **Most policyholders are non-smokers and have lower charges.** Outliers with very high charges are almost exclusively smokers, often older and/or with higher BMI.
- **Business implication:** Focusing on reducing smoking rates among policyholders could yield substantial cost savings for both the insurer and insured. Risk-adjusted pricing should prioritize smoking status, age, and BMI.