**Medical Costs Analysis — Exploratory Data Analysis, Preprocessing, and Insights**

**Dataset:** Medical Cost Personal Datasets (insurance.csv)

**Planned idea and goal:**

The main idea of this project is analyzing medical insurance costs based on various factors such as age, gender, BMI, number of dependents, smoking habits, and residential region. The goal is to explore the “Medical Cost Personal Datasets” to uncover patterns and insights that could inform pricing strategies, risk assessment, and personalized health care planning.

**Executive Summary**

This report summarizes an explOratory data analysis (EDA) and preprocessing pipeline performed on the Medical Cost Personal dataset. The objective was to understand how demographic and health-related features (age, sex, BMI, children, smoking status, region) relate to annual insurance charges.

**Our Solution and Technology**

We analyzed the **Medical Cost Personal Dataset** using advanced data analytics and preprocessing techniques. The solution leverages **Python (Pandas, NumPy, Matplotlib)** in order to:

1. **Explore the dataset (EDA):** Identified cost drivers and patterns.
2. **Preprocess the data:** Encoded categorical features (sex, smoker, region), standardized continuous variables (age, BMI, charges).
3. **Visualization:** Created clear charts to communicate insights (e.g., smokers’ charges vs non-smokers, BMI impact by smoker status).

**1. Data Overview**

* **Rows:** 1,338
* **Columns:** 7 — age, sex, bmi, children, smoker, region, charges
* **Types:** numeric (age, bmi, children, charges), categorical (sex, smoker, region)
* **Missing values:** none detected (0 missing entries in all columns)

The dataset is clean in terms of completeness but contains categorical text columns that require encoding before modeling.

**2. Exploratory Data Analysis (EDA)**

**2.1 Descriptive statistics (numerical)**

* age: distribution across adult ages, mean and spread showing typical insurance-age distribution.
* bmi: distribution with many observations in the overweight range; presence of higher BMI outliers.
* charges: strongly right-skewed with a small set of high-cost outliers (large medical expenses).

**Visualization**

* A **histogram for each numeric column** (age, bmi, charges).
* Helps you **see the distribution**:
  + Is age evenly distributed or skewed?
  + Does bmi look normal or have outliers?
  + Are charges concentrated at lower values or spread out?

**2.2 Categorical summaries**

* sex: roughly balanced between male and female.
* smoker: minority are smokers, but they account for a disproportionate share of high charges.
* region: four categorical regions with similar row counts (northeast, northwest, southeast, southwest).

**What this boxplot shows**

* If smokers have **higher median charges** than non-smokers.
* How **spread out** the charges are in each group.
* Whether there are **extreme outliers** (e.g., very expensive cases).

**What this scatterplot shows**

* The relationship between **BMI** and **medical costs**.
* **Smokers (red)** cluster at **much higher charges** compared to non-smokers (blue), especially for higher BMI.
* It helps detect **interaction effects** → e.g., being both overweight and a smoker may greatly increase charges.

**3.2 Encoding categorical variables**

* sex mapped to binary: male → 1, female → 0.
* smoker mapped to binary: yes → 1, no → 0.
* region converted via **one-hot encoding** (four columns region\_northeast, region\_northwest, region\_southeast, region\_southwest).

**2.4 Key EDA insights**

* **Smoking is the most dominant feature**: median and mean charges for smokers are substantially higher than for non-smokers; smokers are frequent among top charge outliers.
* **BMI matters more for smokers**: scatter plots show that as BMI increases, smokers' charges increase steeply, while non-smokers show a milder increase.
* **Age has a positive relationship** with charges, but less pronounced than smoking and BMI.
* **Children**: little direct visual correlation with charges; small positive or neutral effect.
* **Sex**: small differences after controlling other variables; not a primary driver.

**3. Data Preprocessing Steps**

These steps were implemented in the pipeline and saved to insurance\_processed.csv.

**3.1 Missing values**

* Checked for missing values. None were present.
* If missing values existed, plan: numeric → median imputation; categorical → mode imputation.

**3.2 Encoding categorical variables**

* sex mapped to binary: male → 1, female → 0.
* smoker mapped to binary: yes → 1, no → 0.
* region converted via **one-hot encoding** (four columns region\_northeast, region\_northwest, region\_southeast, region\_southwest).This command takes the **categorical column "region"** and turns it into several **binary columns** (0/1) — one for each region.

Rationale: many models require numeric inputs; one-hot preserves region-specific effects while avoiding implicit ordinal assumptions.

**4. Feature Exploration**

**4.1 Correlation analysis**

* After encoding, smoker had the highest positive correlation with charges.
* bmi and age had positive correlations too; regional dummies had mild correlations.

**5. Regional Analysis (Bonus)**

* Computed mean, median, and count of charges by region.
* Observed moderate regional variation in mean charges. Differences could reflect: regional pricing differences, demographic composition (age/BMI/smoker concentration), or sample variability.

**Key Features & Insights**

Key findings (high level):

* **Smoking status** is the strongest predictor of charges — smokers pay substantially higher average charges than non-smokers.
* BMI correlates with charges, particularly among smokers: higher BMI + smoker → much higher costs.
* **Age** shows moderate positive correlation with charges (older individuals tend to have higher charges), but its effect is smaller than smoker status and some regional patterns.
* **Region** shows small-to-moderate differences in mean charges; differences may reflect demographic or healthcare-cost differences in regions.

**Conclusion**

Our project demonstrates that **data-driven analysis of medical costs** is not only feasible but highly valuable for insurers and healthcare providers. By applying EDA, preprocessing, and modeling, we have identified key cost drivers, challenges, and actionable insights that can guide **business strategy, pricing, and prevention programs**.