**Healthcare Insurance**

This dataset contains information on the relationship between personal attributes (age, gender, BMI, family size, smoking habits), geographic factors, and their impact on medical insurance charges. It can be used to study how these features influence insurance costs and develop predictive models for estimating healthcare expenses.  
**Age**: The insured person's age.

**Sex**: Gender (male or female) of the insured.

**BMI** (Body Mass Index): A measure of body fat based on height and weight.

**Children**: The number of dependents covered.

**Smoker**: Whether the insured is a smoker (yes or no).

**Region**: The geographic area of coverage.

**Charges**: The medical insurance costs incurred by the insured person.

Beginner:

1. Descriptive Statistics: Calculate the average `bmi` and compare it between `male` and `female` groups.

2. Data Visualization: Create a histogram to visualize the distribution of `charges` across the dataset.

3. Data Grouping: Group the data by `region` and calculate the average number of `children` in each region.

4. Basic Filtering: Identify all records where `smoker` is 'yes' and `age` is less than 30.

5. Conditional Calculation: Compute the average `charges` for `smokers` versus `non-smokers`.

Intermediate:

1. Correlation Analysis: Determine if there's a correlation between `bmi` and `charges`. Does a higher BMI result in higher insurance charges?

2. Multivariate Analysis: Analyze how `smoker` status and `bmi` together influence insurance `charges`.

3. Data Cleaning: Handle any missing or outlier values in `bmi` or `charges`. What method did you choose and why?

4. Data Transformation: Create a new column `bmi\_category` based on BMI: 'Underweight', 'Normal', 'Overweight', and 'Obese'.

5. Exploratory Data Analysis (EDA): Conduct EDA on `children` and `charges` to identify trends and outliers. What can be inferred about insurance costs for families with different numbers of children?

Advanced:

1. Predictive Modeling: Build a linear regression model to predict `charges` based on all the other variables in the dataset. What is the performance of the model?

2. Clustering: Perform a cluster analysis to segment customers based on their features. What distinct segments can you identify?

3. Hypothesis Testing: Is there a statistically significant difference in average insurance costs between regions? Use an ANOVA test to validate this.

4. Feature Engineering: Create interaction terms for `smoker` status and `bmi`, and assess if they improve a predictive model for `charges`.

5. Decision Trees: Create a decision tree to understand the key factors that lead to higher insurance charges. What are the main splitting variables?