**Data Analysis of Heart Disease Dataset**

The **Heart Disease dataset** from the UCI Machine Learning Repository is a widely-used dataset in medical data analysis. It contains information about various health-related attributes of patients and whether they have heart disease or not.

**Features in the Heart Disease Dataset:**

The dataset consists of several columns (features), including:

1. **Age**: Age of the patient.
2. **Sex**: Gender of the patient (1 = male, 0 = female).
3. **ChestPainType**: Type of chest pain (4 categories: typical angina, atypical angina, non-anginal pain, asymptomatic).
4. **RestingBP**: Resting blood pressure.
5. **Cholesterol**: Serum cholesterol level (in mg/dl).
6. **FBS**: Fasting blood sugar (1 = true, 0 = false).
7. **RestingECG**: Resting electrocardiographic results (values like normal, ST-T wave abnormality, left ventricular hypertrophy).
8. **MaxHR**: Maximum heart rate achieved.
9. **ExerciseAngina**: Whether the patient has exercise-induced angina (1 = yes, 0 = no).
10. **Oldpeak**: Depression induced by exercise relative to rest.
11. **Slope**: Slope of the peak exercise ST segment (3 categories).
12. **Ca**: Number of major vessels colored by fluoroscopy (0-3).
13. **Thal**: Thalassemia (3 categories: normal, fixed defect, reversible defect).
14. **Target**: Whether the patient has heart disease (1 = yes, 0 = no).

**Task 1: Load and Inspect the Data (15-20 minutes)**

* Load the dataset into a Pandas DataFrame.
* Use basic inspection techniques:
  + df.head() to view the first few rows.
  + df.describe() to get statistical summaries.
  + df.info() to check for missing values and data types.
* Check the distribution of the target variable (Target).

**Task 2: Data Cleaning and Preprocessing (25-30 minutes)**

* **Handle Missing Data**:
  + Check for missing values (df.isnull().sum()).
  + If any columns have missing values, discuss strategies for handling them (e.g., filling with mean/median, or dropping rows).
* **Feature Engineering**:
  + Convert categorical variables like Sex, ChestPainType, RestingECG, etc., into numeric format using encoding (e.g., one-hot encoding or label encoding).
  + Normalize or standardize numerical columns if necessary (e.g., RestingBP, Cholesterol, MaxHR).
* **Create New Features**:
  + Example: Combine Oldpeak and Slope to create a new feature that represents "exercise-induced heart stress".

**Task 3: Exploratory Data Analysis (EDA) (30-40 minutes)**

* **Univariate Analysis**:
  + Plot the distribution of key numerical features (e.g., Age, Cholesterol, MaxHR) using histograms or boxplots.
  + Visualize the distribution of the target variable (Target), using a count plot.
* **Bivariate Analysis**:
  + Explore the relationship between the target variable (Target) and other features:
    - Use a **count plot** or **bar plot** for categorical features like Sex, ChestPainType, FBS, ExerciseAngina.
    - Use a **boxplot** or **violin plot** for numerical features like Age, Cholesterol, MaxHR to see how they relate to heart disease presence.
    - Correlation matrix to explore relationships between numeric features.
* **Visualizing correlations**:
  + Visualize correlations using a heatmap (sns.heatmap), focusing on relationships between features like cholesterol, age, and resting blood pressure.

**Task 4: Aggregation and Insights (20 minutes)**

* Use **groupby** to find the survival rate (presence of heart disease) by different categories:
  + Survival by gender (Male vs Female).
  + Survival by chest pain type (ChestPainType).
  + Survival by maximum heart rate achieved (MaxHR).
* Calculate **average cholesterol levels** for people with and without heart disease, and compare these across categories (e.g., ChestPainType).

**Task 5: Derive Medical Insights**

* Investigate the data set and come up with at least 5 different insights with proof
* Prepare a dashboard for your findings