# Heart Disease Dataset: Data Engineering and Analysis

## 1. Dataset Overview 🗃️

The dataset used is a structured CSV file containing patient heart health indicators such as:  
- Age, sex, chest pain type, resting blood pressure (trestbps)  
- Cholesterol level (chol), fasting blood sugar, ECG results  
- Maximum heart rate achieved (thalach), exercise-induced angina (exang)  
- ST depression (oldpeak), and target (presence of heart disease)

## 2. ECT: Extraction, Cleaning, Transformation ⚙️

### Data Cleaning 🧼

- Removed unnecessary/duplicate entries.  
- Checked and handled missing/null values.  
- Converted categorical values to human-readable labels (e.g., sex: 0 → Female, 1 → Male).

### Data Transformation 🔁

- Binning and categorization:  
 - age\_group (Young, Middle-aged, Senior)  
 - age\_decade (20s, 30s, ..., 70s)  
 - cholesterol levels (Desirable, Borderline High, High)  
 - blood pressure levels (Normal, Elevated, Hypertension Stages)  
 - oldpeak\_category (low, moderate, high)  
- Type conversions for binary classification (e.g., high\_cholesterol, high\_bp).

## 3. Feature Engineering 🧠

New features engineered to enrich analysis and modeling:

- age\_group: Categorical age bins  
- age\_decade: Decade-wise age  
- chol\_per\_age: Cholesterol normalized by age  
- risk\_score: Weighted combination of risk factors  
- exercise\_recovery: Recovery based on thalach and exang  
- age\_oldpeak: Interaction of age and oldpeak  
- high\_cholesterol, high\_bp: Flags for elevated conditions  
- heart\_rate\_reserve: 220 - age - thalach  
- max\_hr\_percent: thalach / (220 - age)  
- hr\_efficiency: heart\_rate\_reserve / (oldpeak + 1)  
- combined\_risk\_flag: True if multiple high-risk conditions  
- sex\_str: Human-readable sex label

## 4. Statistical Summary & Charts 📊

- Descriptive statistics (mean, median, mode, 25%, 50%, 70%, max) were exported.  
- Charts were created to visualize distributions and compare metrics.

## 5. Visualizations 📈

- Summary Bar Chart (mean, median, mode, percentiles)  
- Pie Charts:  
 - Sex distribution  
 - Age group distribution  
- Optional: Correlation heatmap and Streamlit dashboard

## 6. Output Directory 🧪

All outputs are stored in the dw/ directory:  
- heart\_engineered\_features\_extended\_v2.csv  
- heart\_summary\_stats\_extended.csv  
- pie\_sex\_distribution.png  
- pie\_age\_group\_distribution.png  
- plot\_mean\_median\_mode\_v2.png

## 7. Technologies Used 🚀

- Python 3  
- pandas, numpy  
- matplotlib, seaborn  
- os (file handling)

## 8. Next Steps 💡

- Deploy interactive Streamlit dashboard  
- Push data to cloud DW (e.g., Snowflake, BigQuery)  
- Train ML models with engineered features