**Exploratory Data Analysis Report**

**1. Dataset Overview**

The dataset under review is titled **Heart Attack Prediction Dataset** and is intended to analyze or predict the likelihood of a heart attack based on patient data.

* **Number of Rows**: *Provide the count of rows (e.g., 303 rows).*
* **Number of Columns**: *Provide the count of columns (e.g., 14 columns).*

**2. Column Information**

Each column represents a feature or variable in the dataset. Below is the description and type of each column:

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| age | Numerical | Age of the patient in years |
| sex | Categorical | Gender (1 = Male, 0 = Female) |
| cp | Categorical | Chest pain type (0: Typical angina, 1: Atypical angina) |
| trestbps | Numerical | Resting blood pressure in mm Hg |
| chol | Numerical | Serum cholesterol level in mg/dl |
| fbs | Categorical | Fasting blood sugar (>120 mg/dl, 1 = True, 0 = False) |
| restecg | Categorical | Resting electrocardiographic results |
| thalach | Numerical | Maximum heart rate achieved |
| exang | Categorical | Exercise-induced angina (1 = Yes, 0 = No) |
| oldpeak | Numerical | ST depression induced by exercise relative to rest |
| slope | Categorical | Slope of the peak exercise ST segment |
| ca | Categorical | Number of major vessels colored by fluoroscopy (0–3) |
| thal | Categorical | Thalassemia (3 = Normal, 6 = Fixed defect, 7 = Reversible defect) |
| target | Categorical | Predicted outcome (1 = Heart attack risk, 0 = No risk) |

**3. Summary Statistics**

This section summarizes numerical features to understand their distributions:

| **Column Name** | **Mean** | **Median** | **Std. Dev.** | **Min** | **Max** |
| --- | --- | --- | --- | --- | --- |
| age | *value* | *value* | *value* | *value* | *value* |
| chol | *value* | *value* | *value* | *value* | *value* |
| trestbps | *value* | *value* | *value* | *value* | *value* |

For categorical features, include counts of unique values:

| **Column Name** | **Value** | **Count** |
| --- | --- | --- |
| sex | 1 (Male) | *value* |
|  | 0 (Female) | *value* |

**4. Missing Values**

The table below lists columns with missing values:

| **Column Name** | **Missing Count** | **Percentage Missing** |
| --- | --- | --- |
| chol | 5 | 1.65% |
| thal | 3 | 0.99% |

Recommendation: Impute missing values using mean, median, or other imputation techniques.

**5. Correlation Analysis**

Using a correlation matrix, we analyze relationships between numerical columns, particularly focusing on how they relate to the target variable (target). A correlation heatmap is recommended for visual insights.

Example observations:

* thalach has a strong positive correlation with target.
* oldpeak is negatively correlated with target.

**6. Data Distribution**

* **Histograms**: Display distributions for numerical columns like age, chol, and trestbps.
* **Count Plots**: Show frequency distributions for categorical columns like sex, cp, and thal.

**7. Observations**

* The dataset appears imbalanced in gender, with more males than females.
* Cholesterol levels (chol) range significantly, requiring potential normalization.
* Certain features like ca and thal have low variance and may require investigation for impact on prediction.

**8. Next Steps**

To prepare the data for modeling:

1. **Handle Missing Values**: Impute missing values in chol and thal.
2. **Scale Data**: Normalize numerical features like chol, trestbps, and thalach.
3. **Feature Selection**: Analyze importance of features like cp, ca, and thal.
4. **Model Building**: Use algorithms such as Logistic Regression, Decision Trees, or Neural Networks.