Exploratory Data Analysis (EDA) **Report**

Overview

Dataset Name	UCI Heart Disease Data
Dataset source	• Kaggle link
Understanding the Dataset	This dataset contains medical records related to heart disease diagnosis. It includes patient demographics, clinical measurements, and test results. The primary goal is to analyze factors contributing to heart disease and develop predictive models for diagnosis.
important note	This dataset is widely used in machine learning research for heart disease classification. It consists of a subset of 14 key attributes extracted from a larger database of 76 features. The Cleveland dataset is the most commonly used for predictive modelling.

Dataset Structure	The dataset consists of key patient-related attributes categorized as follows: • **Patient Demographics:** Age, Sex, Origin (Place of Study) • **Medical Information:** Chest Pain Type (Typical Angina, Atypical Angina, Non-Anginal, Asymptomatic), Resting Blood Pressure, Serum Cholesterol, Fasting Blood Sugar, Resting ECG Results (Normal, ST-T Abnormality, LV Hypertrophy), Maximum Heart Rate Achieved, Exercise-Induced Angina, Oldpeak (ST Depression), Slope of Peak Exercise ST-Segment, Number of Major Vessels, Thalassemia (Normal, Fixed Defect, Reversible Defect) • **Target Variable:** num (Heart Disease Classification: 0, 1, 2, 3, 4)
Size	• Total rows: 920. • Total columns: 16.
Key Use Case	• A primary application is predicting heart disease severity (0, 1, 2, 3, 4) as a **multi-class classification** problem in healthcare analytics."

Main columns

ID	A unique identifier is assigned to each patient.
Age	The patient's age at the time of their hospital admission.
Origin	The location where the study was conducted.
Sex	The patient's gender is categorized as either "Male" or "Female."
Chest Pain Type (CP)	The type of chest pain experienced by the patient, categorized as "Typical Angina," "Atypical Angina," "Non-Anginal," or "Asymptomatic."
Resting Blood Pressure (Trestbps)	The patient's blood pressure (measured in mm Hg) at the time of hospital admission.
Serum Cholesterol (Chol)	The cholesterol level in the patient's blood is measured in mg/dL.
Fasting Blood Sugar (FBS)	Indicates whether the patient's fasting blood sugar level is greater than 120 mg/dL (True/False).
Resting Electrocardiographic Results (Restecg)	The results of the patient's electrocardiogram (ECG), are categorized as "Normal," "ST-T Abnormality," or "Left Ventricular Hypertrophy."

Maximum Heart Rate Achieved (Thalach)	The highest heart rate was reached by the patient during testing.
Exercise-Induced Angina (Exang)	Indicates whether the patient experienced angina due to exercise (True/False).
ST Depression (Oldpeak)	The degree of ST depression observed during exercise, measured relative to rest.
Slope of Peak Exercise ST Segment (Slope)	The slope pattern of the ST segment during peak exercise.
Number of Major Vessels (CA)	The count of major blood vessels (ranging from 0 to 3) that were colored by fluoroscopy.
Thalassemia (Thal)	A blood disorder classification, categorized as "Normal," "Fixed Defect," or "Reversible Defect."
Heart Disease Severity (Num)	The predicted classification of heart disease severity, ranging from 0 to 4

Statistical Analysis

ID	• Values range from 1 to 920, reflecting a logical sequential numbering. • No missing values (count = 920), which is excellent.
Age	• Ages range from 28 to 77 years, which makes sense. • The distribution appears normal, with most ages concentrated between 47 and 60 years. • No negative or unreasonable values.
Trestbps (Resting Blood Pressure)	• The minimum value is 0, which is not a valid blood pressure reading and likely represents incorrect data that needs cleaning. • Values range from 0 to 200, with a mean of 132 — mostly reasonable except for the zero values.
Chol (Cholesterol)	• The minimum value is 0, which is not a realistic cholesterol level and suggests some data cleaning is required. • Values range from 0 to 603, with a mean of 199 — the presence of zero values indicates some potential outliers.
Thalch (Maximum Heart Rate Achieved)	• Values range from 60 to 202, which is a reasonable range. • The distribution appears normal, with no negative or unreasonable values.
Oldpeak (ST Depression)	• There are negative values (min = -2.6), which do not make sense in this context and likely need cleaning. • Values range from -2.6 to 6.2, indicating the presence of some extreme values.

CA (Number of Major Vessels Colored by Fluoroscopy)	• There is a significant number of missing values (count = 309 out of 920), which needs to be addressed. • Values range from 0 to 3, which seems logical for this type of variable.
Num (Diagnosis of Heart Disease)	• No missing values. • Values range from 0 to 4, indicating varying degrees of disease diagnosis, and the distribution appears balanced

General Observations:

- *Some variables contain unrealistic values (like O for blood pressure and cholesterol, and negative values in Oldpeak), requiring data cleaning.*
- *The CA column has a significant number of missing values and will need a strategy for handling them.*
- *Apart from these issues, most variables show a reasonable distribution without major outliers.*

You can check the results of Descriptive Statistics from the notebook link

Missing Values

Columns with Missing Values:	
CA (Number of Major Vessels Colored by Fluoroscopy)	611 missing values — a significant amount, making up a large portion of the data. This will need a clear strategy for handling.
Thal	486 missing values — also a high percentage of missing data, requiring careful attention.
Slope	309 missing values — a notable amount that could affect analysis.
FBS (Fasting Blood Sugar)	90 missing values — not insignificant and may influence results.
Oldpeak (ST Depression)	62 missing values — important to address as this is a key numerical indicator.
Trestbps (Resting Blood Pressure)	59 missing values — critical because it's a measure of resting blood pressure.
Thalch (Maximum Heart Rate	55 missing values — a key measure of maximum heart rate, requiring attention.

Achieved)	
Exang	55 missing values — exercise-induced angina, also vital for analysis.
Chol	30 missing values — cholesterol levels are an important health metric, so missing data here matters.
Restecg	2 missing values — minimal, but still worth noting.

Observations:

- *The high number of missing values in CA and Thal may affect model performance and overall insights, so we'll need a strong imputation or removal strategy.*
- *We'll need to decide whether to fill missing values with means, medians, or use more advanced techniques, depending on the distribution and importance of each variable.*

You can check the results of Checking Missing values from the notebook link

Unique Values

The data shows some imbalance across certain categories, which could potentially affect the machine learning model's performance and prediction accuracy.

You can check the results of Checking Unique values & Countplots from the notebook link

Check Outlier

Trestbps (Resting Blood Pressure)	There are several outliers on the higher end of the graph, indicating some unusually high blood pressure readings that may need to be investigated or cleaned.
Chol (Cholesterol Levels):	The distribution shows a number of high-value outliers above 400, which could reflect extreme cholesterol levels. These values might need further analysis.
Thalach (Maximum Heart Rate Achieved)	The data is mostly well-distributed, with a few lower-end outliers that could represent unusually low heart rate measurements.
Oldpeak (ST Depressi)	There are several high-end outliers above 4, which may indicate significant deviations from normal results and need closer examination.
Ca (Number of Major Vessels Colored by Fluoroscopy)	The distribution shows clear outliers, particularly on the higher end with values above 3, suggesting rare cases of increased vessel count.
Num (Diagnosis of Heart Disease)	The data is fairly balanced but shows some skewness toward higher values. Outliers appear on the upper end, indicating more severe diagnoses.

You can check the results of Boxplot from the notebook link

Check Correlations

Age	Shows a noticeable correlation with the number of major vessels colored by fluoroscopy (Ca) at (O.37) and with ST depression (Oldpeak) at (O.26), indicating that increasing age may be associated with higher values in these measures. It also has a moderate correlation with heart disease diagnosis (Num) at (O.34).
Resting Blood Pressure (Trestbps)	Displays weak correlations with most variables, with a slight association with age (O.24), suggesting that blood pressure tends to increase with age.
Cholesterol Level (Chol)	Does not show strong correlations with other variables, with its highest association being with maximum heart rate (Thalach) at (O.24), indicating a minimal effect of cholesterol levels on heart performance.
Maximum Heart Rate Achieved (Thalach)	Has a moderate negative correlation with age (-O.37) and heart disease diagnosis (-O.37), reflecting that lower heart rate performance may be a potential indicator of cardiac issues.
ST Depression (Oldpeak)	Moderately correlates with the number of major vessels (Ca) at (O.28) and with heart disease diagnosis (Num) at (O.44), suggesting that higher ST depression levels may signal an increased likelihood of heart disease.

Number of Major Vessels (Ca)	Stands out as one of the most significant predictors of heart disease diagnosis, with a strong correlation of (O.52), highlighting its importance in assessing cardiac health.
Heart Disease Diagnosis (Num)	Strongly correlates with the number of major vessels (O.52) and ST depression (O.44), emphasizing the role of these factors in predicting the presence of heart disease.

Conclusion:

The data shows that the most influential factors in diagnosing heart disease are the number of major vessels colored and ST depression. There's also a notable relationship between maximum heart rate and age, reflecting the importance of these indicators in evaluating cardiac health.

You can check the results of Heatmap from the notebook link

Next Steps

Data Cleaning:

Investigate and address high-end outliers in variables like resting blood pressure (Trestbps), cholesterol (Chol), ST depression (Oldpeak), and the number of major vessels (Ca). Consider removing or transforming extreme values to improve model performance.

Feature Engineering:

Scale numerical features to ensure consistent data ranges, especially for variables like cholesterol, maximum heart rate, and ST depression.

Explore interactions between features like age, heart rate, and vessel count to create more informative predictors.

Categorical Data Transformation:

convert the categorical variables to numerical format using techniques like One-Hot Encoding or Label Encoding to prepare them for machine learning models.

Data Visualization:

Create clear and interactive visualizations to explore feature relationships and model outputs.

Use dashboards to display insights like risk factors, correlations, and prediction probabilities for heart disease diagnosis.

Data Preprocessing Report

1. Outlier Treatment

- *Outliers were detected in several columns, and due to the small dataset size, removing them wasn't an optimal choice. Instead, we applied capping using defined valid ranges.*
- *This approach ensured that extreme values were handled without losing valuable data.*

2. Missing Values Handling

We used different strategies based on the column types and data distribution:

- *Predictive Imputation (Random Forest): For columns with high missing values and potential bias: 'ca', 'slope', 'thal'*
- *Mode Imputation: For categorical columns with fewer missing values: 'restecg', 'exang', 'fbs'*
- *Median Imputation: For numerical columns with outliers: 'trestbps', 'chol', 'thalch', 'oldpeak'*

This mixed approach helped preserve the overall data distribution and reduced the risk of bias.

3. Encoding Categorical Variables

Both One-Hot Encoding and Label Encoding were used depending on the nature of the categorical features:

- *One-Hot Encoding: For unordered categorical variables:*
  - *'sex' (Male/Female)*
  - *'cp' (Chest pain type)*
  - *'restecg' (Resting ECG results)*
  - *'slope' (Slope of ST segment)*
  - *'thal' (Thalassemia types)*
- *Label Encoding: For binary categorical variables:*
  - *'fbs' (Fasting blood sugar)*
  - *'exang' (Exercise-induced angina)*

This ensured that the model could interpret the categorical data properly without introducing order where none exists.

4. Feature Scaling

We applied both Normalization and Standardization based on the distribution of each column:

Normalization: For non-normally distributed columns: 'num', 'ca'

Standardization: For normally distributed columns with outliers: 'age', 'trestbps', 'chol', 'thalch', 'oldpeak'

This approach balanced the feature scales and improved model performance.

5. Feature Selection

*The **id** and **dataset** columns were removed because having no predictive value. All remaining features were considered relevant and retained.*

6. Data Balancing

No data balancing techniques were applied yet, but an initial check of target class distribution ('num') suggested potential imbalance. We plan to address this using:

- *SMOTE (Synthetic Minority Over-sampling Technique) for increasing underrepresented classes*
- *Class weights adjustment in models like RandomForest or LogisticRegression*

This will ensure that the model does not favor the majority class and performs well across all categories.