BIG DATA FINAL PROJECT REPORT

Topic: Heart Disease Data

Group Members:

Aayush Sunilbhai Patel – 1265286

Harshil Lakhamanbhai Bavaliya – 1277378

Yash Mukeshkumar Panchal – 1274893

Table of Content

Project Overview:

This project performs an Exploratory Data Analysis (EDA) on a heart disease dataset using Python. The code identifies trends, relationships, and patterns in the dataset using statistical methods and data visualization tools. Additionally, the code outlines how to implement distributed data processing in Big Data environments using master and worker nodes.

Implementation of V's:

1. Volume:

- Loads and analyzes structured data of various sizes.
- Extracts information about dataset dimensions and memory usage.

2. Variety:

- Processes and visualizes a mix of categorical and numerical features (e.g., age, gender, chest pain type).

3. Velocity:

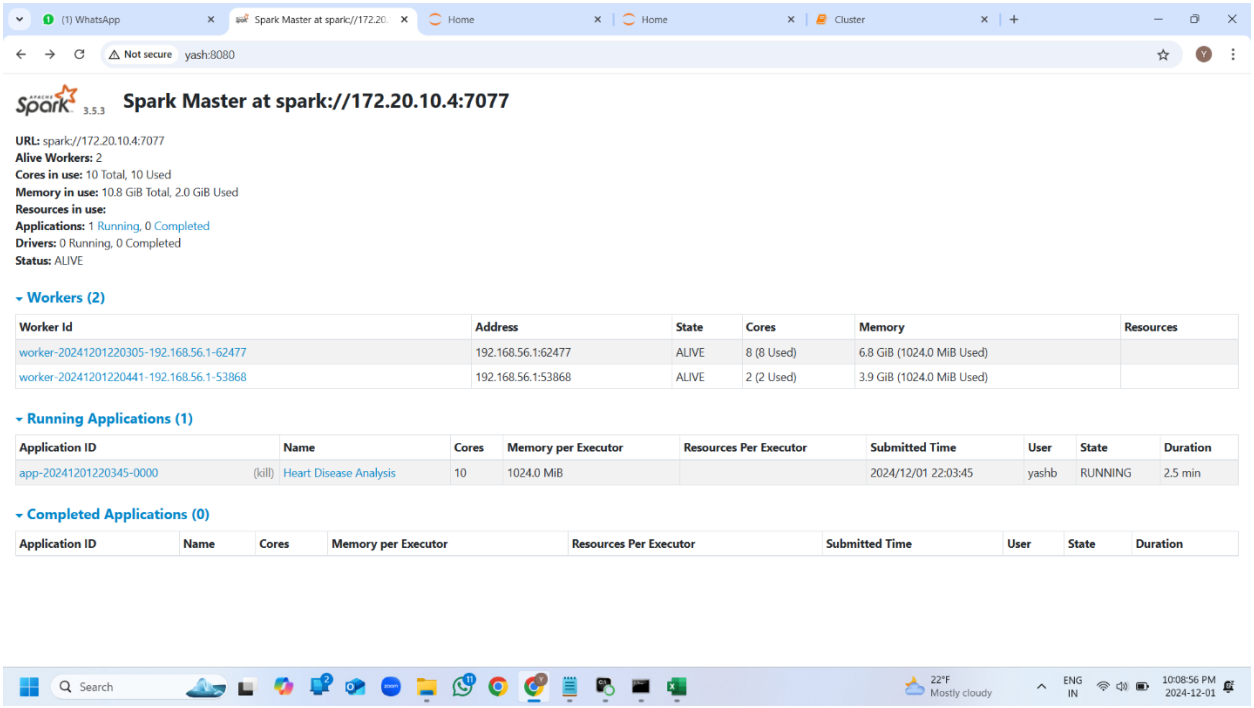
- Performs efficient statistical operations and generates insights quickly.

4. Veracity:

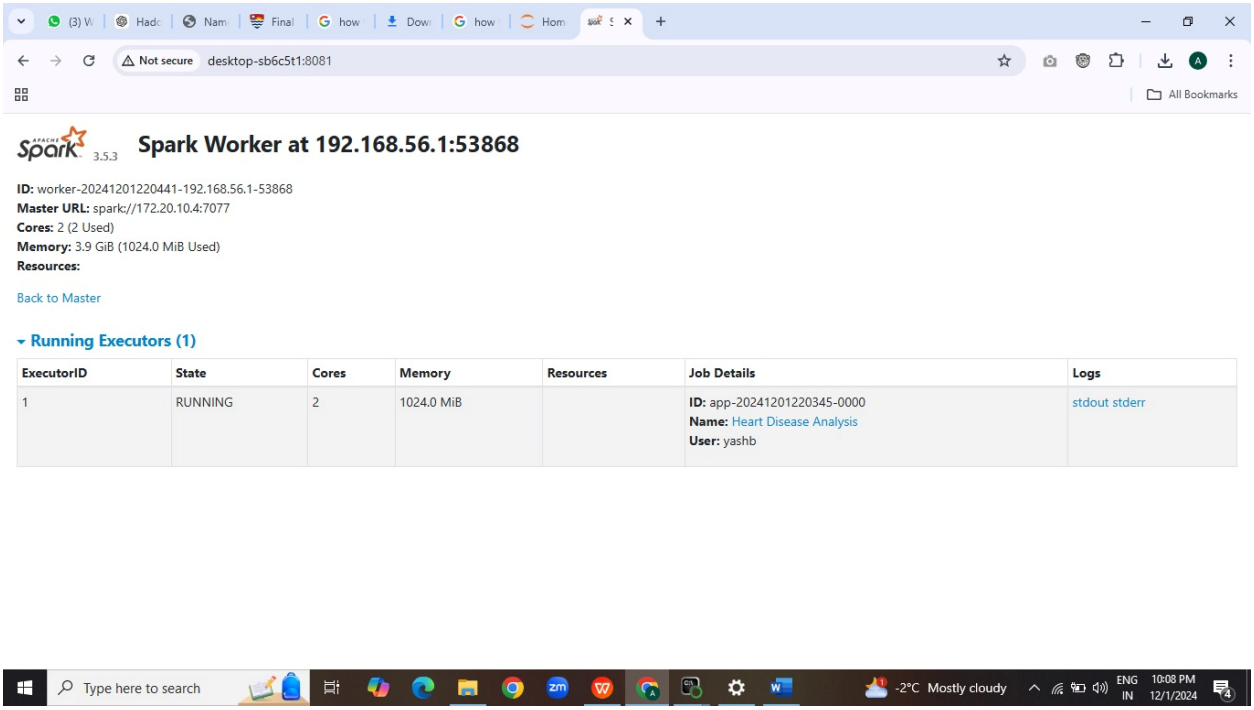
- Identifies missing values, ensures data quality, and visualizes correlations for trustworthy results.

Running Cluster's Screenshot:

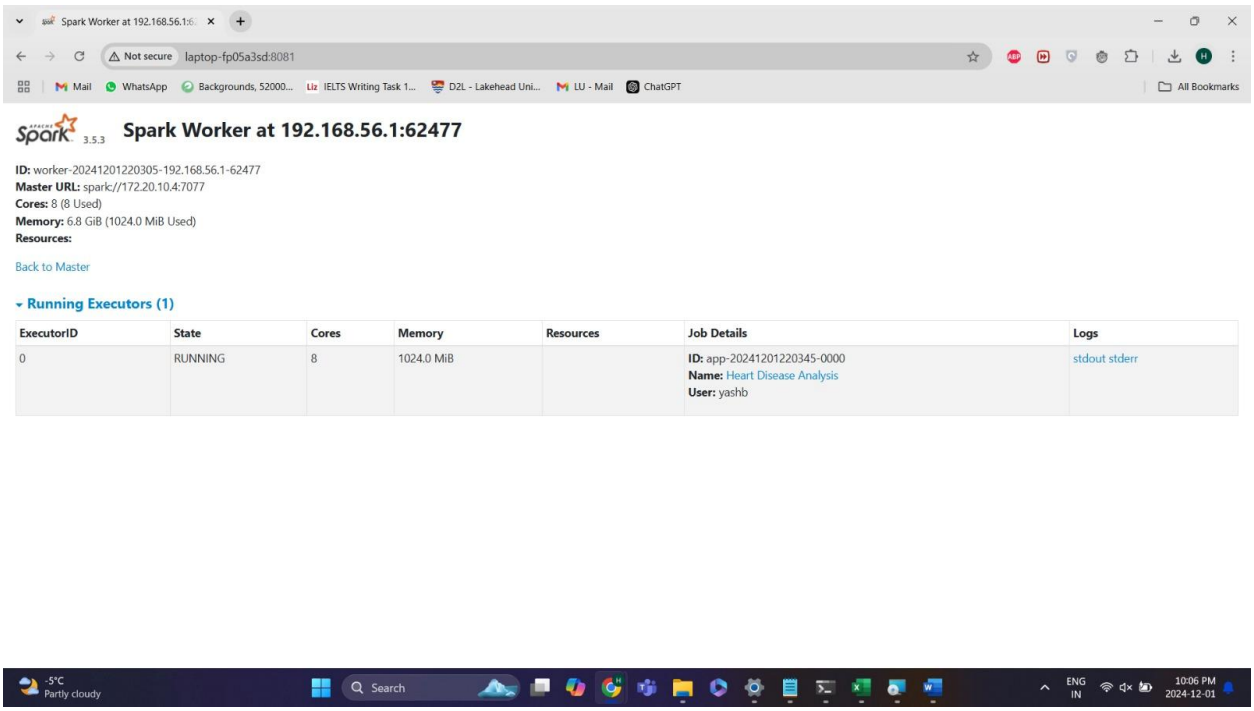
Master-Node:



Worker Node 1:

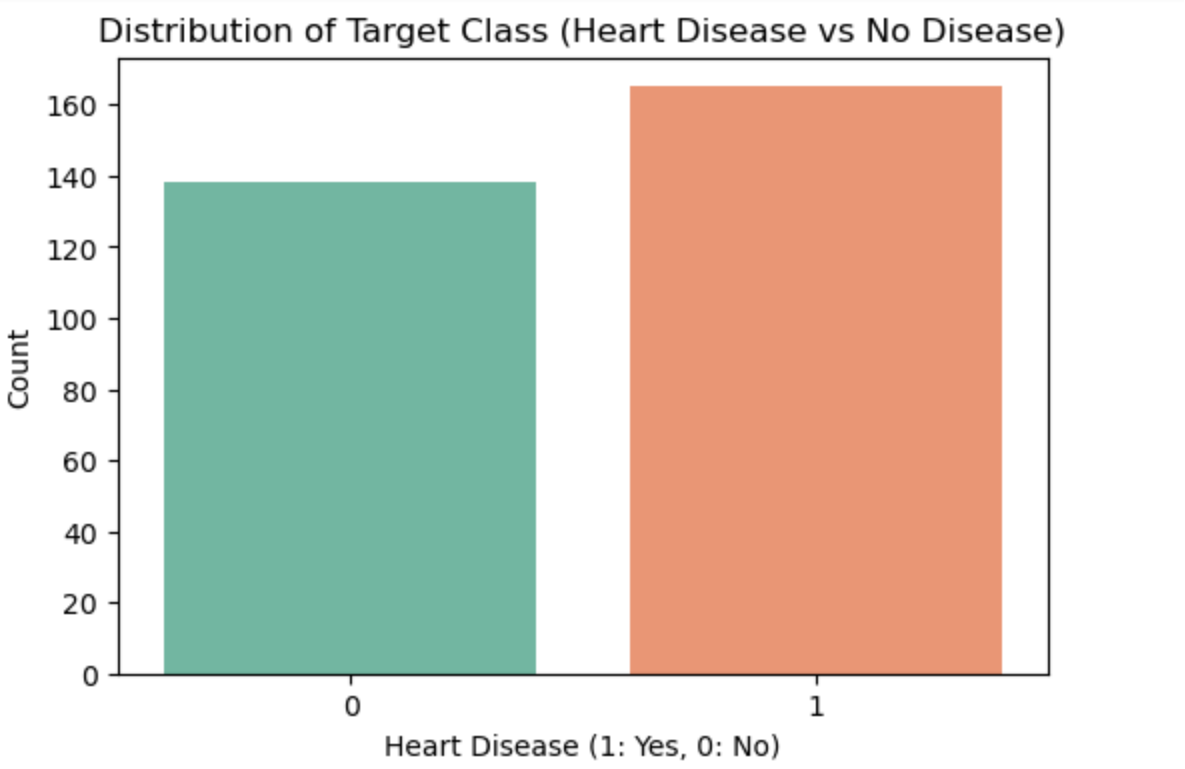


Worker Node 2:

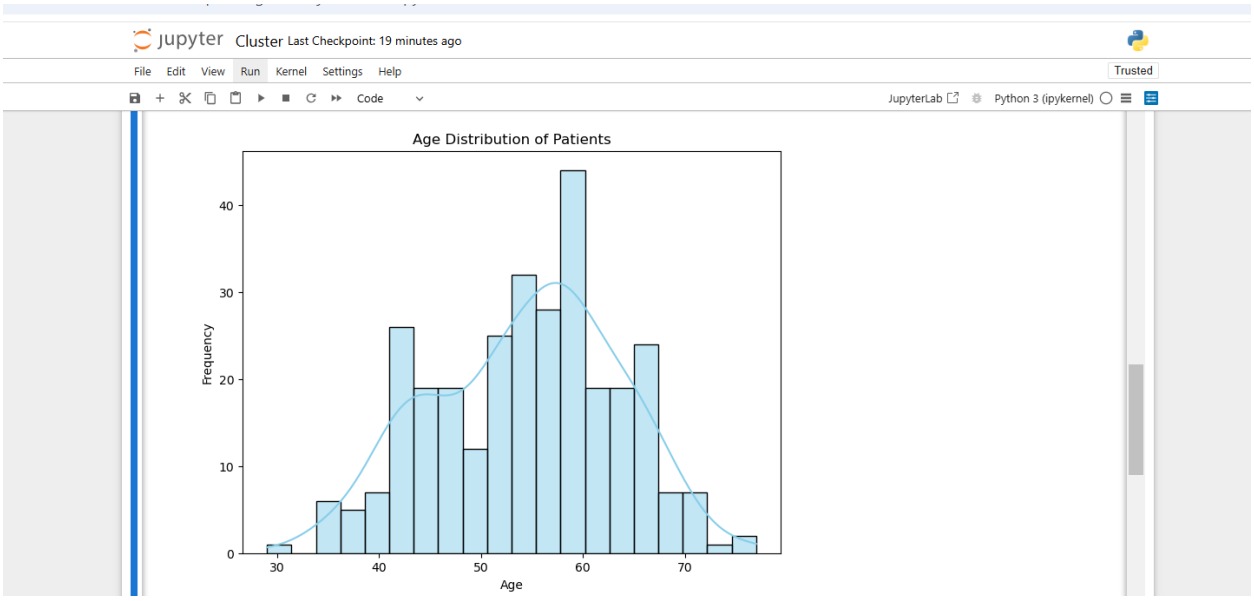


Execution Screenshot:

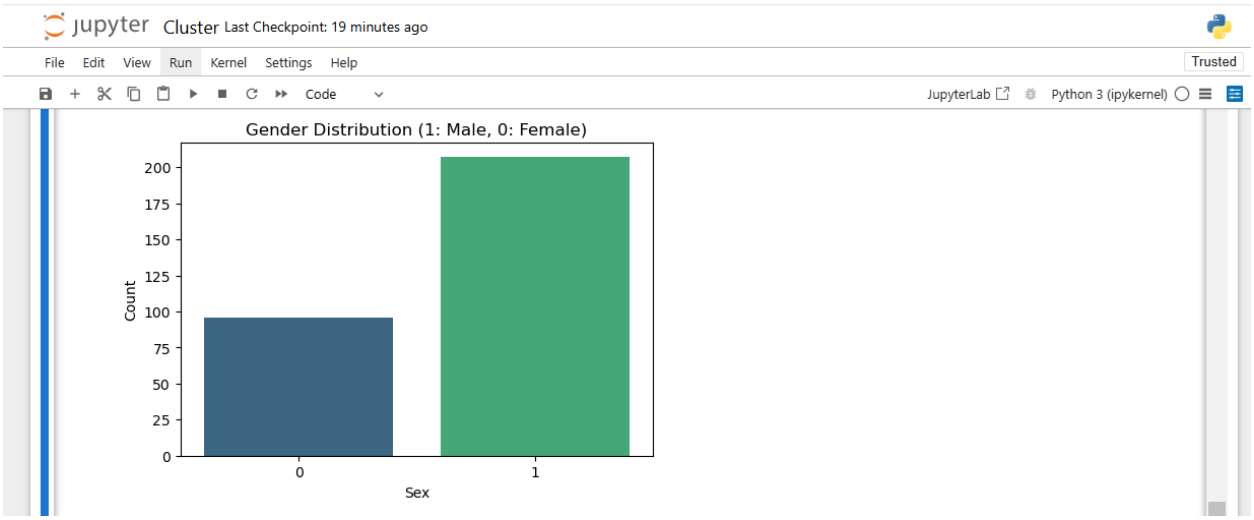
- **Distribution of Target Class:**



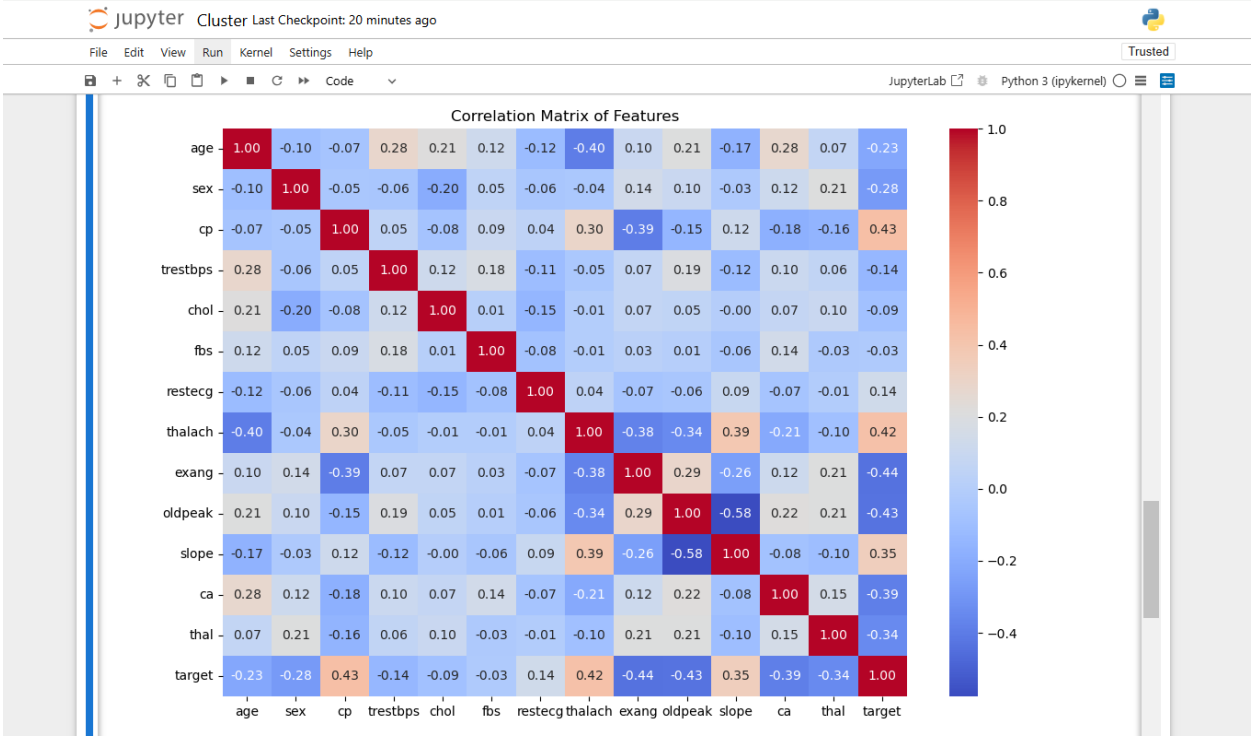
Age Distribution of Patient:



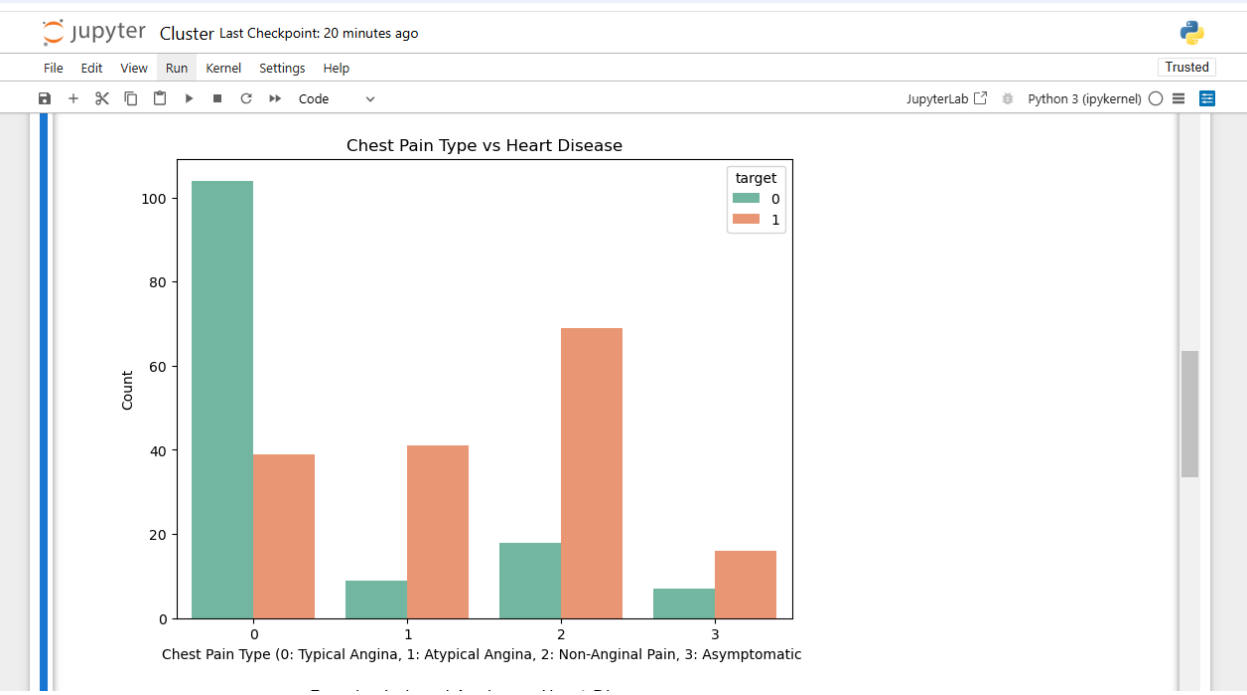
Gender Distribution:



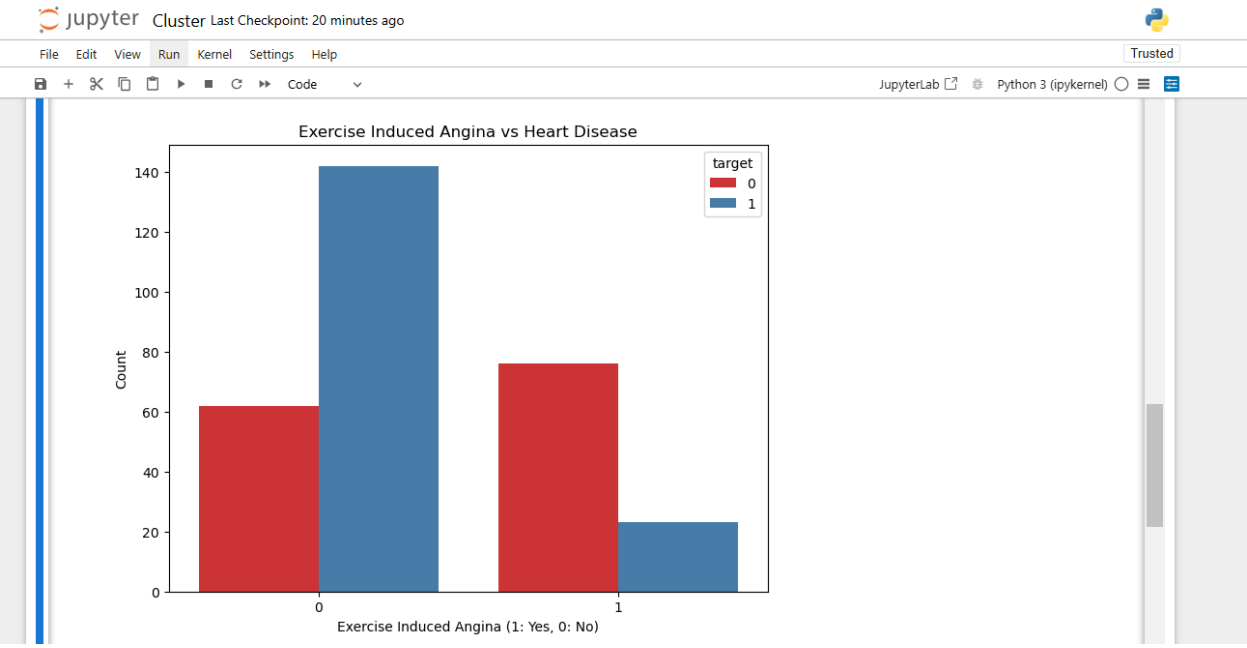
Matrix Features:



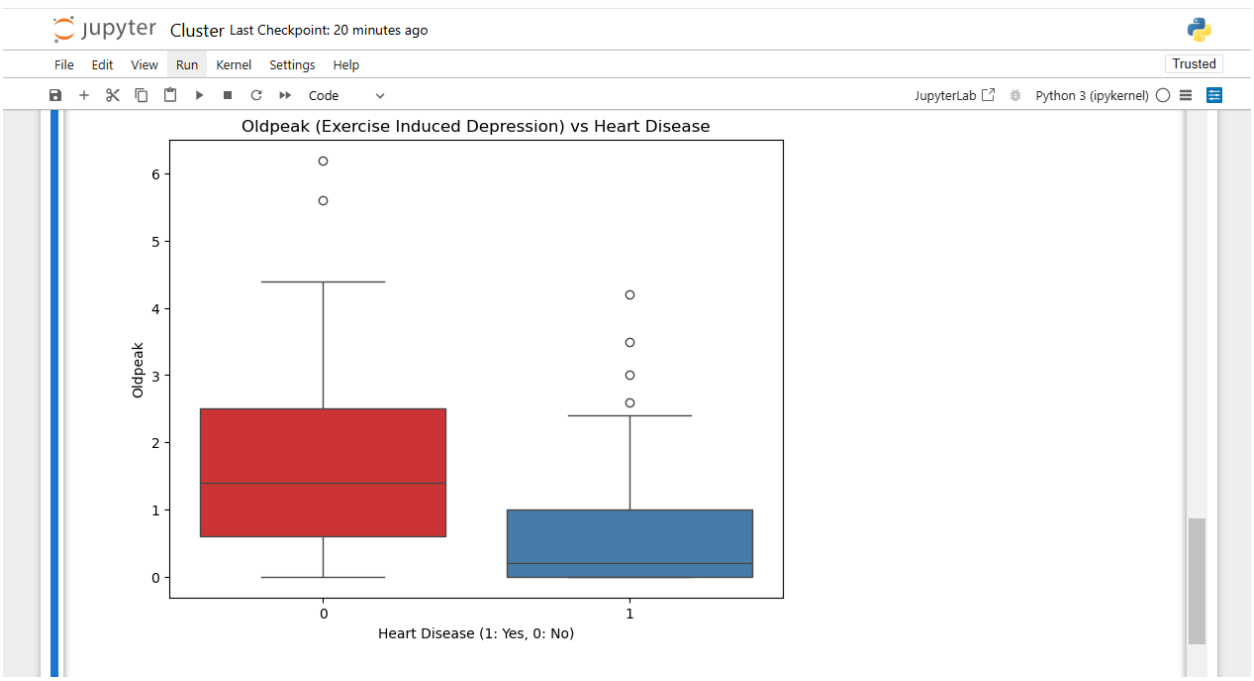
Chest Pain Type:



Exercise included :



OldPeak VS Heart Disease:



Conclusion:

This project analyzes key factors influencing heart disease using PySpark and Jupyter Notebook. Key findings include:

- Age and Gender: Middle-aged individuals, especially males, are at higher risk.
- Chest Pain and Angina: Certain chest pain types and exercise-induced angina significantly correlate with heart disease.
- Correlation Analysis: Features like thalach (maximum heart rate) and oldpeak are strong indicators of heart disease.
- Using PySpark enables efficient, scalable data processing for larger datasets. While the current analysis is insightful, future efforts could include predictive modeling, real-time data integration, and addressing dataset biases for broader applicability.