Healthcare Patient Data Analysis Report

Date: September 25, 2025

1. Introduction

This report details the analysis of a patient dataset related to heart health. The project utilizes PySpark for scalable data processing to extract key metrics and employs Python libraries such as Pandas, Matplotlib, and Seaborn for in-depth data visualization. The objective is to uncover patterns and insights from the patient data.

2. Summary of Key Metrics

- Total Number of Patients: 1025.
- Average Patient Age: 54.43 years.
- Patient Gender Distribution: Male (sex=1): 713 patients; Female (sex=0): 312 patients.
- Heart Disease Diagnosis: Patients with heart disease (target=1): 526; Patients without heart disease (target=0): 499.
- Patient Cholesterol Range: Maximum Cholesterol: 564; Minimum Cholesterol: 126.
- Patient Count by Age Range: 20-30: 4; 31-40: 64; 41-50: 247; 51-60: 438; 60+: 272.

3. Visual Analysis

3.1 Age Distribution of Patients

The histogram below shows the frequency distribution of patient ages. The data is most concentrated in the 51-60 age group.

3.2 Heart Disease by Gender

This count plot illustrates the number of heart disease cases broken down by gender. It compares the counts for patients with heart disease (target=1) and without (target=0) for both males and females.

3.3 Cholesterol Levels vs. Heart Disease

The box plot visualizes the distribution of cholesterol levels for patients with and without heart disease. This helps compare the median and spread of cholesterol for the two groups.

3.4 Average Maximum Heart Rate by Age Group

This bar plot displays the average maximum heart rate for different age groups, showing a clear trend of decreasing heart rate with increasing age.

3.5 Correlation Heatmap of Medical Features

The heatmap provides a visual summary of the correlations between different medical features in the dataset. Lighter shades indicate a stronger positive correlation.

4. High-Risk Patient Export

As part of the analysis, a subset of patients considered 'high-risk' was identified and exported to a file named HighRiskPatients.csv for further review.

5. Conclusion

The analysis provides a comprehensive overview of the patient dataset, revealing key trends in demographics and clinical metrics. Visualizations highlight the relationship between factors like age, gender, and cholesterol with the incidence of heart disease. The statistical summary and correlation analysis serve as a valuable foundation for future predictive modeling and more targeted clinical studies.