Module Summary: Heart Disease Data Analysis

Overview

This project analyzes the Heart Disease UCI dataset, which contains 1025 patient records with 14 clinical attributes related to heart disease diagnosis. The goal was to apply a reproducible data workflow -- including data ingestion, cleaning, exploratory analysis, and visualization -- to uncover patterns associated with heart disease. The dataset is publicly available on Kaggle (Heart Disease UCI Dataset).

Dataset Description

The Heart Disease UCI dataset represents clinical records of patients evaluated for heart disease. It contains 1025 rows and 14 columns, including demographic variables (age, sex), clinical measurements (resting blood pressure, cholesterol, maximum heart rate achieved), and diagnostic indicators (exercise-induced angina, ST depression, number of major vessels). The binary target variable indicates whether a patient has heart disease (1) or not (0). Key variables of focus in this analysis include age, cholesterol, maximum heart rate, and ST depression, as these are well-known risk factors in cardiovascular medicine. After cleaning, duplicate rows were removed, reducing the dataset to unique patient records.

Workflow Description

The data workflow followed a structured pipeline approach:

- Ingestion: The CSV file was loaded into a Pandas DataFrame and initial inspection (shape, data types, missing values) was performed to verify data integrity.
- Cleaning: Two cleaning functions were applied: (1) duplicate row removal to ensure data quality, and (2) column renaming to improve readability and interpretability.
- Exploratory Analysis: An EDA function generated summary statistics, group-level comparisons between heart disease and non-heart disease patients, and a correlation matrix.
- Visualizations: Three visualizations were created -- an age distribution histogram, a cholesterol vs. heart rate scatter plot, and a correlation heatmap.
- Summary: Findings were interpreted in context, noting patterns, limitations, and areas for further investigation.

Key Decisions and Assumptions

Cleaning choices: Duplicate removal was prioritized because the dataset contained a significant number of repeated rows, which could distort statistical summaries and model training. Column renaming was chosen to make the dataset self-documenting, following the principle that readable code and data improve reproducibility (Kazil et al., 2023). No imputation was needed since there were no missing values.

EDA focus: The exploratory analysis focused on comparing clinical measurements between patients with and without heart disease. This grouping directly addresses the clinical question of what distinguishes these populations. Correlation analysis was included to identify multicollinearity and potential predictive features.

Visualization design: Figure 1 (age distribution) was designed to show whether age is a distinguishing factor. Figure 2 (cholesterol vs. heart rate) explores the interaction between two key clinical measurements. Figure 3 (correlation heatmap) provides an overview of feature relationships. These visualization choices follow best practices for exploratory

data analysis (McKinney, 2022).

Results and Interpretation

The analysis revealed several notable patterns. As shown in Figure 1, heart disease patients span a broad age range but show a slightly different distribution compared to non-disease patients. Figure 2 demonstrates that patients with heart disease tend to achieve higher maximum heart rates, with some separation visible between the two groups. The correlation heatmap (Figure 3) confirmed that maximum heart rate has a positive correlation with the target, while ST depression, exercise-induced angina, and the number of major vessels show negative correlations.

Group-level analysis showed that patients with heart disease had a higher average maximum heart rate and lower average ST depression compared to those without. These patterns are consistent with established medical literature on heart disease risk factors.

Responsible Practice (Bias and Data Quality)

Several sources of potential bias exist in this dataset. The data was compiled from multiple clinical databases, and the presence of duplicate rows suggests possible merging artifacts. Removing duplicates could inadvertently remove legitimate repeat observations if patients were measured multiple times. The binary encoding of the target variable simplifies a complex clinical diagnosis, which may obscure nuances in disease severity. Additionally, the dataset lacks lifestyle variables (smoking, diet, exercise) that are significant confounders.

To reduce bias risk, duplicates were removed conservatively (keeping first occurrences), and the analysis presented findings as associations rather than causal claims. Further work should include examining data provenance and validating findings against external datasets.

Reproducibility

This project is designed to be fully reproducible. All code is contained in data_workflow.ipynb, and all dependencies are listed in requirements.txt. To reproduce the analysis: (1) clone the GitHub repository, (2) install dependencies using pip install -r requirements.txt, and (3) run all cells in the notebook in order. The Git workflow includes multiple commits tracking progress and at least one development branch, supporting version control best practices as recommended by Kazil et al. (2023).

References

Kazil, J., Jarmul, K., & Janssens, J. (2023). Reproducible Data Science with Python: An Open Learning Resource. Journal of Open Source Education, 6(65), 186. https://doi.org/10.21105/jose.00186

McKinney, W. (2022). Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter (3rd ed.). O'Reilly Media.