Heart Disease Data Visualization Using Matplotlib

May 8, 2025

Contents

1 Introduction

This document details a data visualization project using the heart.csv dataset to explore patient health metrics related to heart disease. The analysis employs Matplotlib to create histograms, pie charts, box plots, scatter plots, and a combined scatter plot with box plots. These visualizations aim to reveal distributions, proportions, relationships, and patterns in the data, particularly in relation to the presence or absence of heart disease.

The objective is to explain each visualization's purpose and insights clearly for an oral presentation. This document provides a structured explanation to facilitate understanding and effective communication of the analysis.

2 Dataset Overview

The heart.csv dataset contains 1025 patient records with 14 columns: 13 features and 1 target variable. Key features include:

- age: Age in years
- sex: Gender (1 = male, 0 = female)
- cp: Chest pain type (0–3)
- trestbps: Resting blood pressure (mm Hg)
- chol: Cholesterol level (mg/dl)
- fbs: Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
- restecg: Resting ECG results (0–2)
- thalach: Maximum heart rate achieved
- exang: Exercise-induced angina (1 = yes, 0 = no)
- oldpeak: ST depression induced by exercise
- target: Heart disease (1 = yes, 0 = no)

The visualizations focus on both continuous variables (e.g., age, chol) and categorical variables (e.g., sex, target).

3 Visualization Methods

The project includes five visualization types, each designed to highlight specific aspects of the dataset. The ggplot style is used for consistent aesthetics, and each plot is saved as a PNG file.

3.1 Histograms

Objective: Show the distribution of continuous variables.

Five continuous variables (age, trestbps, chol, thalach, oldpeak) are plotted as histograms with 30 bins. Each histogram is displayed in a 3x2 subplot grid, with the last subplot hidden. Key features include:

- Black-edged bars for clarity.
- Titles and labels for each variable (e.g., "Distribution of age").
- Tight layout to prevent overlap.

Insights: Histograms reveal the shape of distributions (e.g., age is roughly normal, oldpeak is right-skewed) and identify potential outliers or multimodal patterns.

3.2 Pie Charts

Objective: Display the proportion of categorical variables.

Six categorical variables (sex, cp, fbs, restecg, exang, target) are visualized as pie charts in a 2x3 subplot grid. Each chart shows:

- Proportions as percentages (e.g., 68.4% male for sex).
- Labels for each category.
- Titles indicating the variable (e.g., "Proportion of sex").

Insights: Pie charts highlight imbalances, such as a higher proportion of males or a balanced target variable (roughly 50% with heart disease).

3.3 Box Plots

Objective: Examine the spread and outliers of continuous variables by heart disease status.

Box plots for the five continuous variables are grouped by target (0 = no disease, 1 = disease) in a 3x2 subplot grid. Features include:

- Median, quartiles, and whiskers to show spread.
- Outliers as individual points.
- Labels for heart disease status and variable names.

Insights: Box plots reveal differences in distributions (e.g., thalach is higher for patients with heart disease) and identify outliers that may affect modeling.

3.4 Scatter Plots

Objective: Explore relationships between pairs of continuous variables.

Four scatter plots are created for the pairs (age, thalach), (trestbps, chol), (chol, thalach), and (age, oldpeak) in a 2x2 subplot grid. Each plot:

- Colors points by target using the viridis colormap.

- Includes a colorbar indicating heart disease status.
- Labels axes and titles (e.g., “thalach vs age”).

Insights: Scatter plots show relationships (e.g., `thalach` decreases with `age`) and how heart disease status correlates with variable pairs.

3.5 Scatter Plot with Box Plots

Objective: Combine distribution and relationship visualization for key variables.

A single figure combines a scatter plot of `age` vs. `thalach` with marginal box plots:

- **Scatter Plot**: Shows `age` vs. `thalach`, colored by `target`.
- **X-axis Box Plot**: Displays `age` distributions for no disease and disease groups.
- **Y-axis Box Plot**: Displays `thalach` distributions for both groups.
- Custom axes positioning to align scatter and box plots.

Insights: This plot integrates relationship (scatter) and distribution (box plots) information, highlighting how `thalach` is generally higher for patients with heart disease across age groups.

4 Results and Discussion

The visualizations provide valuable insights:

- **Histograms**: Show distributions, aiding in understanding variable ranges and skewness (e.g., `chol` has a long tail).
- **Pie Charts**: Reveal categorical imbalances, such as more males or specific chest pain types.
- **Box Plots**: Highlight differences in continuous variables by heart disease status, useful for feature selection in modeling.
- **Scatter Plots**: Indicate potential correlations and how heart disease relates to variable pairs.
- **Scatter with Box Plots**: Combines multiple perspectives, showing both relationships and distributions in one view.

These plots are critical for exploratory data analysis, guiding preprocessing and modeling decisions in a heart disease prediction task.

5 Conclusion

This project demonstrates the use of Matplotlib to create insightful visualizations for the heart disease dataset. Each visualization type serves a unique purpose, from understanding distributions to exploring relationships and categorical proportions. The combined scatter plot with box plots is particularly effective for integrating multiple insights. These techniques are essential for data exploration and can inform subsequent machine learning tasks.

Future work could include interactive visualizations or additional plots (e.g., correlation heatmaps) to further explore the dataset.

6 Presentation Tips for Oral Explanation

To deliver an effective oral presentation:

- **Introduce the Context**: Explain the heart disease dataset and the role of visualizations in understanding it.
- **Simplify Each Plot**: Describe histograms as "showing how data is spread out" and pie charts as "showing category proportions."
- **Use Visual Aids**: Display the saved PNG files or simplified diagrams to illustrate each plot type.
- **Highlight Key Insights**: Emphasize findings, like `thalach` differences in box plots or imbalances in pie charts.
- **Engage the Audience**: Ask questions like, "What might a skewed histogram tell us about the data?"
- **Be Concise**: Focus on one key purpose per visualization to keep the audience focused.
- **Practice Terminology**: Be comfortable with terms like "outliers," "distribution," and "colormap."
- **Conclude Strongly**: Summarize how visualizations guide data analysis and suggest future steps.