

Exploring Disease Symptom and Patient Profile Analysis

Aaryan Samanta
aaryan.samanta@gmail.com

Abstract

This paper features a structured analysis and reflection of the use of data science in healthcare, examining the "Disease Symptoms and Patient Profile Dataset." This paper aims to show a clear understanding of the structure of the datasets, consider what a dataset like this might look like in the world, identify some associated tools, and consider associated ethical issues. The analysis will include visualizations of key metrics such as age and disease among the patients in the dataset, along with an analysis of correlation to show relationships between the variables. The analysis presented shows how data can be used to achieve insights into public health patterns and support in the diagnostic process.

1 Introduction

This analysis will focus on the use of data analysis and machine learning to unravel the connections among patient profiles, symptoms and disease outcomes, a branch of work that is often referred to as health informatics. Informatic systems used in health care are important because they allow health professionals and researchers to detect patterns in patient data that may offer clues for earlier detection of illness, diagnosis, and treatment specificity. The data provided, which includes numerous diseases, symptoms, and patient demographics are an ideal way to practice those principles, but it is also a relevant topic within the overall context of AI as it illustrates a basic example of how supervised learning models should be trained to predict diseases in terms of a patient's symptoms and profile.

2 Data Cleaning & Exploratory Data Analysis (EDA)

The initial analysis of the dataset revealed several issues that required cleaning to ensure the integrity of our findings. The following steps were performed. First, duplicate records were removed to ensure each entry represents a unique patient, preventing skewed results in frequency and distribution analyses. Next, data types were converted. The Age column was converted to a numeric data type to allow for statistical calculations, with any non-numeric values being filled with 0. The Gender, Blood Pressure, and Cholesterol Level columns were converted to categorical data types, as they contained non-numeric text values that needed to be encoded for analysis. Lastly, missing values in the Age column were handled by filling them with 0. For Blood Pressure and Cholesterol Level, if they were of object type, they were converted into numerical codes using categorical encoding. This ensures the dataset is complete and ready for numerical operations.

After cleaning, a comprehensive EDA was conducted to understand the dataset's characteristics. Key descriptive statistics were calculated to summarize the central tendency and dispersion of the numerical data, including the mean, median, and standard deviation. As an

example, the median age of the patients was found to be in the mid-40s, with a standard deviation indicating a wide spread of ages within the patient population. Visualizations were also used to explore the data. The Age Distribution histogram (Figure 1) shows that the dataset contains a wide range of ages, with a relatively normal distribution centered around the middle-aged population. The Blood Pressure Distribution histogram (Figure 2) indicates a bimodal distribution, with concentrations at the low and high ends of the scale, and fewer patients in the middle range. The Age by Gender and Blood Pressure by Gender boxplots (Figure 3 and Figure 4) show the median and spread of these metrics across male and female patients, with Age showing a similar distribution between genders. The Top 20 Diseases chart (Figure 5) provides a clear view of the most frequent diseases, with Asthma, Osteoporosis, and Stroke being the most common in the dataset. The Correlation Heatmap (Figure 6) provides a quantitative summary of the relationships between numerical features, with a positive correlation of 0.31 between Blood Pressure and Cholesterol Level suggesting a moderate relationship.

3 Real-World Applications

Example 1: Clinical Support Systems for Diagnosis One important real-life application of this type of data is for the development of a clinical support system for diagnosis. In a hospital setting, a clinician support system could assess a patient's age and onward blood pressure (see the distributions in Figure 1 and Figure 2) to assess a preliminary diagnosis of the patient. The Age by Gender and Blood Pressure by Gender boxplots (Figure 3 and Figure 4) also provide valuable context as you can visualize how each of these key variables are distributed by patients. Clinical support systems should not be viewed as a replacement for physician expertise, rather, this serves as a strong second opinion to clinical expertise, where you can reduce time to diagnosis and possibly identify rare diseases that a clinician might overlook.

Example 2: Analyze Trends in Public Health. The second important use is effectively analyzing trends in public health. When considering diseases, government health agencies may have the same data, especially if it's been aggregated from multiple hospitals or locations, to track the spread of diseases. One way to analyze might be to look at the frequency of symptoms (e.g., Fever and Cough), Age and Location (if available), and then report could signal, at least, a localized outbreak of a disease like influenza or even some novel virus. A simple but excellent example would be to examine the Top 20 Diseases plot (Figure 5) that reports the most frequently reported diseases. This could help facilitate actions that could limit the spread of an epidemic, including distributing medical surge capacity with limited resources, public awareness campaigns, and executing specific vaccination programs.

4 Tools, Libraries, or Systems

The Python programming language and its libraries represent a central tool for analyzing this kind of data. Particularly critical are the Pandas and Matplotlib/Seaborn libraries. Pandas is an open-source library designed for data manipulation and analysis, and provides powerful data structures like the DataFrame for working with structured data, such as our dataset. Data cleaning and preprocessing - dropping duplicates and changing data types, for example - can all

be done efficiently in Pandas. Matplotlib and Seaborn are used for data visualizations, as they allow us to make useful plots including the Age Distribution histogram (Figure 1), Blood Pressure Distribution histogram (Figure 2), Age by Gender and Blood Pressure by Gender boxplots (Figure 3 and Figure 4), and the bar chart of the Top 20 Diseases (Figure 5). In particular, we can use the Correlation Heatmap (Figure 6) to visualize the correlation coefficients for all numerical variables - a great feature that allows us to quickly see the relationships between variables, such as the positive correlation of 0.31 between Blood Pressure and Cholesterol Level.

5 Ethical Consideration

A major ethical issue when using a dataset such as this (no matter the reviewer location) is data privacy and data security, which are both very real issues. Patient data, many of the demographic(s), symptoms, and health outcomes all fall under very sensitive data. If patient data is mishandled, breached, or poorly anonymized, data privacy could far exceed any reasonable expectation. The data collected could be used to injure a patient in accessing insurance coverage or employment, for example. Appropriate data governance procedures have to be used to prevent this type of behavior. Data anonymization is one aspect of data governance, which is removing or obfuscating personally identifiable data (e.g., (patient) names and/or exact addresses). Data encryption also governs data privacy by preventing data access in the event of a data breach. Data governance must govern the management of sensitive health data, including only using secure and compliant databases (e.g., databases compliant with HIPAA in the US).

6 Conclusion

To summarize, the work on the "Disease Symptoms and Patient Profile Dataset" illustrates how data science can be a powerful force in healthcare. Through the application of Python libraries such as Pandas and Seaborn, we were able to clean the data, visualize important distributions and analyze the relationships between multiple health metrics. For instance, we used visualizations to examine the correlation between diseases and age distribution of patients, and used a correlation heatmap to show numerical features and their relationships at a glance, Age and Blood Pressure. The biggest surprise by far was the range of diseases, which further addressed the complexity of making diagnostics and how the structured data contributed to incorporating the complexity of diagnostics. Moving forward, this will be an enormous focus for the future of AI. As more longitudinal datasets (with appropriate privacy features) become available, we will be seeing AI-enabled diagnostic tools that are more accurate and incorporated into regular practice. This will improve outcomes for patients and provide more equitable access to quality advice from physicians in resource limited environments.

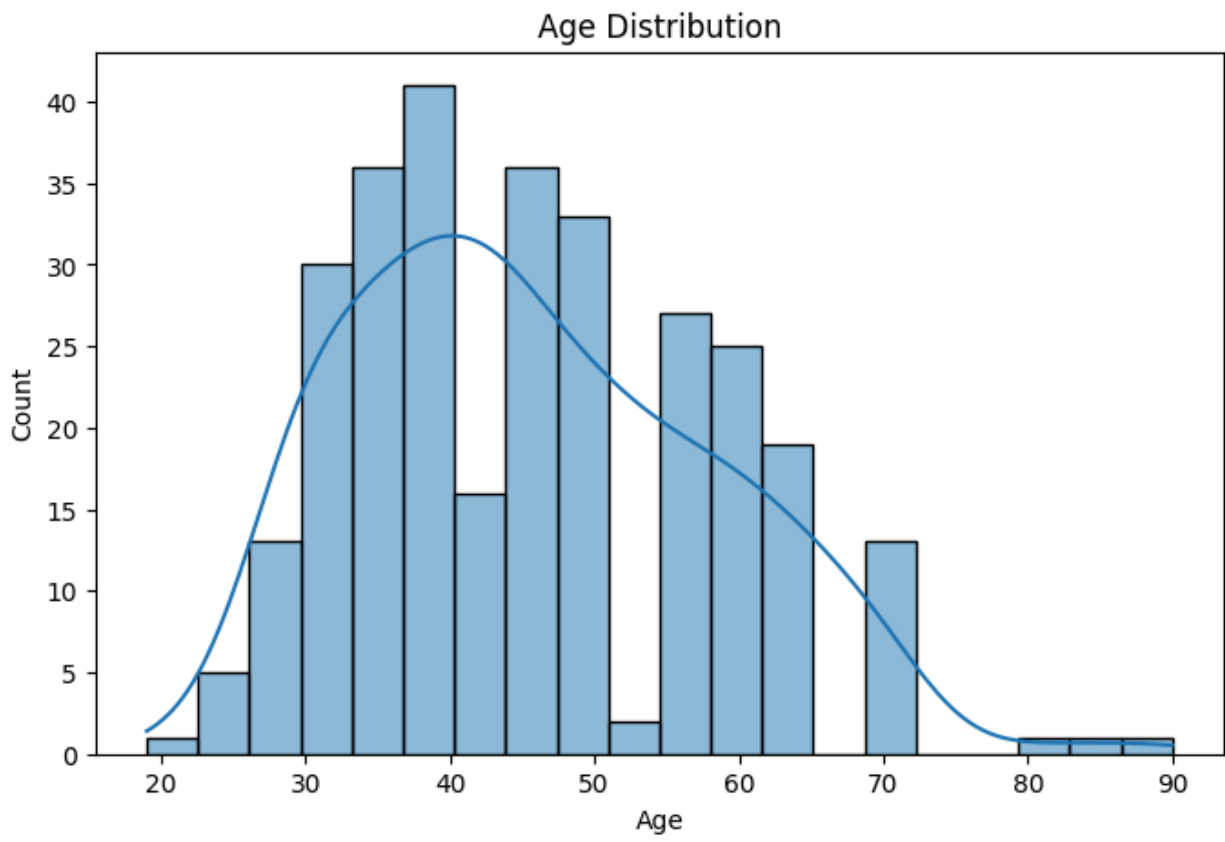


Figure 1

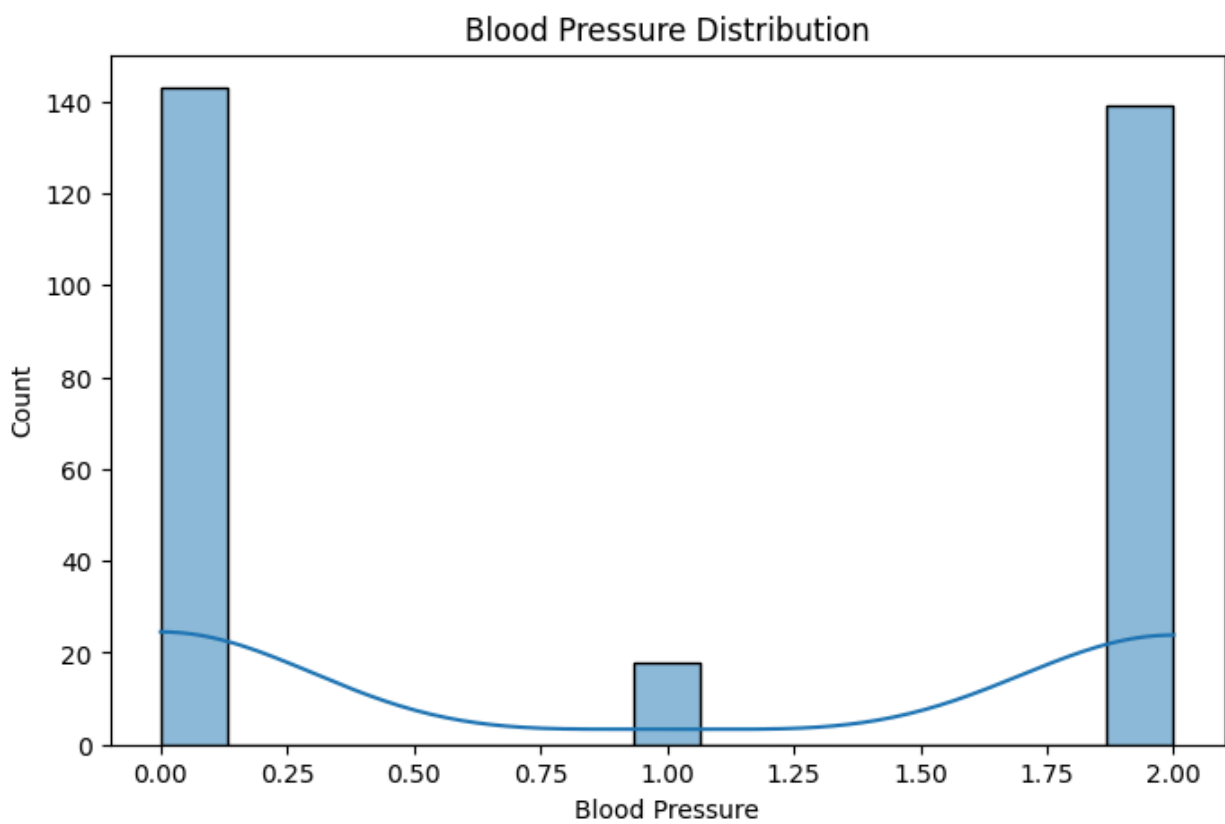


Figure 2

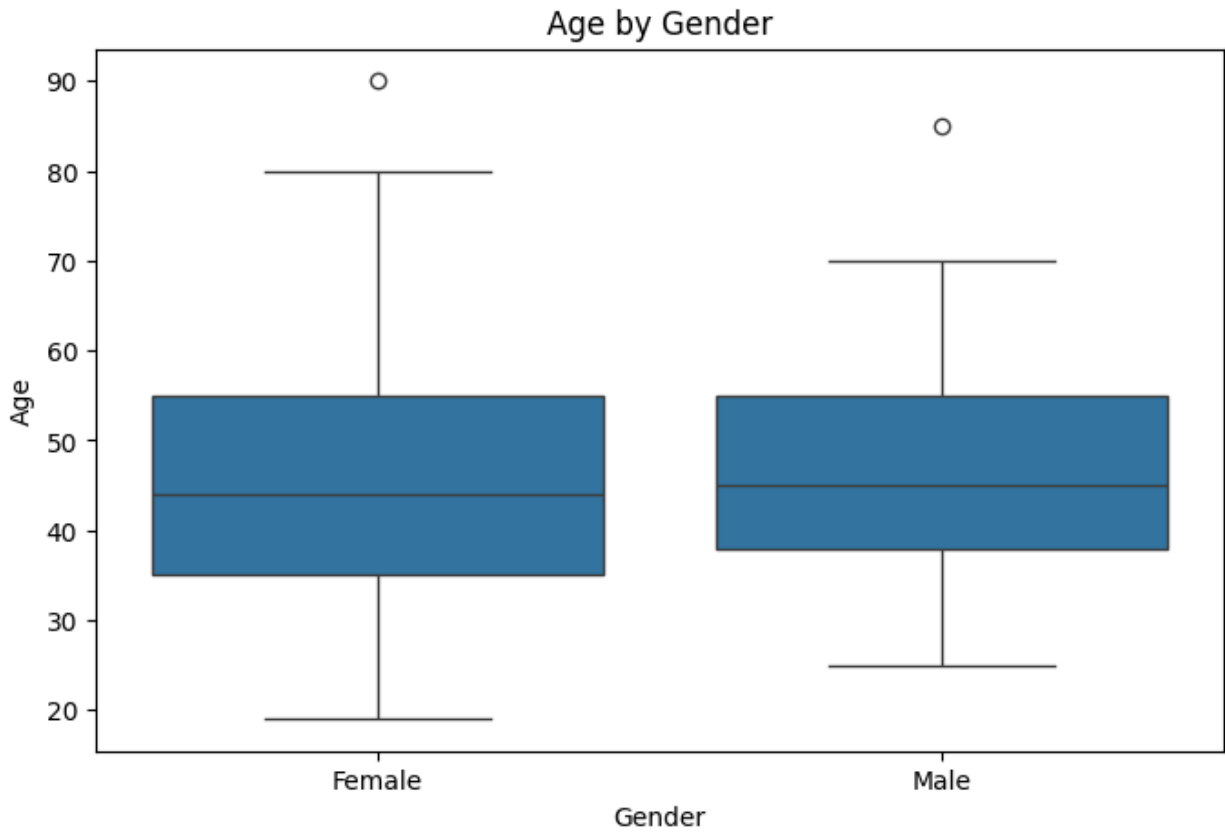


Figure 3

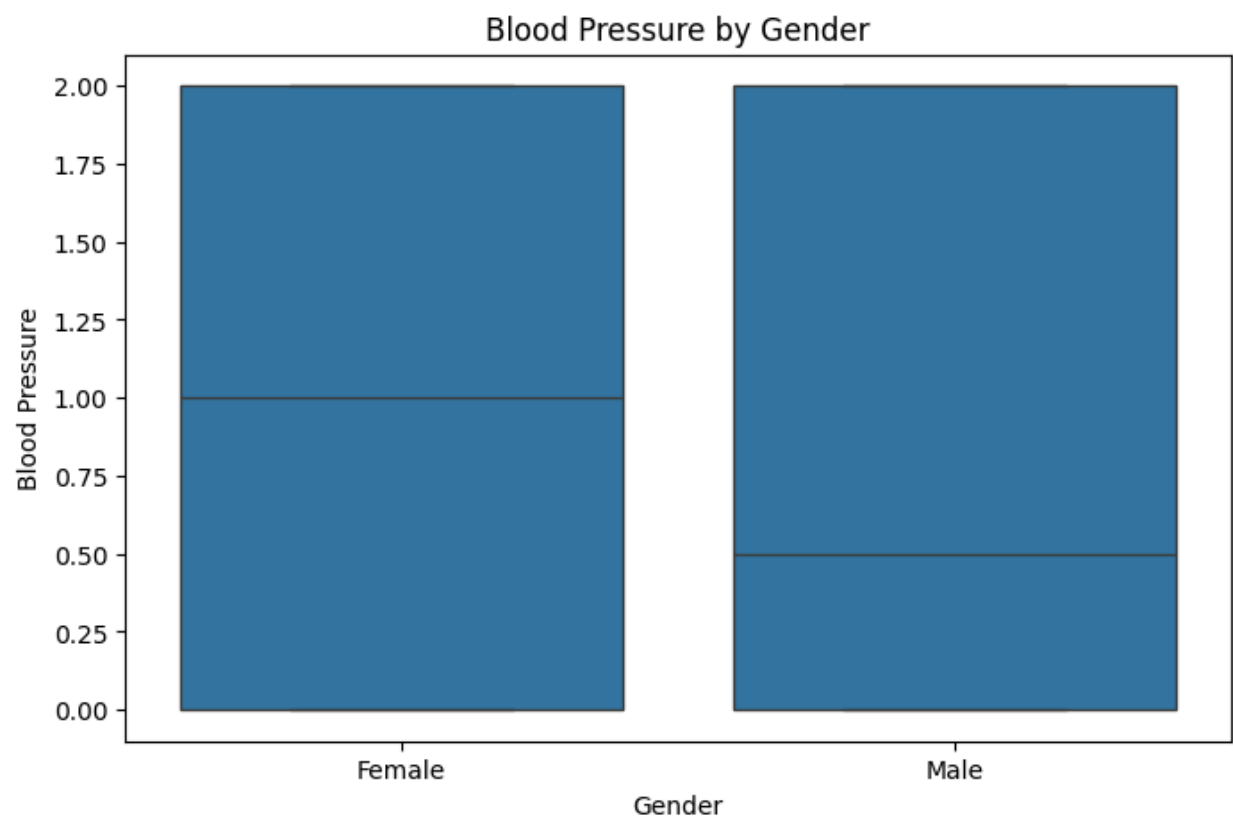


Figure 4

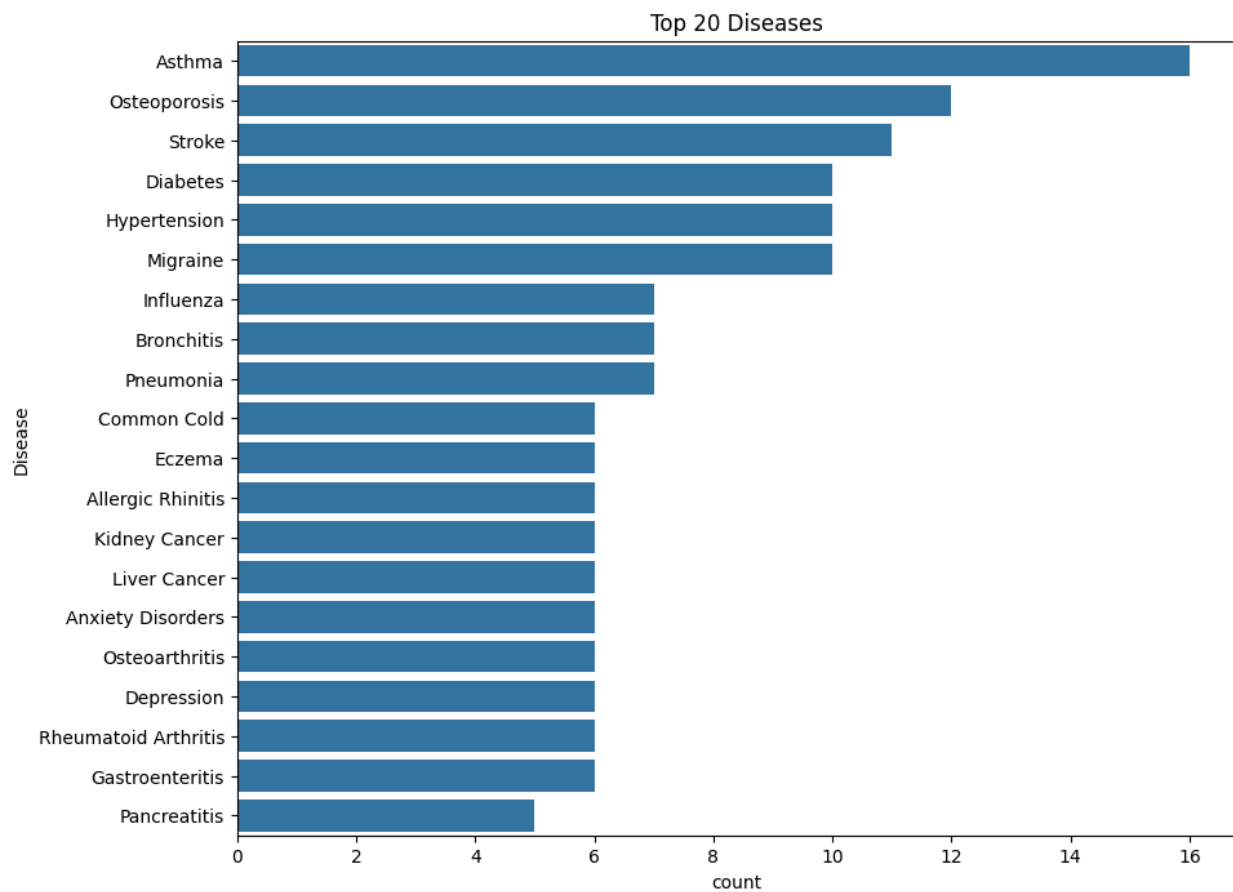


Figure 5

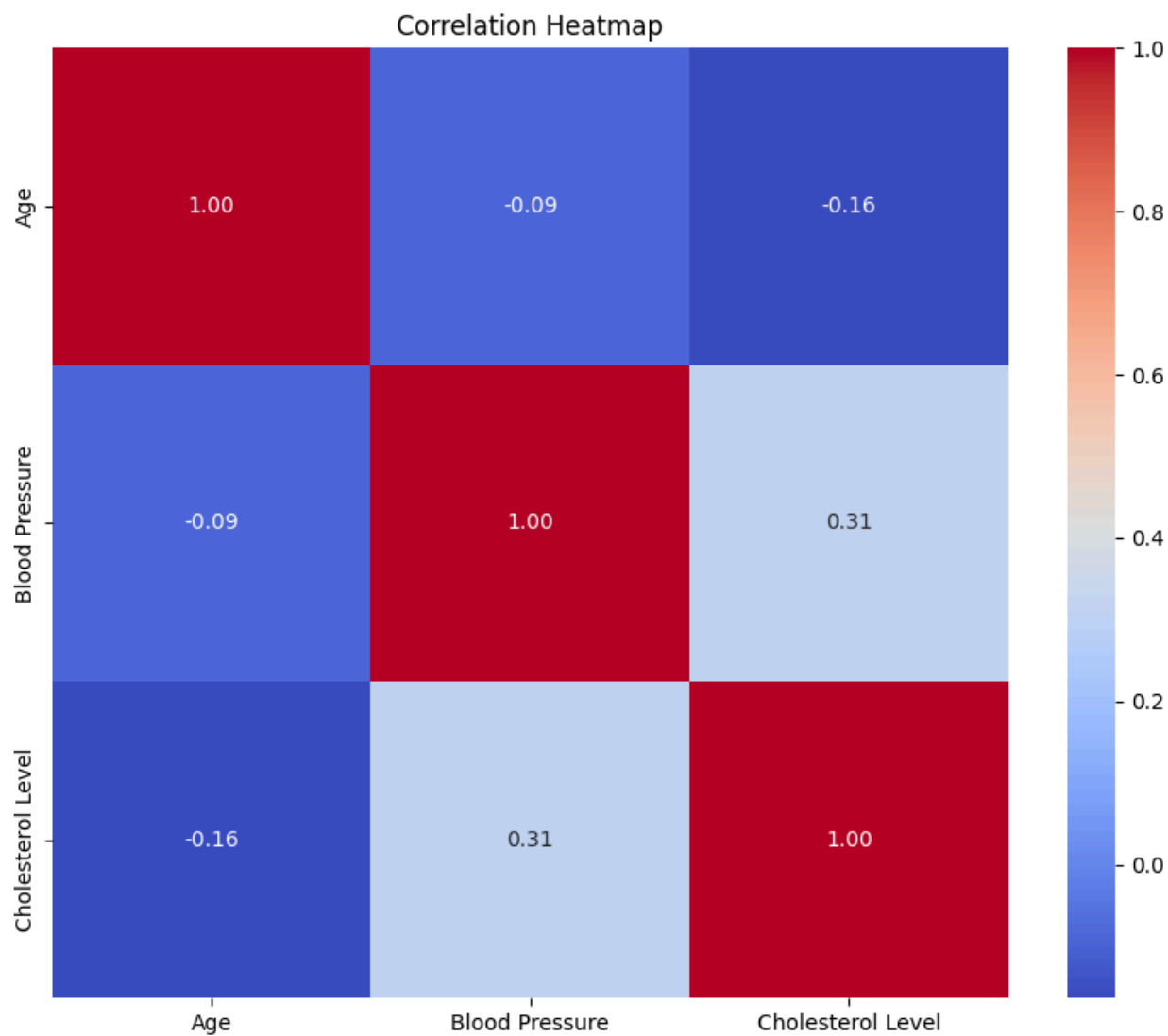


Figure 6