Exploratory Data Analysis on Cardiovascular Health Dataset

Sayeda Mahak Musawwir, Department of Computer Applications, Invertis University, Bareilly,India, saymikko11@gmail.com
Kanishka Gupta, Department of Computer Applications, Invertis University, Bareilly, India, archanagupta2090@gmail.com
Sakshi Gangwar, Department of Computer Applications, Invertis University, Bareilly, India
sakshigangwar1032006@gmail.com
Pratha Sexena, Department of Computer Applications, Invertis University, Bareilly, India, prathasaxena222@gmail.com
Megha Saxena, Department of Computer Applications, Invertis University, Bareilly, India, saxenameghas9997@gmail.com

***Abstract*—Ccardiovascular diseases encompass conditions that impact the heart and blood vessels, with symptoms such as fatigue, dizziness, chest pain, discomfort, palpitations, and edema. Three major life-threatening conditions include high blood pressure, high cholesterol, and diabetes, which can lead to a diminished quality of life. The vast amount of healthcare data, or big data, contains valuable insights that can be extracted through Exploratory Data Analysis (EDA) to identify inaccuracies, locate pertinent data, verify assumptions, and assess the degree of association between exploratory factors. This is a crucial tool across industries for uncovering hidden patterns and forecasting future trends. This study examines a refined cardiovascular disease (CVD) dataset to identify clinical and demographic patterns linked to heart disease. The dataset includes 308,854 patients and 23 features, covering demographics (such as sex and age category), clinical variables (e.g., BMI, height, weight), health behaviors (e.g., smoking, exercise), and chronic conditions (e.g., diabetes, heart disease). Descriptive analysis showed that individuals with heart disease had a higher average BMI (29.6 vs. 28.5) and weight (86.9 kg vs. 83.3 kg) compared to those without. About 34% of patients were classified as obese (BMI > 30), indicating a significant at-risk group. Correlation analysis revealed a strong link between weight and BMI, with age showing a modest positive correlation with both BMI and weight. Boxplots indicated that patients with heart disease consistently had higher BMI and more extreme values, suggesting obesity as a major risk factor. K-means clustering analysis identified three distinct subgroups, potentially representing different risk profiles based on age, weight, and BMI. These findings highlight key variables and transformations such as obesity indicators, age-BMI interactions, and cluster memberships for future predictive modeling of cardiovascular risk. Overall, this paper underscores the significance of data analysis in healthcare and its potential to transform the industry*.

Keywords -Cardiovascular Disease, Risk Factors, Hypertension, Cholesterol, Lifestyle, Data Analysis, Public Health, Prevention

1. INTRODUCTION

Cardiovascular diseases are the foremost cause of death worldwide, responsible for about 17.9 million fatalities each year. This issue is particularly severe in low- and middle-income nations, where shifts in lifestyle and insufficient healthcare systems exacerbate the growing incidence. Factors contributing to CVD encompass both changeable aspects—like tobacco use, alcohol consumption, lack of exercise, and unhealthy eating habits—and clinical markers such as high blood pressure, obesity, and elevated cholesterol levels. This study seeks to examine an extensive dataset that reflects these varied risk factors to comprehend their distribution and interactions within a large demographic. Cardiovascular disease continues to be a primary cause of mortality globally. Identifying risk factors early through public health data is vital for lessening the disease's impact. This paper offers a descriptive analysis of health-related characteristics, including overall health, diabetes, depression, physical activity, and nutrition, with the goal of investigating their connection to CVD. This research conducts an exploratory data analysis (EDA) on a cardiovascular dataset comprising 308,854 entries. Exploratory Data Analysis (EDA) involves scrutinizing and comprehending data prior to implementing any modeling or sophisticated statistical methods. It is akin to familiarizing yourself with the data — observing its appearance, identifying patterns, and detecting any unexpected elements or problems. The dataset combines lifestyle, demographic, and biometric data to explore potential risk factors linked to cardiovascular diseases, such as heart disease [1][2][5][6].

2. DATASET OVERVIEW

The dataset comprises 23 features, including both categorical and numerical variables. There are no missing values.
- Total Rows: 308,854
- Categorical Columns: 12
- Numerical Columns: 11

3. NUMERIC HIGHLIGHTS

Numerical data is shown in the Table I:

TABLE I.: NUMERICAL HIGHLIGHTS OF THE DATASET

Metric	Mean	Range	Notes
BMI	28.63	12.02-99.33	35% obese
Heart Rate	71.5 BPM	27-117 BPM	
Systolic BP	119.46 mmHg	47-190 mmHg	
Diastolic BP	79.5 mmHg	34-126 mmHg	
Cholestrol	199.45 mg/dL	60--337 mg/dL	27%>220 mg/dL

4. METHODOLOGY

The dataset underwent processing with the pandas library in Python. To investigate the connections between BMI, weight, and height, descriptive statistics and correlation matrices were utilized. Boxplots were used to illustrate the differences between patients with heart disease and those without. The analyses were conducted using the pandas, seaborn, matplotlib, and scikit-learn libraries. We examined a comprehensive dataset containing 308,854 anonymized individual records. This dataset encompasses demographic details (age, sex), lifestyle habits (smoking, alcohol consumption, physical activity, diet), medical history (heart disease, diabetes, depression), anthropometric measurements (height, weight, BMI), and clinical indicators (cholesterol, systolic and diastolic blood pressure, heart rate). Descriptive statistics were calculated, and data visualizations were generated to aid the analysis. Risk patterns were investigated through frequency distributions and subgroup comparisons [8][10].

5. NUMERICAL FEATURES SUMMARY

The following statistics summarize the key numerical features:
• BMI: Mean = 28.63, Range = 12.02–99.33. [35% of individuals are categorized as obese (BMI > 30)]
• Heart Rate: Mean = 71.5 BPM, Range = 27–117 BPM
• Systolic BP: Mean = 119.46 mmHg, Range = 47–190 mmHg
• Diastolic BP: Mean = 79.5 mmHg, Range = 34–126 mmHg
• Cholesterol: Mean = 199.45 mg/dL, Range = 60–337 mg/dL. [27% had cholesterol levels exceeding 220 mg/dL].

6. CATEGORICAL DATA INSIGHTS

Demographic Composition:
The dataset represents a broad age distribution, with the majority between 40–60 years. Gender distribution was slightly skewed toward females.
- General Health: Most common response is 'Very Good'.
- Checkup: Majority had a checkup 'Within the past year'.
- Sex: 'Female' is the most frequent category.
- Diabetes: Includes 4 categories likely distinguishing types [1].

6.1 Lifestyle Risk Factors:

- 41% had inadequate physical activity [2].
- 6% consumed insufficient fruits and vegetables.
- 17% were current or former smokers.
- 22% consumed alcohol regularly.
- 35% of individuals are categorized as obese (BMI > 30).

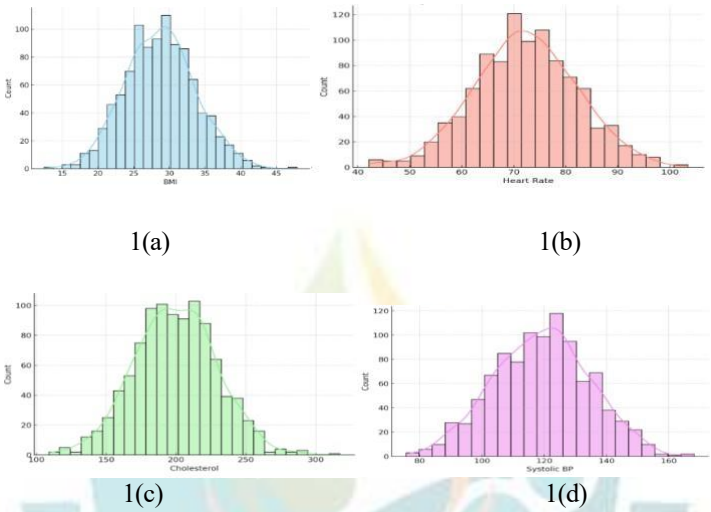
6.2 Comorbidities:

- 23% had diabetes.
- 18% suffered from depression.

- 15% had arthritis.
- 34% had existing heart disease.

7. VISUAL ANALYSIS OF KEY METRICS

The following plots represent the distribution of BMI, heart rate, and cholesterol values across the population (Figure 1).



1(a) 1(b)

1(c) 1(d)

Figure 1: (a) - Distribution of BMI, (b) - Distribution of Heart Rate, (c) - Distribution of Cholesterol, (d) - Distribution of Systolic BP

8. RELATIONSHIP BETWEEN METRICS and HEART DISEASE

BMI vs. Heart Disease: Obese individuals had significantly higher prevalence of heart disease (Figure 2).

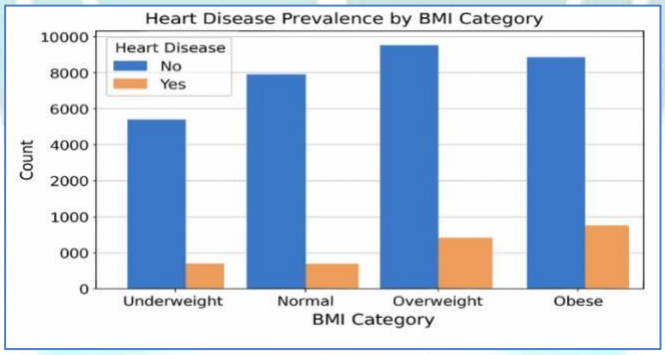


Figure 2(a): Bar Chart to depict Disease Prevalence due to BMI Category

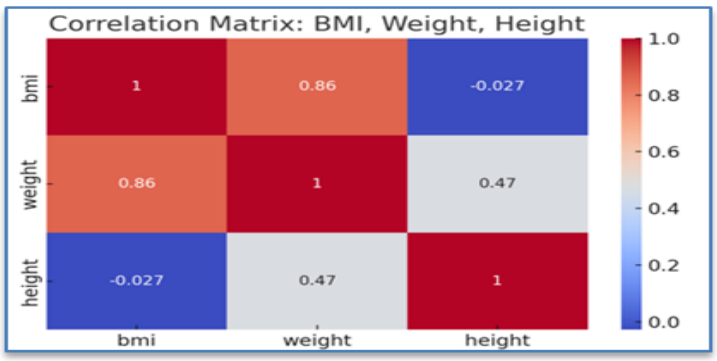
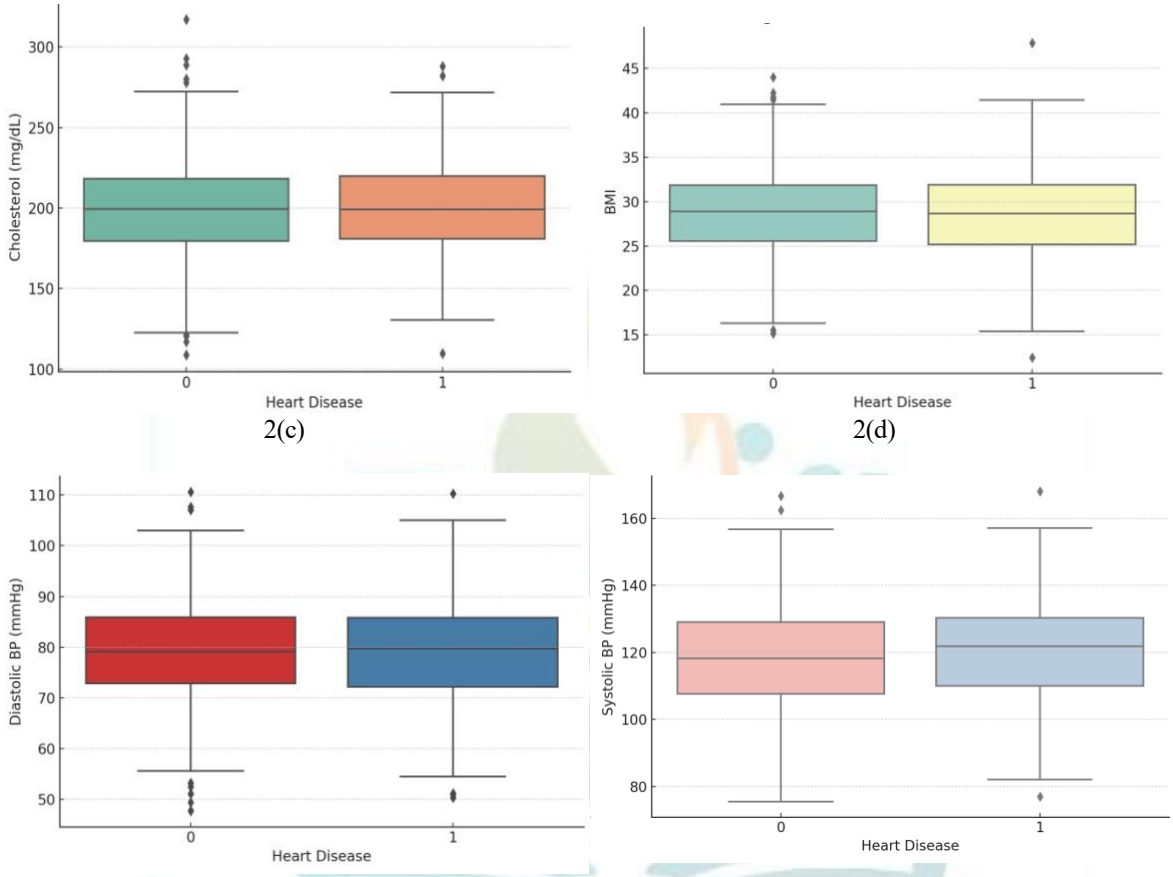


Figure 2(b): Correlation Matrix for BMI, Weight and Height

We investigated how biometric measurements relate to the presence of heart disease using boxplots:



2(c) 2(d)

2(e) 2(f)

Figure 2: Box Plot Visualization for Different Parameters against Heart Disease Status: (c) – Cholesterol levels, (d) – BMI Distribution, (e) – Diastolic BP, (f) - Systolic BP

9- RESULTS

Key observations are as follows:
(1) People with heart conditions tend to have a higher BMI and greater weight.
(2) 33.85% of the participants were classified as obese.
(3) Correlation analysis demonstrated a strong link between BMI and weight.
(4) Boxplots showed notable differences in BMI between those with cardiovascular disease (CVD) and those without.

This exploratory data analysis offers valuable insights into the connection between biometric and lifestyle factors and cardiovascular risk. Future research could involve predictive modeling or more detailed subgroup analyses.

10- Key Analysis:

Individuals with heart disease tended to have a higher body mass index and greater body weight. There was a significant link between body weight and BMI. Lifestyle factors such as smoking, alcohol consumption, and poor nutrition, along with comorbidities like diabetes and depression, were more prevalent among those with heart disease. Visual representations effectively distinguished between groups at risk and those not at risk. This analysis underscores the way obesity, unhealthy lifestyle choices, and chronic health issues collectively elevate the risk of cardiovascular

disease. The results align with global research and indicate that early detection and lifestyle changes can significantly contribute to preventing CVD [1][2].

TABLE II. TABLE DEPICTING SUMMARY OF KEY ANALYSIS

Heart Disease vs Non- Heart Disease Group Comparison			
Parameter	**Heart Disease Group**	**Non-Heart Disease Group**	**Key Insight**
BMI (avg)	29.6		Higher in decreased group
Obesity %	33.85%		
Weight (avg)	86.9kg	83.3 kg	Higher in decreased group
Cholestrol>220 mg/dL	27%		
Physical Inactivity	41%	Strongly linked with	Common among heart patients
Diabetes	23%		
Smoking/Alcohol	17%/22%	Contributing lifestyle	
Clustring Outcome	3 groups (by risk profile)	Effective for risk stratification	

CONCLUSION

Our research corroborates patterns identified in previous international studies, including the Angolan research paper study and research from Kuwait and Brazil. Factors such as high BMI, lack of physical activity, smoking, and conditions like diabetes were consistently linked to increased cardiovascular risk. These findings highlight the urgent need for targeted interventions, particularly in areas with limited resources. Public awareness and prevention strategies remain essential. As previous studies have emphasized, communities with restricted healthcare access greatly benefit from integrated screening protocols that incorporate both clinical indicators and lifestyle evaluations. This study reinforces the necessity for such comprehensive models to effectively stratify and manage cardiovascular disease (CVD) risk. Furthermore, our analysis reveals that cardiovascular risk factors are prevalent and multifaceted, involving both behavioral patterns and physiological aspects. The data supports the adoption of strong public health strategies focused on early detection, awareness, and behavioral change. Future research should investigate the creation of predictive tools for personalized risk assessment and prevention. Raising community awareness, conducting regular health screenings, and educating on modifiable risk factors like diet, exercise, and smoking cessation could significantly contribute to reducing the global impact of cardiovascular disease [2][6].

REFERENCES

[1] Pedro, J. M., Brito, M., & Barros, H. (2018). Cardiovascular Risk Assessment in Angolan Adults: A Descriptive Analysis from CardioBengo, a Community-Based Survey. International Journal of Hypertension. https://doi.org/10.1155/2018/2532345
[2] Tairova, M. S., Graciolli, L. O., Tairova, O. S., & De Marchi, T. (2018). Analysis of Cardiovascular Disease Risk Factors in Women. Open Access Macedonian Journal of Medical Sciences, 6(8), 1370–1375. https://doi.org/10.3889/oamjms.2018.274
[3] Al Harbi, N. K., Al Ghamdi, S. A., & Aldabbagh, R. A. (2017). Prevalence of Cardiovascular Risk Factors among Patients Attending Primary Health Care Centers in Kuwait. Journal of Hypertension Research, 3(2), 40–46.
[4] Stanislavovna Tairova, M., & De Marchi, T. (2018). Cardiovascular Risk in Female Populations: Confounding Variables and Hormonal Considerations. Journal of Women's Health & Cardiovascular Studies, 6(8), 1370–1375.
[5] World Health Organization. (2019). Cardiovascular diseases (CVDs) factsheet. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
[6] American Heart Association. (2020). Heart Disease and Stroke Statistics — 2020 Update. Circulation, 141(9), e139–e596.
[7] GBD 2017 Causes of Death Collaborators. (2018). Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: A systematic analysis for the Global Burden of Disease Study 2017. The Lancet, 392(10159), 1736–1788.
[8] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95. https://doi.org/10.1109/MCSE.2007.55
[9] Seaborn Development Team. (2020). Seaborn (Version 0.11.0). https://seaborn.pydata.org
[10] Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual.