Exploring Risk Factors for Cardiovascular Disease

Kumar Karthik Ankasandra Naveen

Sai Preetham Nagulapalli

Hari Priya Ramamoorthy

College of Professional Studies, Northeastern University

ALY 6040 Data Mining Applications

Professor. Justin Grosz

February 28, 2025

Abstract

Cardiovascular disease (CVD) remains the leading cause of mortality in the U.S., necessitating a data-driven approach to identify key risk factors. This study analyzes the Personal Key Indicators of Heart Disease dataset from the CDC using exploratory data analysis (EDA) to examine demographic, medical, and behavioral contributors to CVD.

Data preprocessing involved feature engineering, categorical encoding, and standardization. EDA revealed an imbalanced distribution of cardiovascular events and highlighted key trends. Correlation analysis and hypothesis testing confirmed that age, smoking, and diabetes significantly increase CVD risk, while BMI alone is not a strong predictor.

These findings emphasize the need for targeted public health strategies focused on aging populations, smoking cessation, and diabetes management. Future research will refine predictive models to enhance risk assessment and healthcare interventions.

Introduction

Heart disease remains the leading cause of death in the United States, with approximately 697,000 fatalities in 2021 (CDC, 2023). This study analyzes the Personal Key Indicators of Heart Disease dataset from Kaggle, derived from CDC surveys, to identify key risk factors across U.S. states. Given the influence of both medical and behavioral factors, the findings aim to inform public health strategies for early intervention and prevention.

The report details data cleaning, exploratory data analysis (EDA), and hypothesis testing to assess the impact of demographic, health, and lifestyle attributes on heart disease prevalence. It also discusses future research directions to refine predictive models and enhance public health recommendations.

Data Cleaning

The dataset consists of 246,022 observations across 40 columns, capturing a wide range of demographic, health, and lifestyle factors. To provide insights to preventive health department, focus variables include age, race, BMI, diabetes status, arthritis, asthma, pulmonary diseases, physical activity levels, and cardiovascular event outcomes, alongside lifestyle factors such as smoking, alcohol consumption, sleep hours, and mental health status. The data cleaning process involved merging redundant columns and encoding categorical variables to enhance data quality.

Missing Value: The first step involved checking for missing values, and no missing data was detected, indicating that the dataset was already pre-processed.

Feature Engineering: To simplify the dataset's structure and reduce redundancy, several columns were mapped together. For instance, Heart attack, stroke, and angina history were consolidated into a single binary Cardio Event column, indicating whether an individual experienced any cardiovascular event.

Categorical Variable Encoding: To prepare dataset for analysis, binary yes/no columns mapped to 0/1, ordinal columns (like general health) were numerically encoded to reflect their natural order, and nominal columns (like race/ethnicity) label encoded.

Standardization and Deduplication: Column names were standardized to ensure consistent naming conventions and improved readability. Unwanted columns were dropped from the dataset. To Duplicate observations were removed to enhance data quality for further analysis.

By the end of this, the dataset prepared for Exploratory Data Analysis (EDA), comprised of 25 columns with a particular focus on identifying risk factors associated with cardiovascular events.

Exploratory Data Analysis

For the prepared dataset, the key summary statistics for numerical columns were explored. Outliers were identified across several numerical columns, highlighting potential extreme values beyond 2 standard deviations that could impact the analysis using boxplots. Figure-1 Boxplot captures outliers in sleep hours highlighting outliers beyond 12 hrs. These outliers were evaluated to determine if they reflected plausible extreme health and lifestyle conditions or data entry errors, helping to decide whether they should be retained, transformed, or removed based on their impact on model performance and data distribution. The bar chart (Figure 2) shows the distribution of the Cardio Event variable, revealing a significant class imbalance, with far fewer individuals experiencing cardiovascular events compared to those who did not. This highlights the need for resampling techniques during modeling to improve predictive performance.

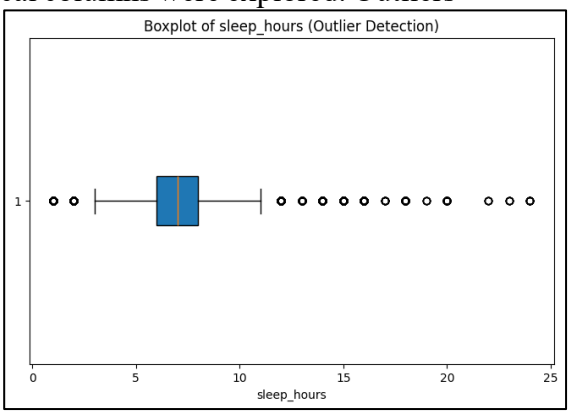

Figure 1. Boxplot of sleep hours

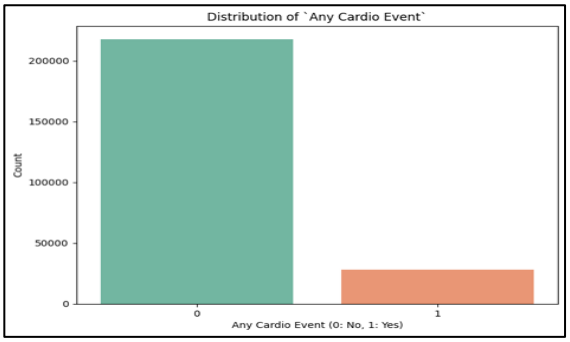

Figure 2. Distribution of Cardiac Events

Correlation Analysis and Insights

This analysis examines the relationships between health, lifestyle, and demographic factors with cardiovascular disease (Any Cardio Event). Given the categorical nature of variables such as Sex, General Health, Alcohol Drinkers, and smoking status, categorical variable encoding was applied during preprocessing to ensure numerical suitability for statistical analysis.

A correlation heatmap was generated to visualize relationships among features, with darker shades indicating stronger correlations. The highest observed correlation was 0.42 between Mental Health

Days and Had Depressive Disorder, which aligns with expectations. Other notable correlations included 0.38 between Had Arthritis and Age Category and 0.25 between Age Category and Any Cardio Event, reflecting weak to moderate associations.

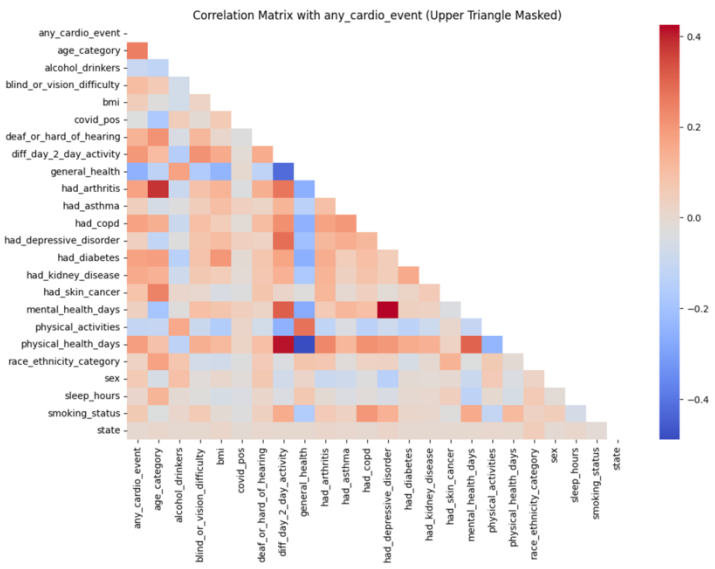

Figure 1. Correlation Matrix

Crucially, Any Cardio Event did not exhibit strong correlations with any single variable, suggesting that cardiovascular disease is influenced by multiple interrelated factors rather than simple linear relationships. Given these findings, correlation analysis alone is insufficient for drawing conclusions.

To deepen our understanding, hypothesis testing will be conducted to assess the statistical significance of lifestyle factors including smoking, alcohol consumption, sleep duration, and physical activity on heart disease prevalence. These tests will help determine whether observed trends are meaningful or occur by chance, providing stronger evidence beyond correlation analysis.

Hypothesis Testing

Hypothesis 1: Older individuals have a higher risk of experiencing a cardiovascular event.

Rationale: Cardiovascular diseases are strongly associated with aging due to factors such as declining heart function, increased arterial stiffness, and the cumulative impact of risk factors over time. This analysis examines whether older individuals in the dataset exhibit a higher frequency of cardiovascular events.

Findings from the Plot: The age distribution boxplot indicates that individuals who experienced a cardiovascular event tend to be older. The median age for those with heart disease is notably higher than for those without, and the interquartile range (IQR) is skewed toward older age groups.

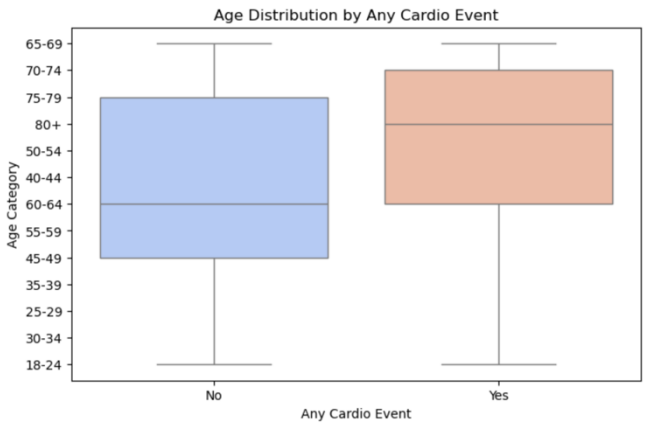

Figure 2. Age distribution of cardio events

Conclusion: The visualization provides strong evidence supporting the hypothesis that cardiovascular risk increases with age. The clear shift in age distribution confirms that cardiovascular events are more prevalent among older individuals, aligning with established medical research.

Hypothesis 2: Higher Body Mass Index (BMI) increases the likelihood of a cardiovascular event.

Rationale: Obesity and excess body weight are well-established risk factors for cardiovascular disease, as they contribute to elevated blood pressure, cholesterol levels, and an increased risk of diabetes. This analysis examines whether individuals with higher BMI are more prone to experiencing cardiovascular events.

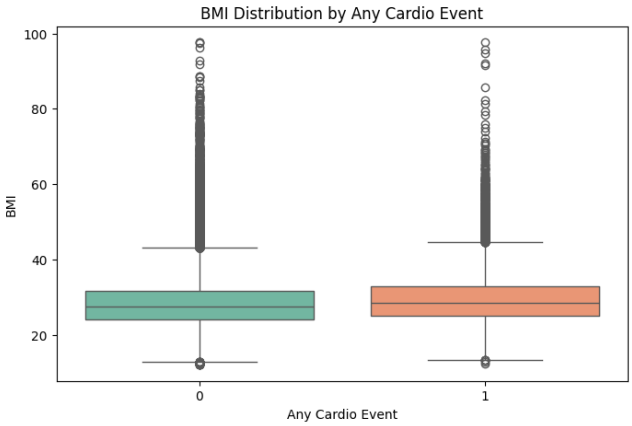

Figure 3. BMI Distribution of cardio events

Findings from the Plot: The BMI distribution boxplot reveals a slight difference in BMI between individuals with and without cardiovascular events. However, there is substantial overlap between the two groups, indicating that while BMI may contribute to cardiovascular risk, it is not a definitive predictor. Both groups also contain outliers with extremely high BMI values.

Conclusion: The results do not strongly support the hypothesis that BMI alone is a major predictor of cardiovascular disease. While individuals with cardiovascular events tend to have slightly higher BMI, the association is not as pronounced as with age. This suggests that BMI may be a contributing factor but is unlikely to be the sole determinant of cardiovascular disease.

Hypothesis 3: Smokers have a significantly higher risk of experiencing a cardiovascular event compared to non-smokers.

Rationale: Smoking is a well-documented risk factor for cardiovascular disease, as it damages blood vessels, raises blood pressure, and increases the likelihood of blood clots.

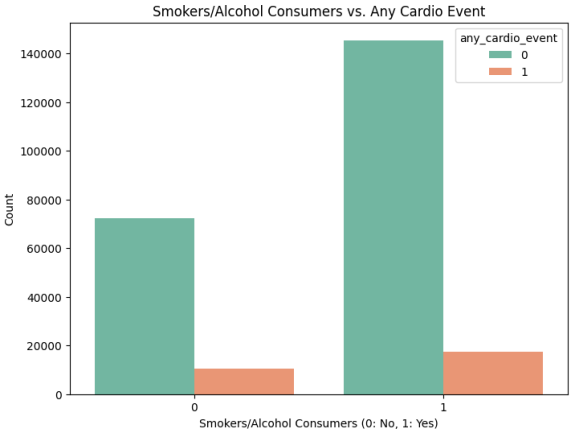

Figure 4. Smokers/Alcohol Consumers vs Cardio Events

Findings from the Chart: The bar chart comparing smoking status and cardiovascular events shows that while most individuals in the dataset are non-smokers, a notable proportion of smokers have experienced cardiovascular events. The trend suggests that smokers are overrepresented in the cardiovascular disease group compared to non-smokers.

Conclusion: The results strongly support the hypothesis that smoking is associated with an increased risk of cardiovascular disease. The higher proportion of smokers with cardiovascular events aligns with established medical research, reinforcing the well-known link between smoking and heart disease.

Hypothesis 4: Individuals with diabetes (including gestational and pre-diabetic conditions) have a higher risk of experiencing a cardiovascular event.

Rationale: Diabetes, especially when uncontrolled, can lead to high blood sugar levels that damage blood vessels and nerves regulating heart function. This increases the risk of heart disease, stroke, and other cardiovascular complications.

Findings from the Chart: The bar chart comparing diabetes status and cardiovascular events reveals that individuals diagnosed with diabetes have a higher proportion of cardiovascular events than non-diabetics. This trend is particularly pronounced for those with diagnosed diabetes, highlighting a clear association between diabetes and cardiovascular risk.

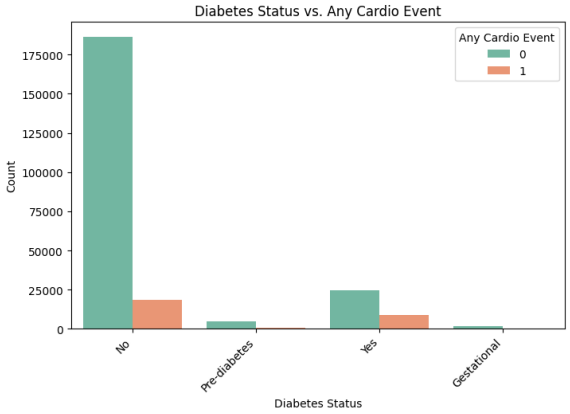

Figure 5.Diabetes Status vs Cardio Events

Conclusion: The findings strongly support the hypothesis that diabetes increases the likelihood of cardiovascular disease. The elevated prevalence of cardiovascular events among diabetics confirms that diabetes is a significant contributing factor to heart disease, aligning with established medical research.

Hypothesis 5: Racial and Ethnic Differences in Cardiovascular Disease Risk

Rationale: Cardiovascular disease (CVD) prevalence varies among racial and ethnic groups due to genetic predisposition, socioeconomic factors, and lifestyle differences. This analysis examines certain racial experience a higher Cardio events.

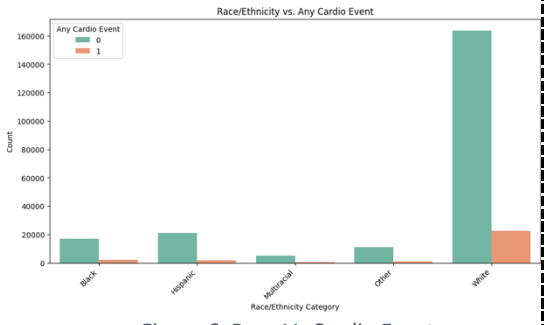

Figure 6. Race Vs Cardio Events

Findings from the Chart: The bar chart shows that the White population has the highest number of CVD cases. However, this may be influenced by their larger representation in the dataset rather than an inherent risk factor. Other racial groups show lower counts, but proportional differences are unclear.

Conclusion: The chart alone does not confirm whether race is an independent risk factor for CVD. Further statistical analysis, such as a Chi-square test, is necessary to determine if these differences remain significant after accounting for confounding variables like age, BMI, and smoking status.

Conclusion & Next Steps

Our analysis explored four key hypotheses related to cardiovascular disease: the impact of age, BMI, smoking, and diabetes on heart disease occurrence. The results strongly support the hypothesis that older individuals are more likely to experience cardiovascular events, as the median age of those with heart disease is significantly higher. Similarly, smoking status exhibits a clear trend, with a greater proportion of smokers experiencing cardiovascular disease, reinforcing the well-established link between smoking and heart-related health risks. Diabetes also emerges as a critical factor, as individuals with diabetes, particularly those with a confirmed diagnosis, show a noticeably higher prevalence of cardiovascular events compared to non-diabetics. These findings validate the role of diabetes in increasing heart disease risk. However, BMI does not exhibit a strong correlation with cardiovascular disease. While individuals with higher BMI values are slightly more represented in the heart disease group, the overlap between those with and without heart disease suggests that BMI alone may not be a definitive risk factor. Overall, our analysis confirms that age, smoking, and diabetes have a statistically significant impact on cardiovascular disease, while BMI alone is not a primary predictor. Further hypothesis testing will help quantify these relationships more precisely for predictive modeling and health risk assessments.

Reference

- Centers for Disease Control and Prevention. (2023). Heart disease facts and statistics. U.S. Department of Health & Human Services. https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html
- Pytlak, K. (2023). Personal key indicators of heart disease [Data set]. Kaggle. https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease