Risks for Heart Disease

# Introduction

Heart disease is the leading cause of death in the US and around the world according to the U.S. Centers for Disease Control (CDC). Understanding the various risk factors associated with heart disease can help us to better understand what public health interventions should target to prevent chronic heart disease. In this analysis, we examine the Framingham Heart Study data to identify the most strongly correlated risk factor, and how that risk factor varies by different groups.

# Background

Heart disease is an umbrella term that refers to several different heart conditions ranging from heart attacks, strokes, and arrhythmia (i.e. abnormal heart rhythm). However, coronary heart disease is the most prevalent (CDC). Coronary heart disease develops when a waxy substance called plaque builds up inside a person’s arteries. This process is also known as atherosclerosis. The plaque buildup restricts blood flow, making it hard to circulate blood to different parts of the body, including the heart. The symptoms a person with coronary heart disease will experience varies, but the disease can cause heart failure, heart attacks, and strokes (Mayo Clinic).

While many large public health organizations such as the CDC and the American Heart Association run broad public health campaigns to raise awareness of coronary heart disease, the disease is not equally prevalent across the U.S. population (Citation needed). Furthermore, smaller public health organizations at the state and local levels may not have the resources to conduct extensive campaigns. As a result, it may be necessary to target a subset of the population most at risk for coronary heart disease in a public health intervention that has a limited budget.

# Research Objective

Our research objective is to understand what is the top major risk factor that can play a paramount role in whether a patient might develop coronary heart disease in the next 10 years, and whether differences between groups can be made to further target public health interventions.

# Methodology

## Dataset

To better understand what characteristics are most associated with high coronary heart disease risk, we leveraged a publicly available extract of data from the Framingham Heart Study. Mahmood,et al. point out that the Framingham Heart Study is a longitudinal study established in 1948. 5,209 residents of Framingham, Massachusetts, both men and women from 28 to 62 years old, were recruited in a prospective epidemiologic cohort study. The objectives of the Framingham Study are to study the incidence and prevalence of heart disease and its risk factors over time.

Study participants continued to return to the study every two years for a detailed medical history, physical examination, and laboratory tests. This was the first prospective study of cardiovascular disease and identified the concept of risk factors and their joint effects. The factors studied include age, gender, cigarette-smoking behavior, blood cholesterol, blood pressure, left ventricular hypertrophy, diabetes, and whether the patient developed coronary heart disease (CHD).

There are three potential limitations in using the Framingham Heart Study data:

1. The Framingham heart study only measures a limited amount of behavioral and medical factors, missing out on environmental factors that may contribute to heart diseases.
2. The Framingham heart study population is almost all Caucasian, and thus the sample is biased in that it does not reflect the broader population.
3. Since age is a prominent determinant of the coronary heart disease risk score, the 10-year hazards of coronary heart disease are, on average, high in older persons. This may over-identify candidates for aggressive interventions.

## Prior Research

The Framingham Heart Study has contributed transformative discoveries related to the treatment of heart disease. Through this study, scientists have learned risk factors for heart disease, and they now know that many of those risks can be changed. It is why, in routine physicals, doctors check for high blood pressure, high cholesterol, unhealthy eating patterns, smoking, physical inactivity, or unhealthy weight. Researchers also know that these conditions can affect people differently depending on a patient’s sex or race. Findings such as these pave the way for new interventions to preempt, prevent, or treat this condition more effectively. According to the Kannel report, In 1971, researchers conducting the Framingham Heart Study analyzed over 10 years of follow-up data to demonstrate increased risk of coronary heart disease morbidity with rising baseline blood pressure. When systolic and diastolic pressures were studied for association with coronary heart disease events, systolic pressure demonstrated a stronger link than diastolic pressure.

## Exploratory Data Analysis

## Data Pre-Processing

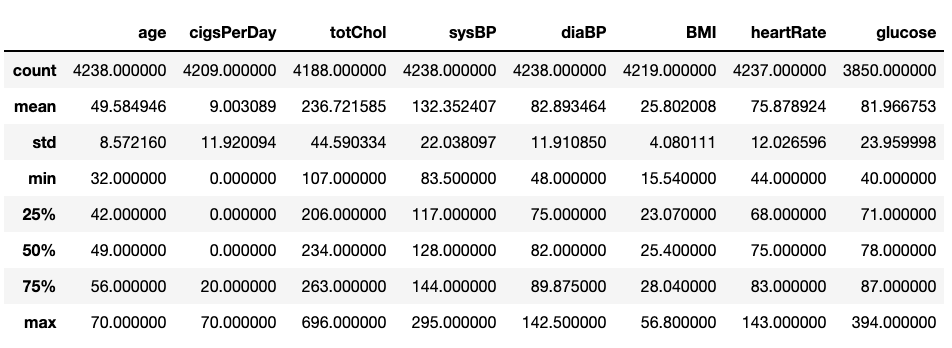
In the initial exploration of the data we found that there were 4,238 observations of 16 different variables. We started processing the data by renaming the first column to be gender instead of male. We also printed the first ten rows of the data as a summary of the dataset. Moreover, we have covered the descriptive of the data by demonstrating the count, mean, std, max, and min value for each interval value column [Figure1].

Figure 1: A quick demonstration about the count, mean, std, max, and min value for each column.

Furthermore, we explored missing values and reduced the data set to ensure non-null rows across each column. By only using complete rows, we were left with 3656 total observations. As a part of the summary of the dataset we plotted the data distribution to visualize and to see how the data was measured. The [figure 2] below illustrates each column and how it measures.

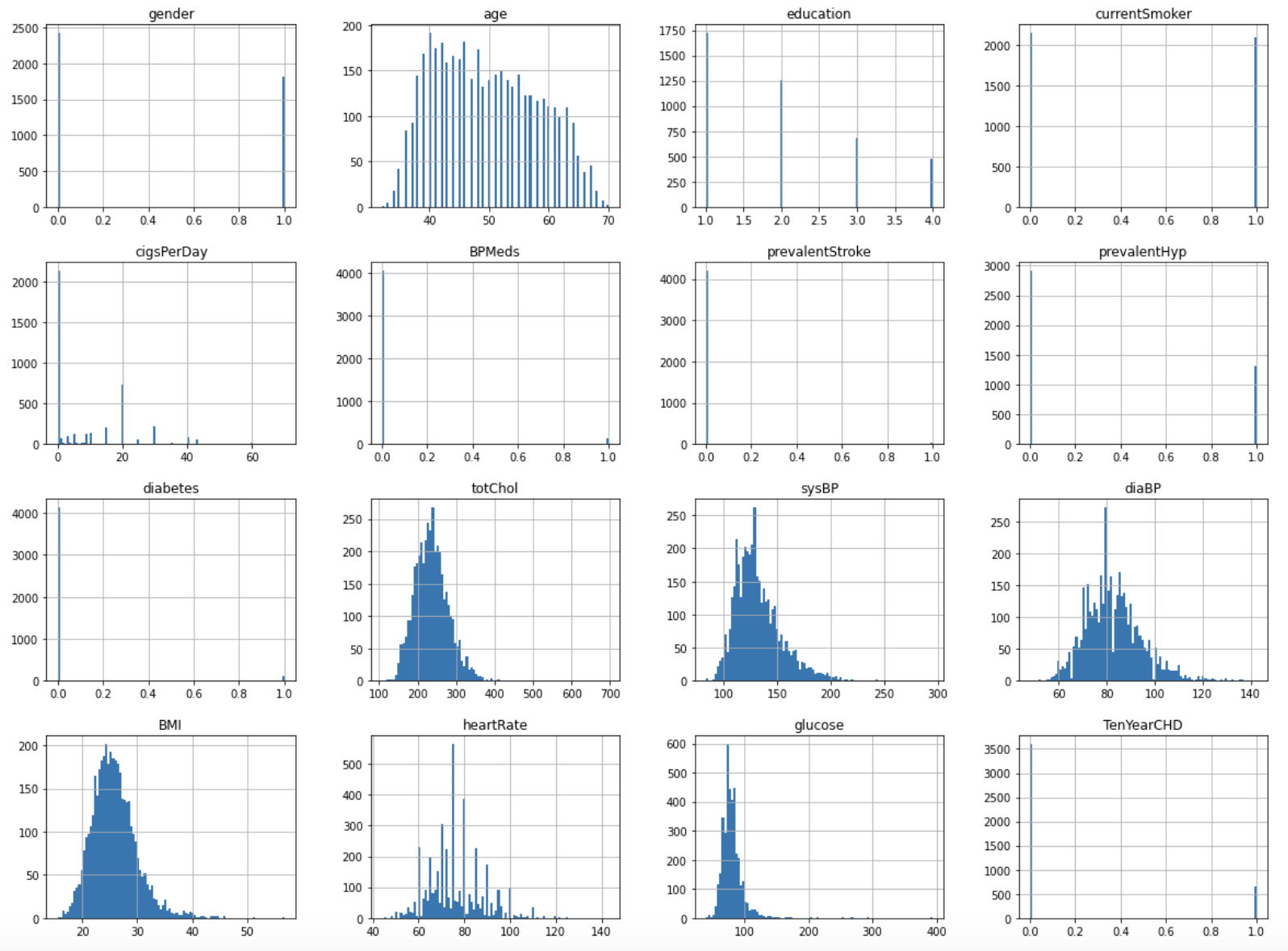
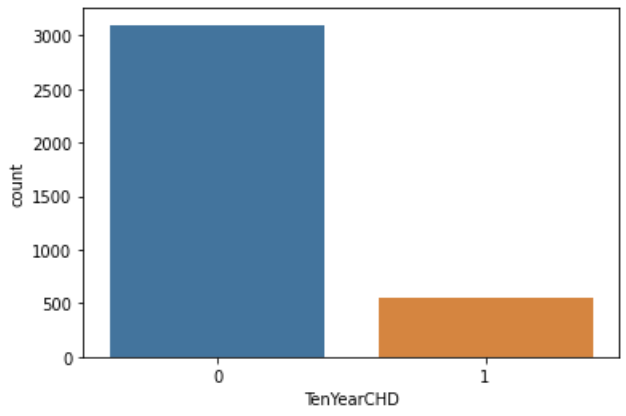


Figure 2: Data distribution

After having a clean and completed dataset we can start answering our biological questions. First, we found that from the distribution of the target column 557 who developed CHD in the ten years where 3099 did not develop CHD [Figure3].

Figure 3: The distribution of the target column (TenYearCHD)

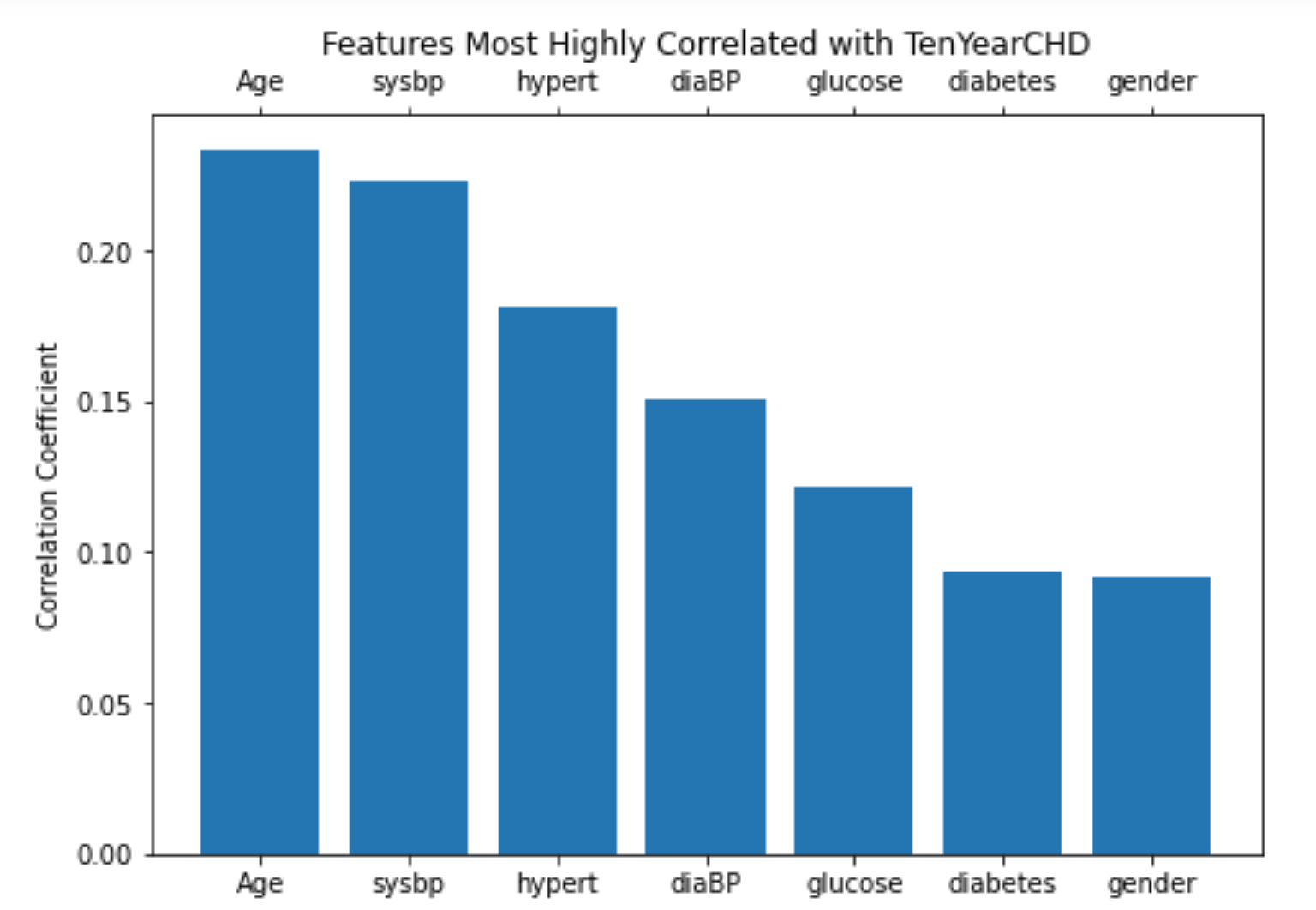
Continuing with EDA, we correlated the data to see what is the strongest relation between the columns. There were around 7 highlighted correlations that we focus on to answer our biological question. Looking at the [Figure 4], we can see that the top variables that are correlated with CHD are age, systolic blood pressure, a previous diagnosis of hypertension, diastolic blood pressure, blood sugar levels, previous diagnosis of diabetes, and gender.

Figure 4: The most highly correlations between the observed columns

## Research Questions

Based on our findings on the variables most correlated with CHD, we identified two key research questions:

* **Question 1:** Do patients who developed CHD have a higher systolic blood pressure on average compared to patients who did not develop CHD?
* **Question 2:** Are patients who developed CHD on average older compared to patients who did not develop CHD?

## Statistical Analysis

In this research, our independent variables focused on patients’ age and systolic blood pressure, which are both interval values. For variables measured at the interval level, we used mean and standard deviation as our primary descriptive statistic. The sample was separated into two groups, being patients who developed CHD within 10 years and patients who did not. We constructed a 95% Confidence Interval (C.I.) for the population proportion of patients who developed CHD. We also constructed a 95% C.I. for population mean systolic blood pressure and age for all with or without CHD. We used a z-test to determine whether the two population means (for age and systolic blood pressure respectively) are different because the variances are known and the sample size is large.

Our exploratory data analysis revealed that there was some correlation between our interval variables. Indeed, prior research by Elisabete Pinto found that the increase in blood pressure with age is mostly associated with structural changes in the arteries and especially with large artery stiffness. Age-related increases in blood pressure have been observed in almost every population. To better explore this relationship, we used linear regression between age and systolic blood pressure.

# Results

According to the data, we have a total of 3656 individuals as observation objects. Among them, the number of patients with coronary heart disease was 557, and the number of patients with non-coronary heart disease was 3099. With 95% confidence, the population proportion of people having CHD is estimated to be between 14.10% - 16.26%. With 95% confidence, the population mean systolic blood pressure for all non CVD Patients is estimated to be between 129.57mm Hg and 131mm Hg With 95% confidence, the population mean age for all non CHD Patients is estimated to be between 48.4 and 48.99 years.

## Age and Coronary Heart Disease

The average age of patients who developed coronary heart disease is 54 years old and patients who didn’t develop CHD were 49 years old on average (see figure 5). To test if this difference was statistically significant, we proposed the following hypotheses：

* **Null hypothesis:** There is no difference in mean age.
* **Alternative hypothesis:** There is a significant difference in mean age.

Using the same statistical method, we get the P value. P-Value is5.698450859102773e-48. Our p-value is smaller than the 0.05 significance level, which means there is strong evidence against the null. Thus, we reject the null. Based on our estimated difference in sample means, we support that there is a significant difference between the population mean age for patients who developed CHD and those who didn’t.

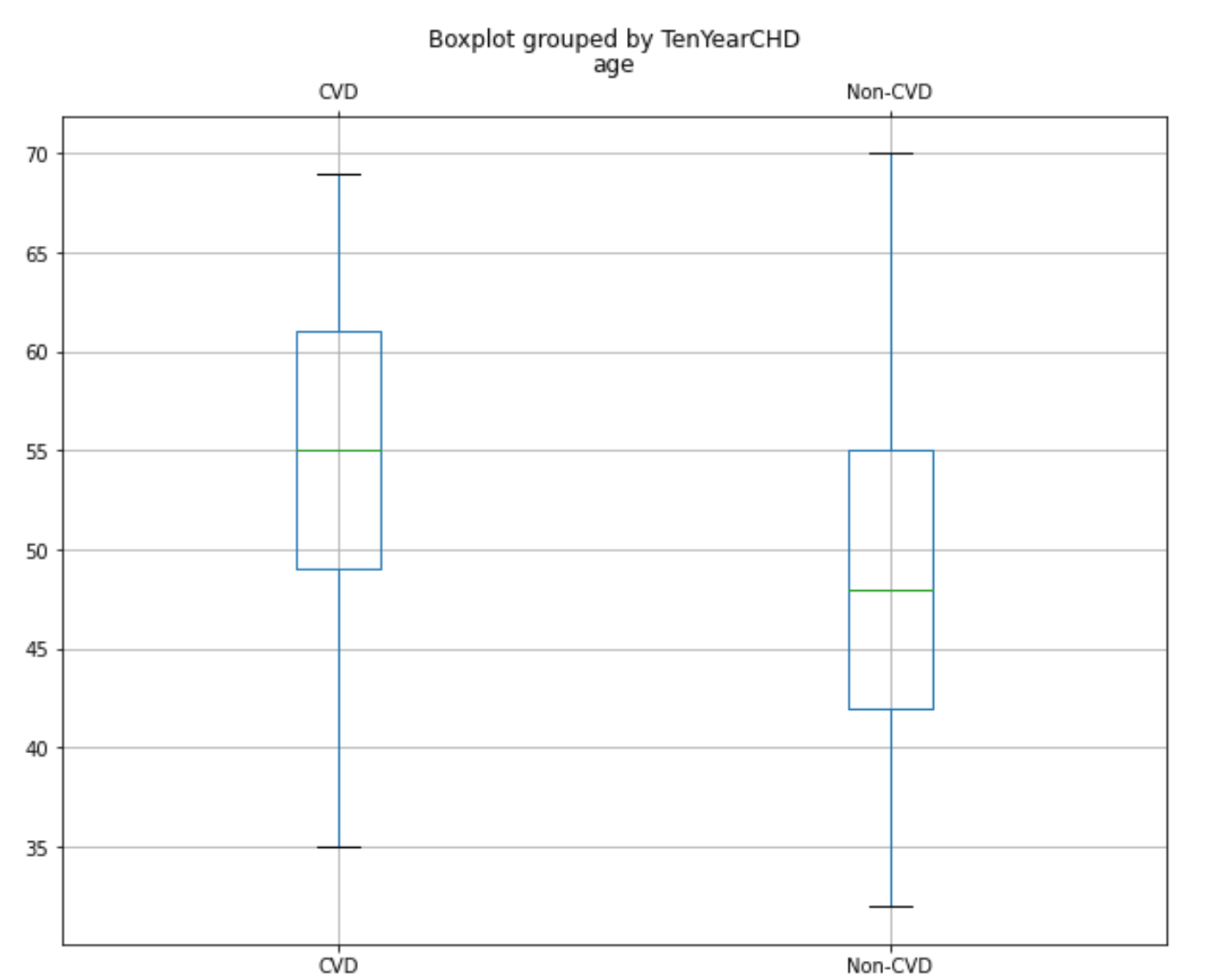


Figure 5 : Average age of patients with or without coronary heart disease

## Systolic Blood Pressure and Coronary Heart Disease

The average blood pressure of patients with coronary heart disease was 143.98mmHg, which was significantly higher than that of the non-coronary heart disease group (see figure 6). To test if this difference was statistically significant, we proposed the following hypotheses：

* **Null hypothesis:** There is no difference in mean systolic blood pressure.
* **Alternative hypothesis:** There is a significant difference in mean systolic blood pressure.

We use z-test with a significant level of 5%. The p value obtained is 1.922670905699858e-43. Our p-value is smaller than the 0.05 significance level, which means there is strong evidence against the null. Thus, we reject the null. Based on our estimated difference in sample means, we support that there is a significant difference between the population mean systolic blood pressure for patients who developed CHD and those who didn’t.

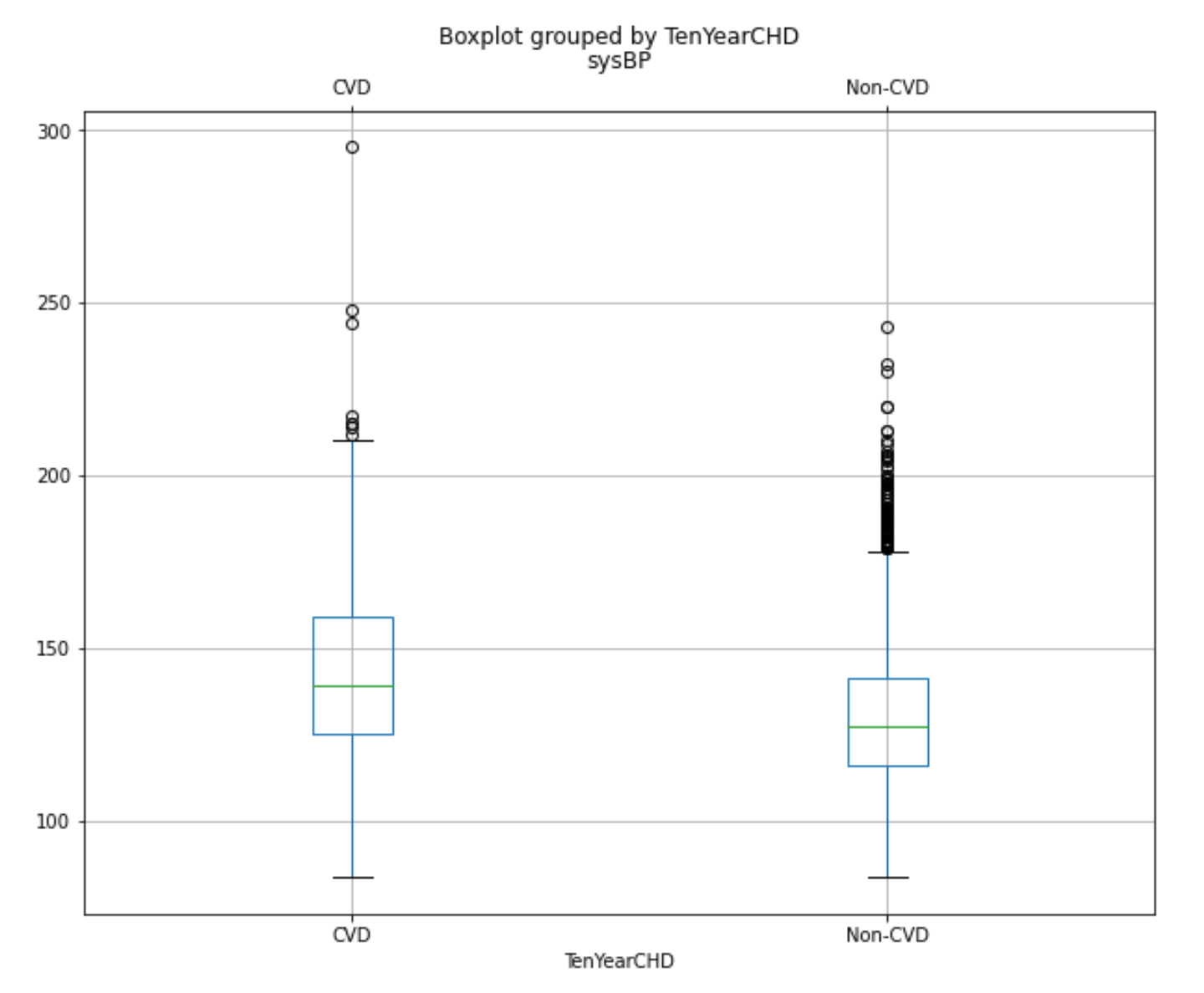


Figure 6: Average blood pressure of patients with or without coronary heart disease

## Age and Systolic Blood Pressure

Given the prior research shows a relationship between age and blood pressure, we conducted a linear regression with age as independent variable and systolic as dependent variable. First, we find the number of people who have heart disease according to age (see figure 7). Then our linear model found that for every year a person ages, their systolic blood pressure increases by 1 mm Hg. Our R-squared value was .38, which means that nearly 40% of the variation in systolic blood pressure can be explained by age. For our visualization purpose, we used a fit line showing this relationship among patients who developed CHD and those who didn’t (see figure 8).

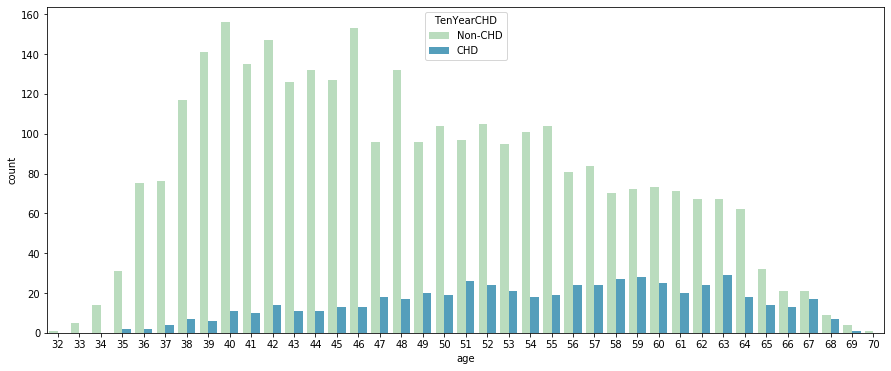


Figure 7: The number of people who have heart disease according to age

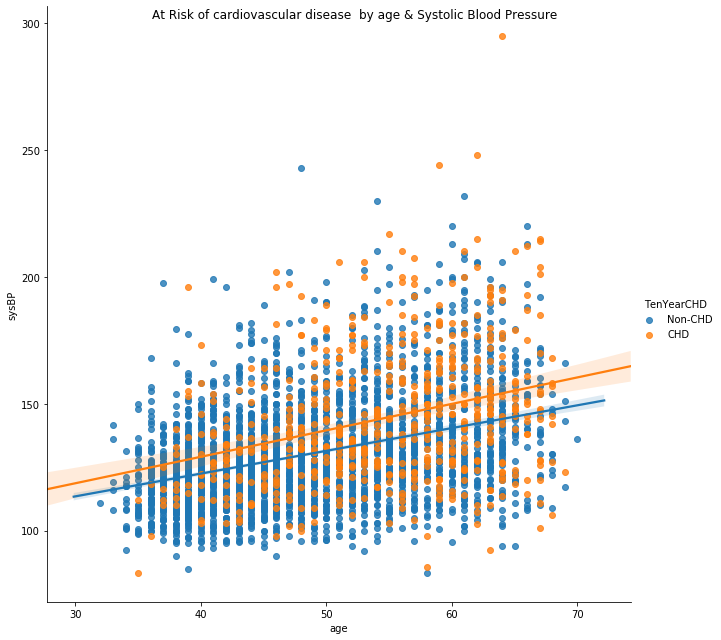


Figure 8: At Risk of cardiovascular disease by age and Systolic Blood Pressure

# Conclusions

For state and local public health officials who have limited budgets, it is helpful to know which populations are most at risk for coronary heart disease. This allows officials to ensure tax payer’s dollars are spent most effectively. The results of our analysis showed that both age and blood pressure have strong positive correlations with coronary heart disease. However, our analysis also uncovered that age and blood pressure are also correlated. Given that age has the strongest relationship with coronary heart disease, we suggest public health officials primarily prioritize elderly populations, focusing on the elderly with high blood pressure where possible.

# 

# Citations

Centers for Disease Control and Prevention. 2020. *Heart Disease Facts*. [online] Available at: <https://www.cdc.gov/heartdisease/facts.htm> [Accessed 22 November 2020].

Centers for Disease Control and Prevention. 2020. *About Heart Disease*. [online] Available at: <https://www.cdc.gov/heartdisease/facts.htm> [Accessed 22 November 2020].

Mayo Clinic. 2020. *Heart Disease - Symptoms And Causes*. [online] Available at: <https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118> [Accessed 22 November 2020].

S. S.,Mahmood,D.Levy, R.S.,Vasan, & T.J. Wang. "The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. "*Lancet*, vol. 383,9921 ,pp. 999–1008. 2014.

WB. Kannel , T. Gordon , MJ. Schwartz. “Systolic versus diastolic blood pressure and risk of coronary heart disease. The Framingham study.” *Am J Cardiol*. Vol. (4):335-46. 1971.

E. Pinto. “Blood pressure and ageing. “ *Postgrad Med J.* Vol. 83(976):109-14. 2007