Caitlin McDonold
Springboard Data Science Career Track, Oct 2018 Cohort

Predicting Risk of Cardiovascular Disease

June 14th, 2019

INTRODUCTION

Cardiovascular diseases, such as coronary artery disease, are the leading cause of death worldwide for both men and women. In the United States, 1 in every 4 deaths is caused by cardiovascular disease (CVD, otherwise known as heart disease). Every year, roughly 790,000 Americans suffer from a heart attack - that's one heart attack every 40 seconds. However, it is estimated that up to 90% of heart disease cases are preventable. Known risk factors for heart disease include high cholesterol, high blood pressure, inactivity, poor diet, and smoking. Early identification of those at higher risk of developing heart disease is critical so that preventative interventions such as lifestyle changes and medication can be implemented.

The goal of this project is to use a large cardiovascular dataset to examine trends between the presence of cardiovascular disease and different health data such as gender, blood pressure, BMI, and cholesterol level. I will then build predictive models that use the most relevant features to classify patients with heart disease and predict risk of heart disease based on the combination of risk factors.

DATA OVERVIEW

The data used for this project was obtained during medical examinations and includes objective data such as height, weight, and blood pressure that was taken at the time of the examination as well as subjective data such as alcohol intake and physical activity level that was reported by the patient. There are a total of 70,000 observations (i.e. patient records) in the dataset and 11 features in addition to the target variable.

Features:

1. Age | Objective Feature | age | int (days)
2. Height | Objective Feature | height | int (cm) |
3. Weight | Objective Feature | weight | float (kg) |

4. Gender | Objective Feature | gender | categorical code |1: female, 2: male
5. Systolic blood pressure | Examination Feature | ap_hi | int |
6. Diastolic blood pressure | Examination Feature | ap_lo | int |
7. Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
8. Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
9. Smoking | Subjective Feature | smoke | binary | 0: no, 1: yes
10. Alcohol intake | Subjective Feature | alco | binary |
11. Physical activity | Subjective Feature | active | binary |
12. Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

The dataset for this project is available from Kaggle: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

DATA CLEANING

The dataset was relatively clean overall, and a first examination of the dataset revealed that there were no null entries for any of the features. Before analyzing the data further, I first converted the units of age from days to years and of height from centimeters to meters to be more easily interpretable. Additionally, I converted the gender feature from a numerical to a categorical data type and renamed the categories 'F' and 'M,' also for easy readability. I also added a new column that calculated the body mass index (BMI) for each individual. The BMI is a useful measure for this dataset as it uses an individual's weight in relation to their height as a rough estimate of body fat; BMIs are used to categorize individuals as underweight, normal weight, overweight, or obese.

Next, I plotted histograms of the continuous features to examine their distributions and identify outliers (Figure 1). The height distribution immediately jumps out, as there are a number of individuals that are either very short (52 individuals under 3.9 ft) or very tall (1 over 8 ft). A number of different medical conditions that may result in either a very short or very tall stature could also disproportionately affect cardiovascular health and thus complicate analysis of the dataset; I have therefore removed any individuals whose height is less than 1.2 m (~4 ft) or more than 2.1 m (~6.9 ft) from the dataset. There were also many blood pressure values that did not make sense, such as a systolic blood pressure of 16020 or -150. Such values are most likely the cause of transcription errors such as dropped decimal points (i.e. 160.20, not 16020) or incorrect negative signs (i.e. 150, not -150), but as I cannot be sure, all such values were dropped from the dataset.

Finally, I also dropped the observations with the five lowest BMIs, as they were unrealistic and likely the result of a transcription error (for example, a 5’8” tall man who

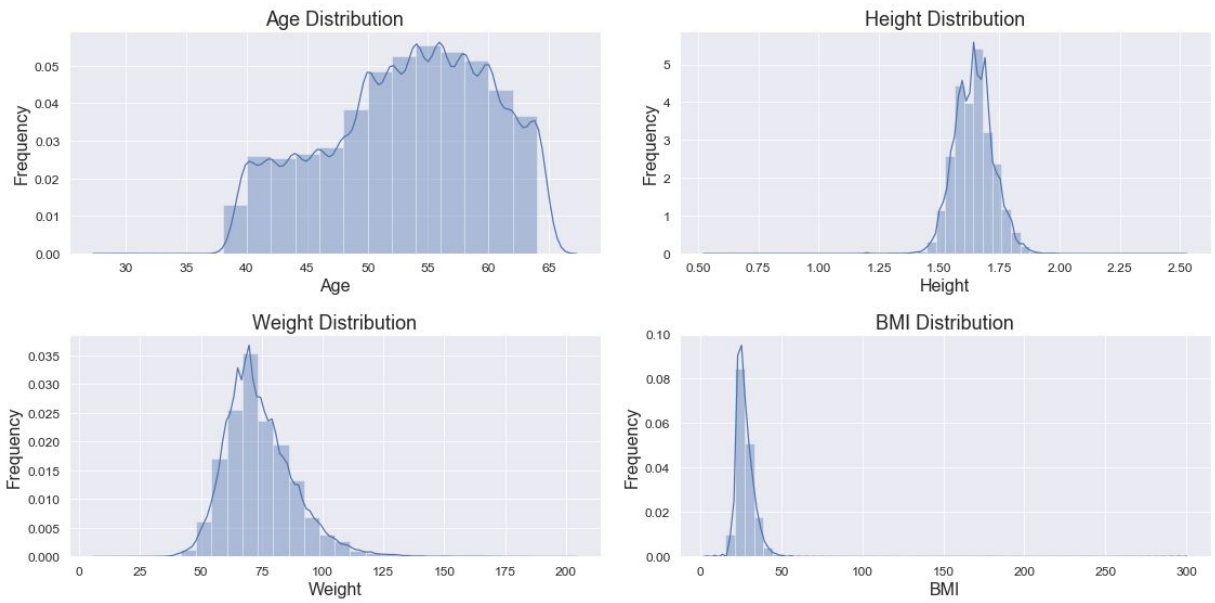
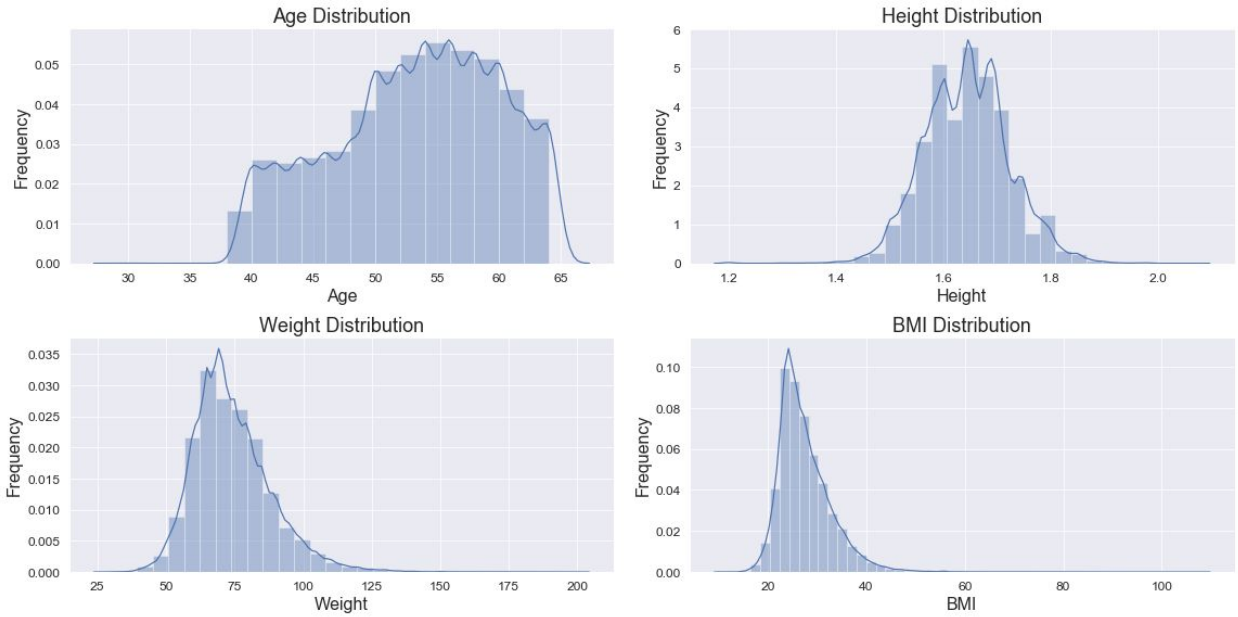


Figure 1: Histograms of age, height, weight, and BMI using full dataset.

weighed only 24 pounds). I then re-plotted the histograms to check the distributions after removing these outliers (Figure 2). As you can see, the height distribution in particular looks much better, though the data is still heavily skewed towards larger weights and BMIs. However, this is also true of the global population these days, so the filtered dataset is likely a reasonable representation of modern human populations.



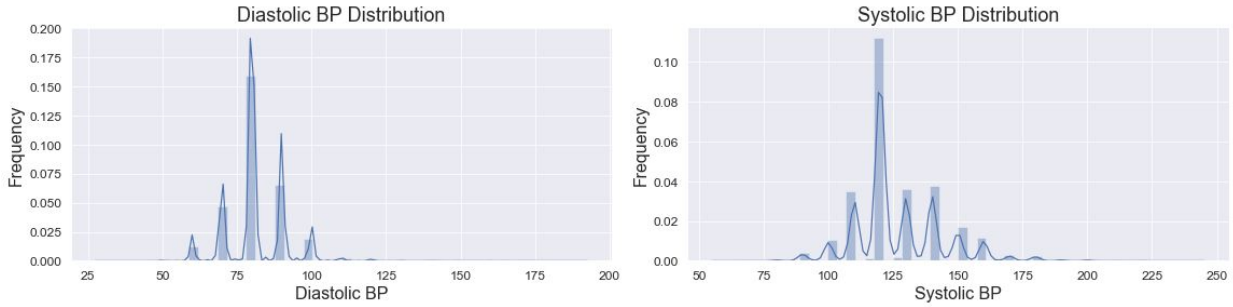


Figure 2: Histograms of age, height, weight, BMI, and blood pressure (BP) using filtered dataset with outliers removed.

A total of 1,274 outlier observations were removed from the dataset, leaving 68,726 observations for subsequent data analysis and model building.

EXPLORATORY DATA ANALYSIS

The next several sections explore the dataset by groups of related features. First, I will explore connections between age, gender, and cardiovascular disease, then move on to look at the effects of height, weight, and BMI, followed by measures of cholesterol and glucose levels in the blood. Finally, I will wrap up with a discussion of the self-reported features around smoking, alcohol consumption, and physical activity.

I. Age, Gender, and Cardiovascular Disease

I began by first examining the number of men and women in the dataset and found that women comprise 65% of the dataset, while men only comprise 35% - in other words, women outnumber men by nearly 2:1 (Figure 3, plot 1). This is interesting in itself - are

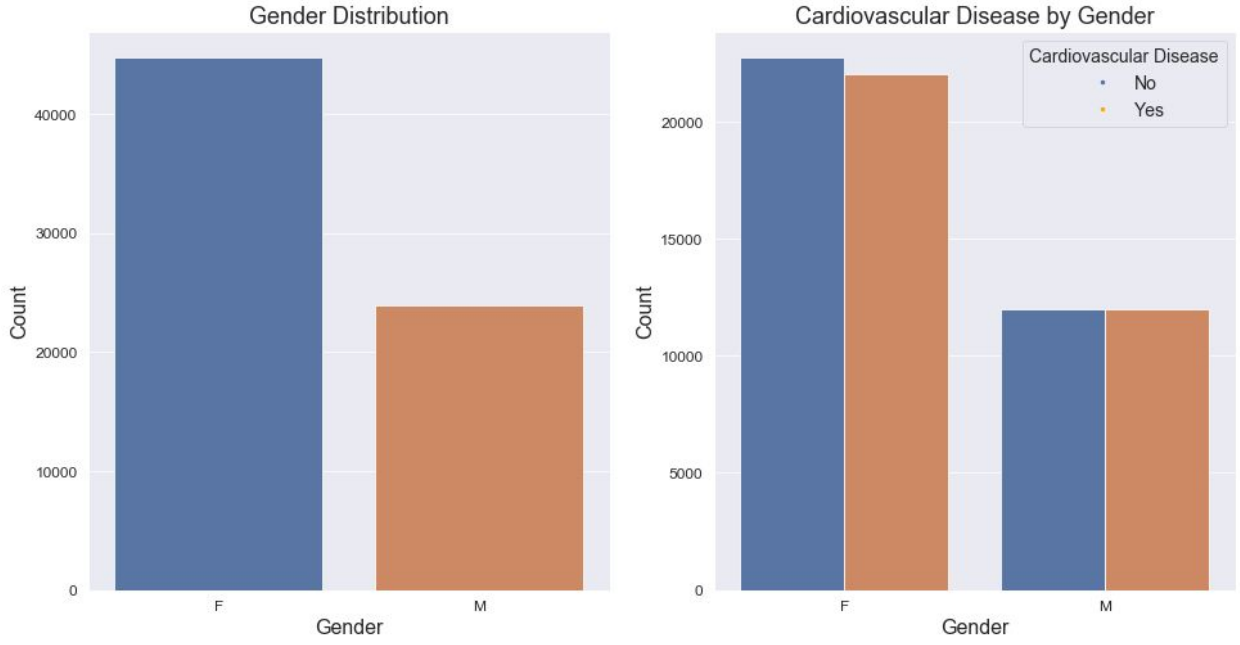


Figure 3: Left - Overall gender distribution. Right - Gender distribution by cardiovascular disease.

women more likely to visit the doctor than men? As the source of the dataset was not given, this question must remain unanswered. By dividing gender further into those with cardiovascular disease and those without, however, we can see that women may have a slightly lower risk of cardiovascular disease than men (Figure 3, plot 2). A chi-square test was then performed to assess whether gender and cardiovascular disease are independent, yielding a chi-square value of 3.62 and a p-value of 0.057, just above the 95% threshold for significance. Therefore, gender is not correlated with cardiovascular disease (i.e. gender and cardiovascular disease are independent).

Moving on to age, we can see from Figure 2 that the dataset is heavily biased towards older individuals. There are only 4 individuals out of over 68,000 who are 30 years old or younger; the next youngest age in the dataset is 39 years old. The oldest person in the dataset is just under 65 years old, so the dataset is only representative of adults from middle-age through to retirement age. Plotting the empirical cumulative distribution functions (ECDFs) for age by cardiovascular disease shows a definite age shift from those without cardiovascular disease and to those with it, indicating that the risk of cardiovascular disease increases with age (Figure 4). The significance of this age shift can be assessed using Student's t-test:

H_0: The mean age of those with or without cardiovascular disease is the same.

H_A: The mean age of those with or without cardiovascular disease is different.

T-statistic: -64.68 p-value: 0.0

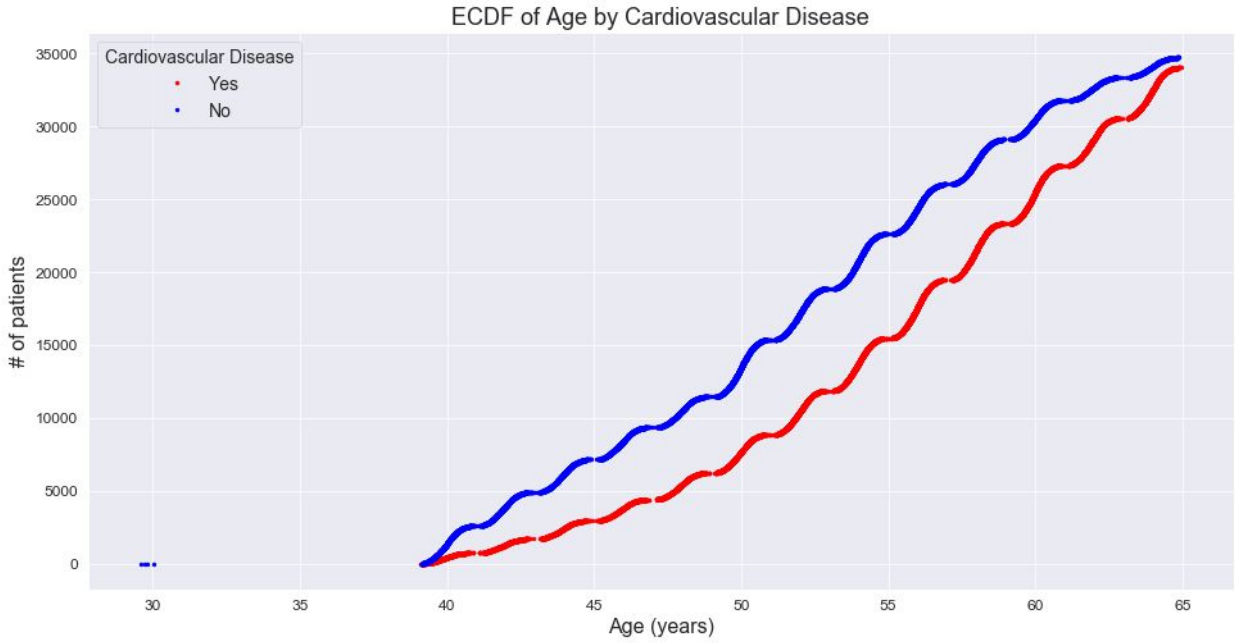


Figure 4: ECDF of age by presence or absence of cardiovascular disease.

As the p-value is well below the 95% significance threshold (0.05), we reject the null hypothesis. Thus, the average age of those with heart disease is higher than for those without heart disease.

The age distribution for those with or without cardiovascular disease can be further visualized using boxen plots. Boxen plots, also known as letter-value plots, are ideal for visualizing the distributions of large datasets, as they afford more precise estimates of quantiles beyond quartiles. A boxen plot is more useful than a normal box plot in this situation because the increased number of quantiles provide more detailed information about the shape of the distribution, especially in the tails. From this plot, we can see not only that older individuals are more likely to have cardiovascular disease on average, but also that the entire age distribution is heavily shifted to the higher age ranges (Figure 5, left plot). We can further subdivide the age distribution by gender, revealing that women tend to develop cardiovascular disease later than men (Figure 5, right plot).

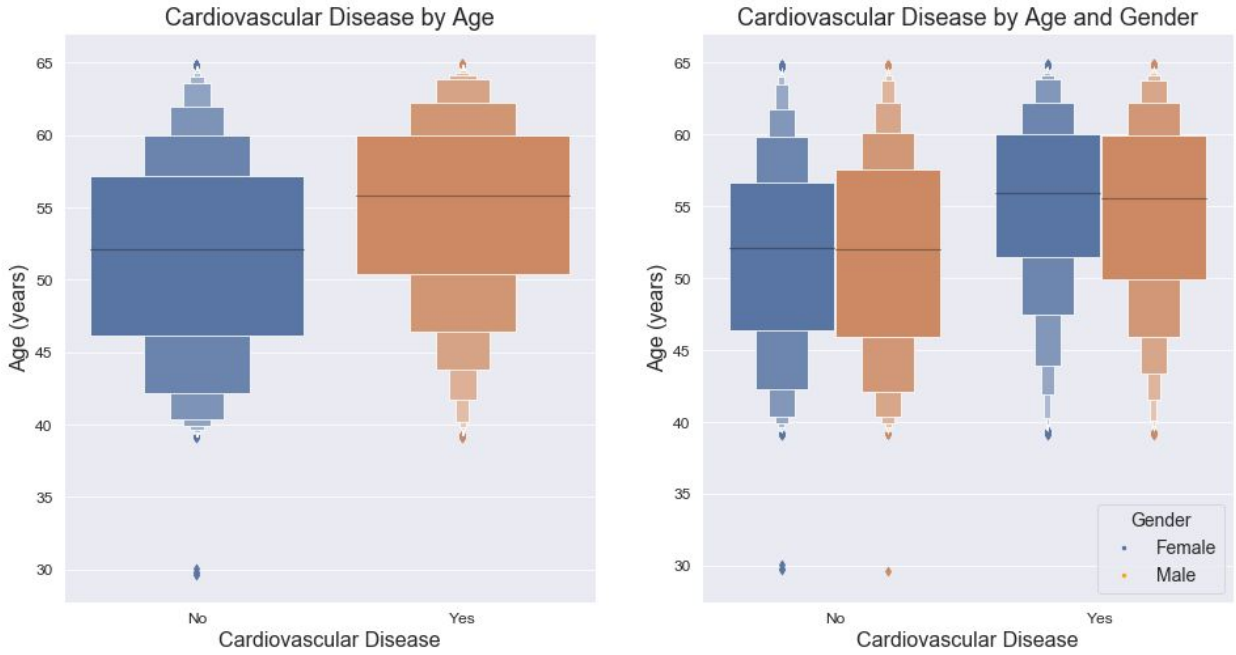


Figure 5: Left - Boxen plot distribution of age by cardiovascular disease. Right - Boxen plot distribution of age by gender and cardiovascular disease.

A t-test can again be used to assess the hypothesis that the average age of women with heart disease is higher than the average age of men with heart disease, yielding a t-statistic of 8.42 with a p-value of 0.0 (or more precisely, 3.9×10^{-17}). Therefore, women with heart disease do have a higher average age than men with cardiovascular disease. Performing the analogous t-test for men and women without heart disease results in a t-statistic of 1.37 and a p-value of 0.17, demonstrating that men and women *without* heart disease have the same average age, statistically speaking.

II. Effect of Height, Weight, and BMI on Cardiovascular Disease

Next, I examined the relationships between height, weight, BMI, and cardiovascular disease. A scatterplot of height vs. weight, colored by cardiovascular disease, shows that heart disease is more prevalent at higher weights, which is perhaps unsurprising (Figure 5). The linear regression lines were also determined and have been overlaid

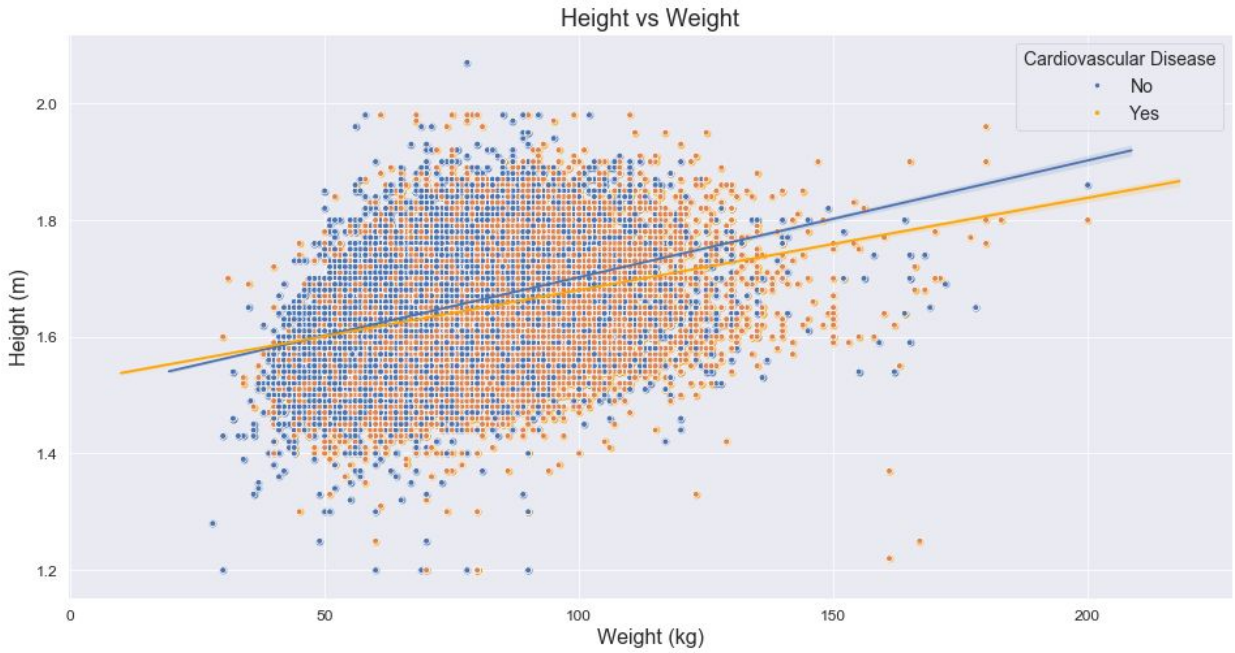


Figure 6: Scatterplot of height vs weight overlaid with linear regression lines; blue: no heart disease, orange: heart disease.

on the scatterplot. The equations for these lines are as follows:

No Heart Disease: $y = 0.0020x + 1.50$

Heart Disease: $y = 0.0016x + 1.52$

The slopes of the lines also seem to indicate that weight has more of an effect on cardiovascular disease than height.

ECDFs can be used to further explore potential differences in height and weight by cardiovascular disease. First, the ECDFs for height by cardiovascular disease look fairly similar, indicating that height might not be an important factor in cardiovascular disease (Figure 7). The ECDFs for weight, however, show a definite shift to higher weight for those with cardiovascular disease (Figure 8). As BMI is a function of both height and weight, it is not surprising to see that the ECDF for those with cardiovascular disease also show a shift to higher BMIs compared to the ECDF for those without cardiovascular disease (Figure 9).

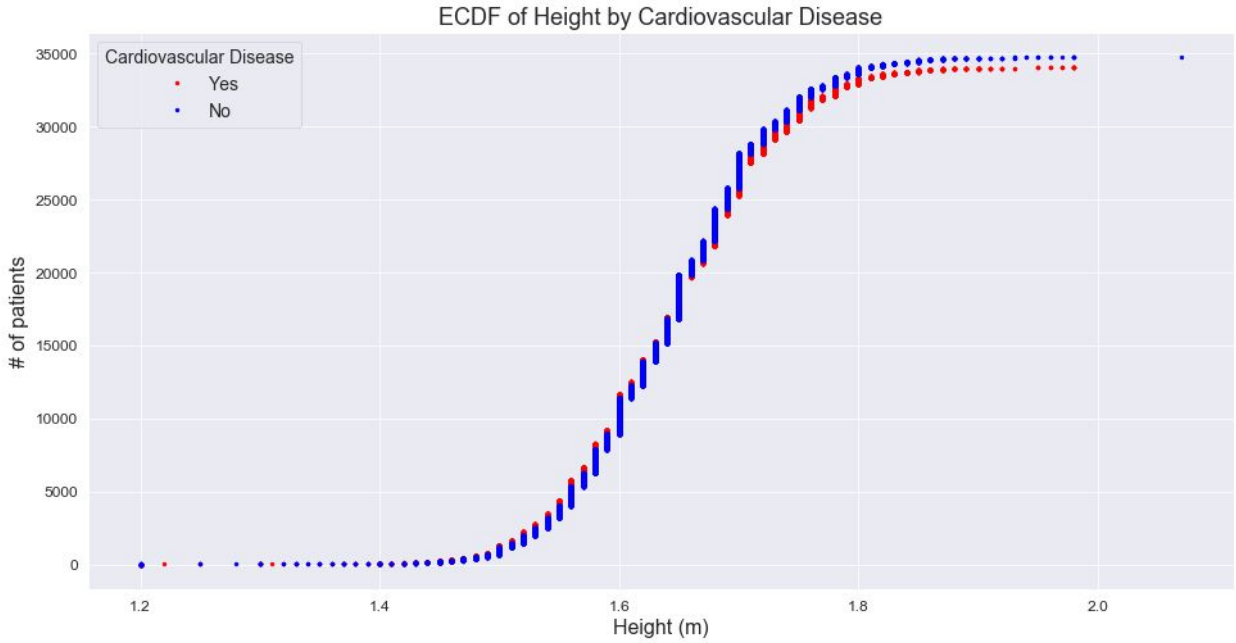


Figure 7: ECDF for height by cardiovascular disease.

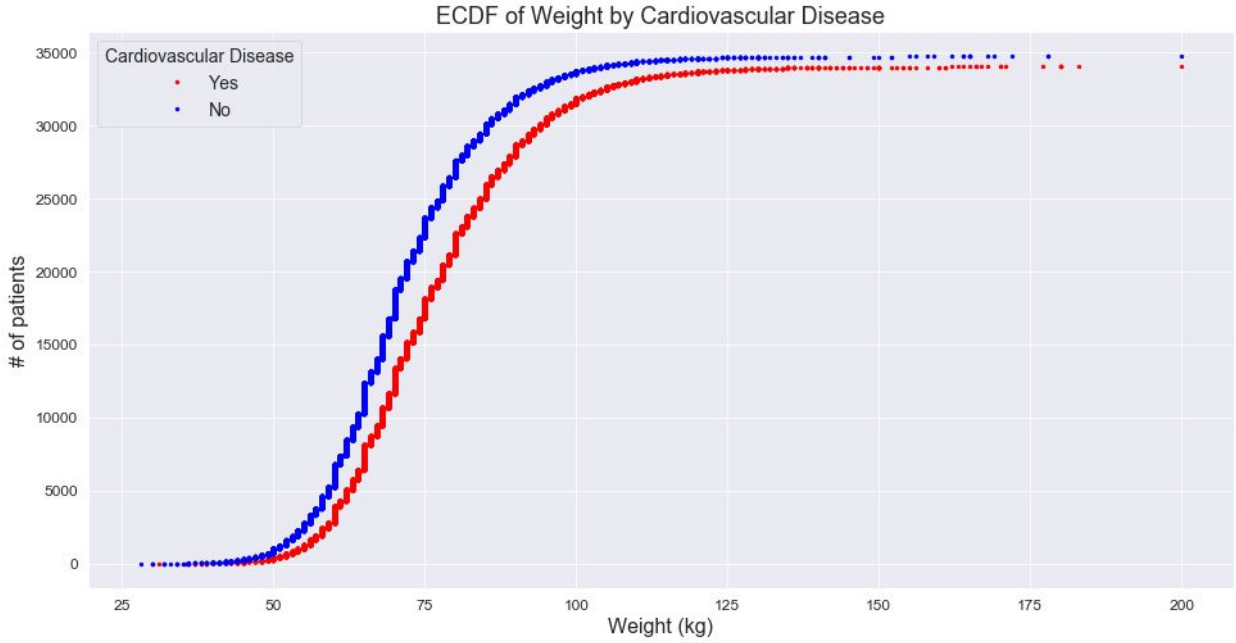


Figure 8: ECDF for weight by cardiovascular disease.

One of the main advantages of working with such a large sample size, increased precision, can also be a disadvantage by enabling detection of differences that are so small as to be meaningless in a practical sense. In this case, the mean height of those without cardiovascular disease is 1.645 m, while the mean height of those with cardiovascular disease is 1.643 m. While a t-test indicates that the mean height is statistically different between those with cardiovascular disease and those without (t-statistic: 3.10 ; p-value: 0.0019), the difference, at only 2 cm, is not *practically* significant. In addition, the boxen plots for height by cardiovascular disease are nearly

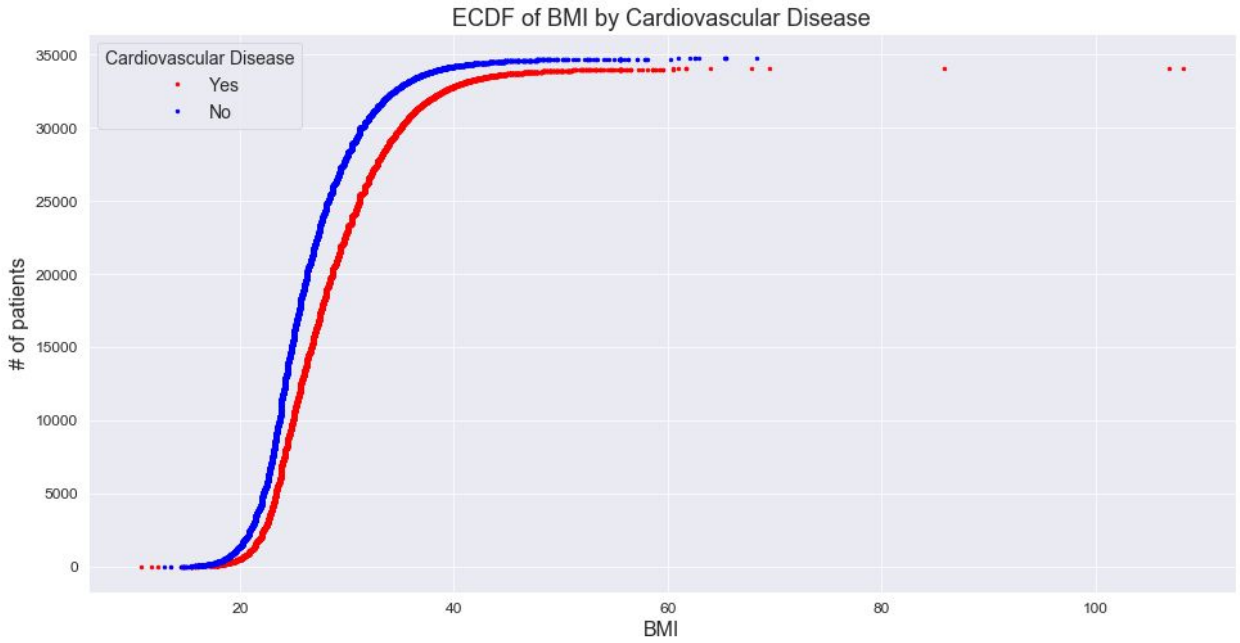


Figure 9: ECDF for BMI by cardiovascular disease.

indistinguishable, further emphasizing that the height distributions are not different in a meaningful sense (Figure 10, plot 1). In contrast, the boxen plots for both weight and BMI segregated by cardiovascular disease show clear shifts in both the mean values and the overall distributions between those with or without cardiovascular disease (Figure 10, plots 2 and 3). Performing t-tests confirms that the shifts in the average weight (t-statistic: 47.99; p-value: 0.0) and average BMI (t-statistic: 50.61; p-value: 0.0) are statistically significant, and both p-values are far smaller than the p-value for height.

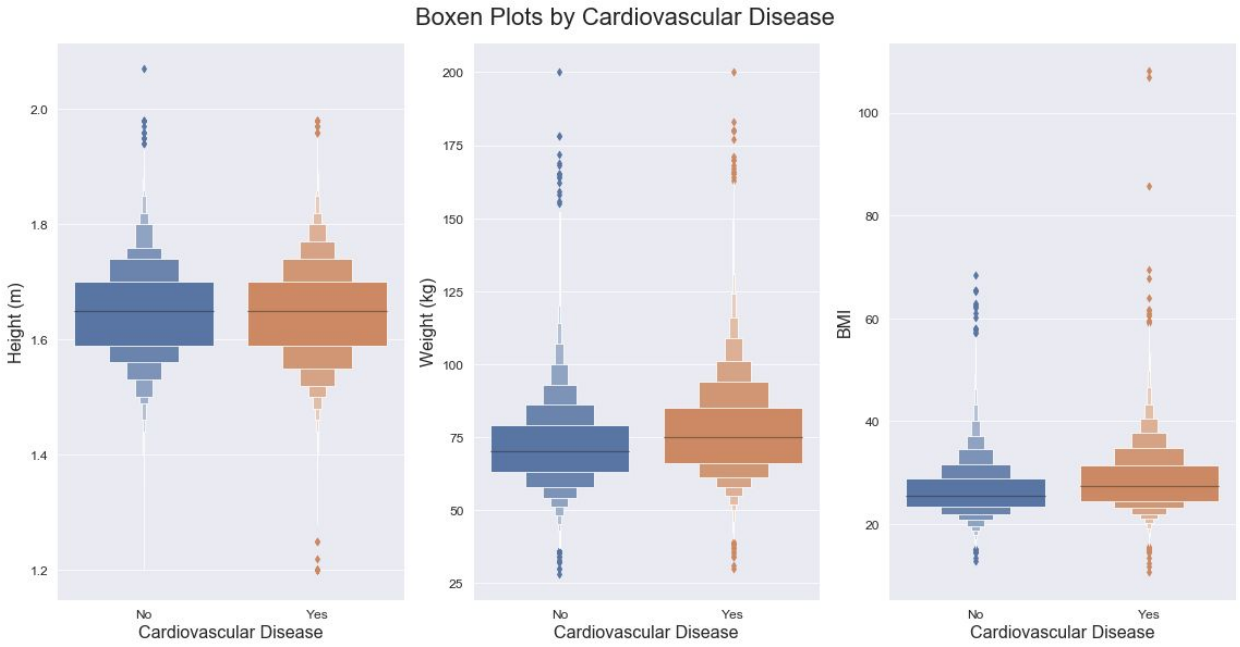


Figure 10: Boxen plots of height (plot 1), weight (plot 2), and BMI (plot 3) by cardiovascular disease.

III. High Blood Pressure Increases Risk of Cardiovascular Disease

Blood pressure is a measure of how much pressure your blood exerts on the walls of your arteries. Systolic blood pressure indicates the amount of pressure when your heart beats, while diastolic blood pressure indicates the amount of pressure when your heart is resting between beats. As expected, a scatterplot of systolic vs diastolic blood pressure shows a clear positive correlation (Figure 11), which can be confirmed by calculating the Pearson correlation coefficient (0.697, p-value = 0.0).

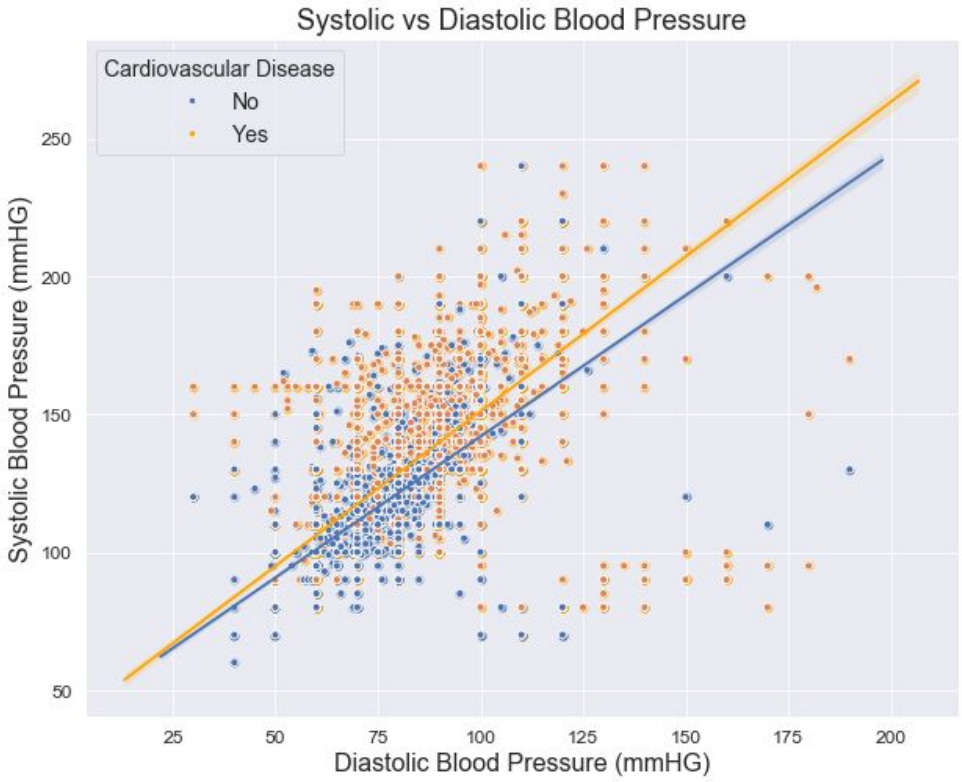


Figure 11: Scatterplot of systolic vs diastolic blood pressure overlaid with linear regression lines; blue: no heart disease, orange: heart disease.

The linear regression lines for the patients with cardiovascular disease (in orange) and those without it (in blue) were also determined and overlaid on the scatterplot (without CVD: y = 1.0233x + 39.57; with CVD: y = 1.1215x + 38.87). While patients with cardiovascular disease have higher systolic blood pressure on average, there is an interesting subset that have high diastolic blood pressure higher and low systolic blood pressure (Figure 11, bottom left). As it is unlikely that blood pressure would be *higher* when the heart is resting than when it is beating, the data points may indicate transcription errors in the dataset.

The ECDFs for systolic (Figure 12) and diastolic (Figure 13) blood pressure for those with cardiovascular disease are both shifted to the left (higher blood pressure) than the

ECDFs for those without cardiovascular disease; however, the shift to higher blood pressure is particularly pronounced for systolic blood pressure.

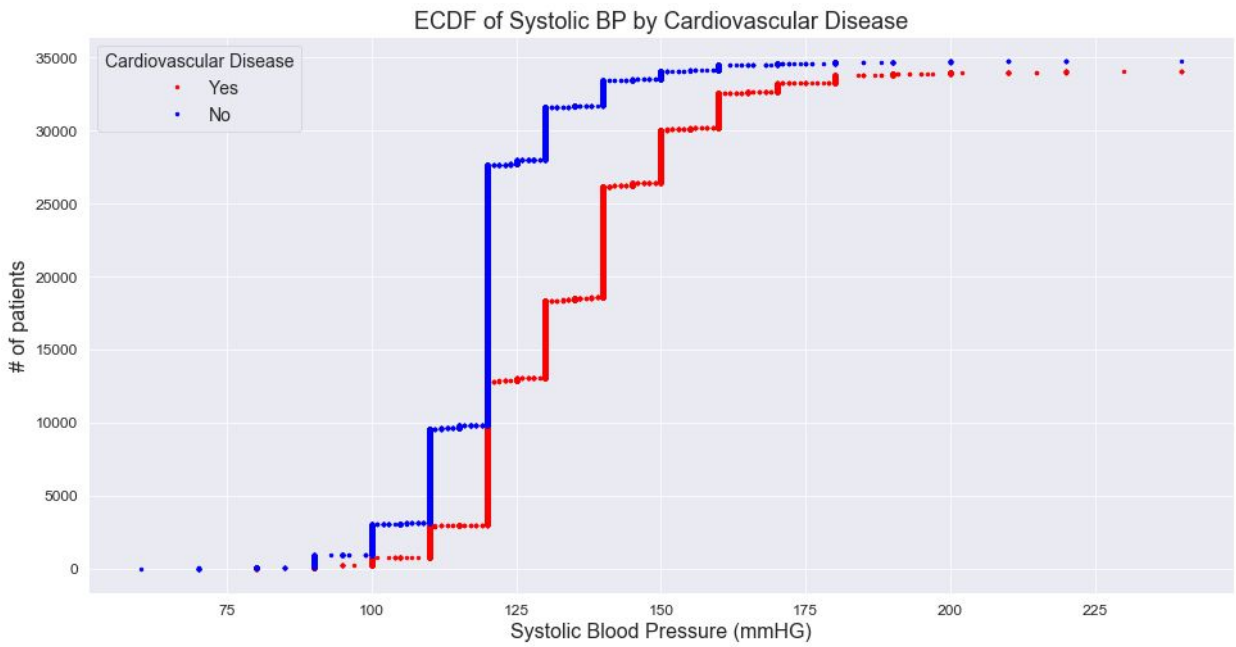


Figure 12: ECDF of systolic blood pressure for patients with or without CVD.

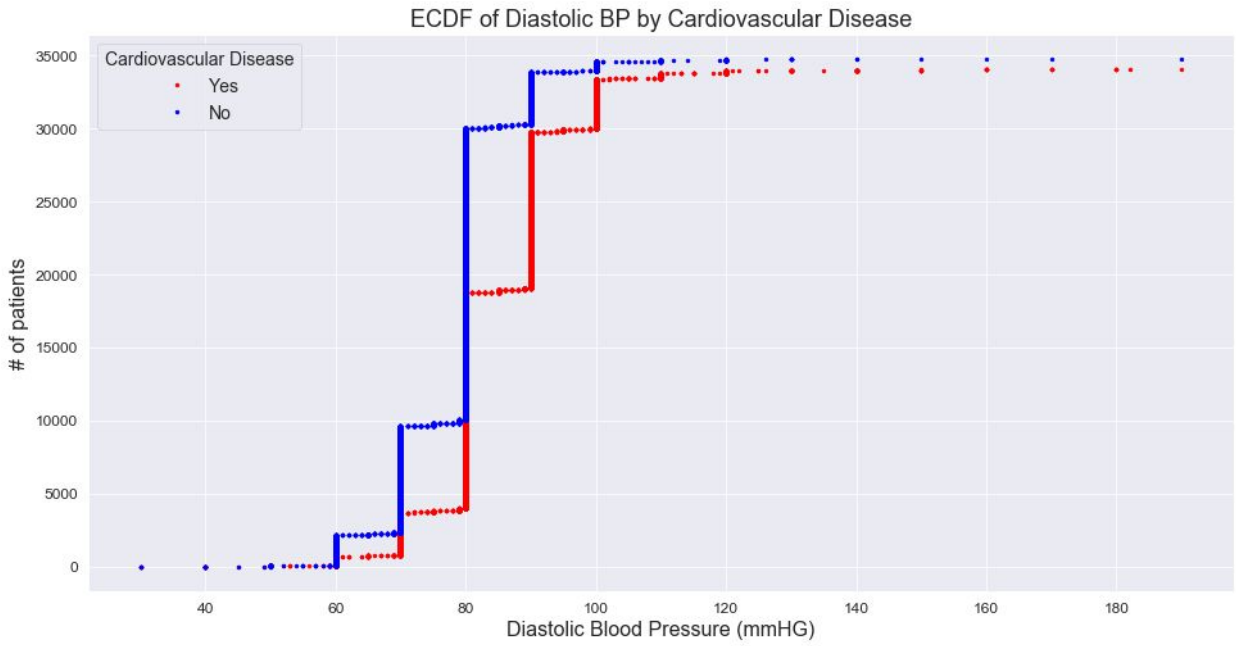


Figure 13: ECDF of diastolic blood pressure for patients with or without CVD.

The shift in mean blood pressure observed in patients with cardiovascular disease is even more apparent in boxen plots of the diastolic and systolic blood pressure (Figure 14). Student's t-tests confirmed that the mean shifts are statistically significant for both diastolic and systolic blood pressure (Table 1). The strength of the correlation

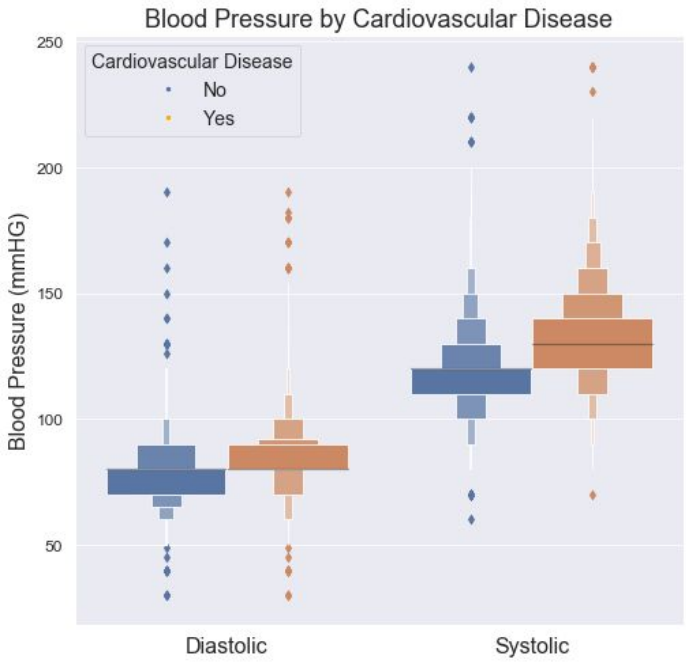


Figure 14: Boxen plots of diastolic and systolic blood pressure by cardiovascular disease.

between blood pressure and cardiovascular disease can also be measured using the point biserial correlation coefficient, which measures the relationship between a binary variable (i.e. cardiovascular disease) and a continuous variable (i.e. blood pressure). The point biserial correlation coefficients show that both systolic and diastolic blood pressure are strongly correlated with cardiovascular disease (Table 1). The point biserial correlation coefficients for the continuous features discussed previously (age, height, weight, and BMI) were also calculated and compared to those for systolic and diastolic

	Mean without CVD	Mean with CVD	T-statistic	p-value	Point Biserial Correlation	p-value
Age	51.72	54.96	64.68	0.0	0.425	0.0
Height (m)	1.645	1.643	3.1	0.0019	-0.012	0.0019
Weight (kg)	71.57	76.72	47.99	0.0	0.18	0.0
BMI	26.47	28.46	50.61	0.0	0.19	0.0
Systolic BP (mmHG)	119.56	133.82	123.14	0.0	0.425	0.0
Diastolic BP (mmHG)	78.17	84.65	93.11	0.0	0.335	0.0

Table 1: Summary statistics for continuous features for patients with and without cardiovascular disease.

blood pressure. Of the continuous features, age and systolic blood pressure are most strongly correlated with cardiovascular disease, followed by diastolic blood pressure.

IV. Cholesterol and Glucose Levels

Cholesterol and glucose levels have been categorized in the dataset as either 'normal,' 'above normal,' or 'well above normal' instead of reporting the exact values. While having the data in this form is slightly less useful than the raw values would be, analysis of the data can still yield valuable insights. Count plots for both the cholesterol and glucose data were used to examine the distribution across the three levels for patients with or without cardiovascular disease, revealing that the proportion of patients with CVD increases as either the cholesterol or glucose levels increase (Figure 15). The

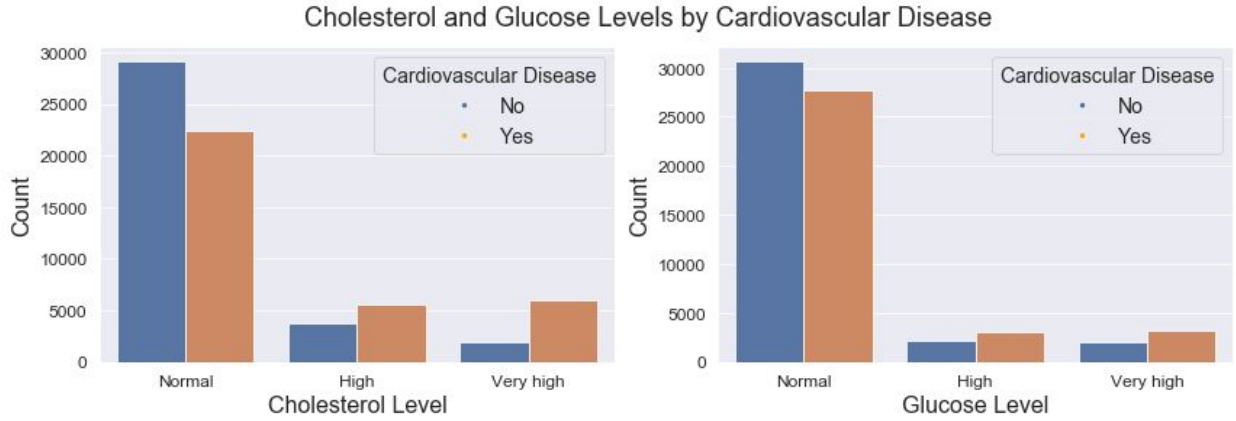


Figure 15: Count plots of individuals with or without CVD by cholesterol or glucose level.

relationship between cholesterol or glucose level and cardiovascular disease is even more apparent if we instead plot the percentage of patients with cardiovascular disease at each level (Figure 16). Among individuals with normal cholesterol levels, 43.56% also have cardiovascular disease. However, this number drastically increases to 76.28% among individuals with the highest cholesterol levels (Table 2). A similar, though less dramatic, increase is observed between individuals with normal glucose levels (47.57%)

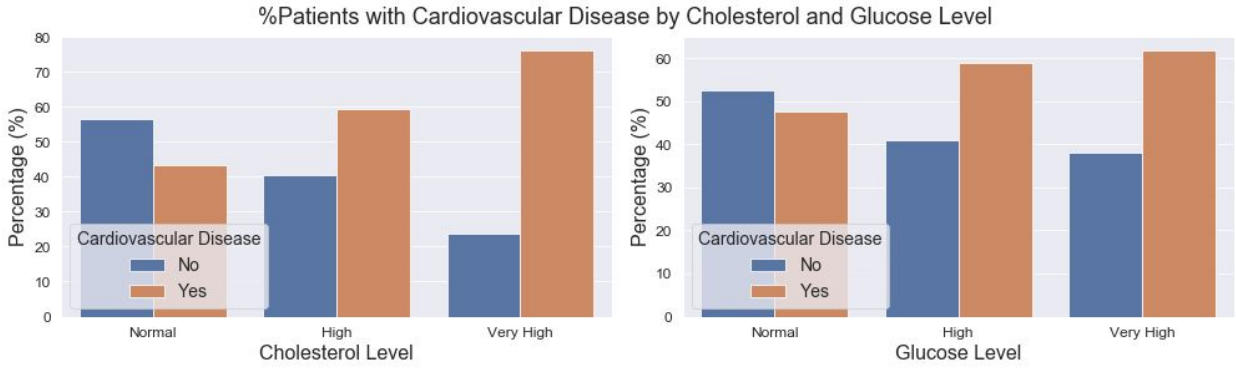


Figure 16: Percentage of patients with or without CVD at different cholesterol and glucose levels.

and those with very high levels (61.88%). Chi-square tests for independence can be used to confirm the statistical significance of the observed increases in cardiovascular disease between cholesterol and glucose levels. The p-values for both chi-square tests were essentially 0, providing strong evidence that cardiovascular disease is linked to both cholesterol and glucose level (Table 2). Additionally, the chi-square values were used to calculate the contingency coefficient, which is another measure of the strength

	Cholesterol Level			Glucose Level		
	Normal	**High**	**Very High**	**Normal**	**High**	**Very High**
% with CVD	43.56%	59.63%	76.28%	47.57%	58.87%	61.88%
% without CVD	56.44%	40.37%	23.72%	52.43%	41.13%	38.12%
χ^2	3372.1			586.33		
p-value	0.0			0.0		
χ^2 **correlation**	0.2163			0.092		

Table 2: Percentage of patients with or without CVD at different cholesterol and glucose levels and results of chi-square tests.

of an association between two variables, similar to Pearson's correlation coefficient or the point biserial correlation coefficient. The disadvantage of the contingency coefficient is that it does not reach a maximum value of 1, and cannot therefore be directly compared to other correlation coefficients. However, we can see from the contingency coefficients that cholesterol level and cardiovascular disease are more strongly associated than glucose level and cardiovascular disease.

V. Trends in Self-Reported Features: Smoking, Alcohol Intake, and Physical Activity

The last few features of the dataset were self-reported by the patients and are therefore inherently subjective, making them much less reliable than the objective physical and medical features analyzed previously. In addition, the dataset does not include the questions that the patients were responding to. Take smoking as an example: we might assume that 'No' means that the patient did not consider themselves to be a smoker when the data was collected. However, it is possible that the patient had been a smoker in the past, or perhaps just smokes occasionally. For alcohol intake, it is unknown whether a negative response indicates that the patient never drinks, or is instead a measure of the level of alcohol intake (i.e. 'light' or 'moderate' vs. 'heavy'). There is also no guarantee that the respondents were truthful in their responses. For example, if we examine the count plot for alcohol intake, split into those with CVD and those without,

only a small proportion of the patients responded positively, but it seems highly unlikely that only this small fraction of the patients drink alcohol (Figure 17, plot 1). In general, the distributions across the three features give the impression that most people exercise regularly and don't drink or smoke (Figure 17), but the reliability of these data is questionable. Despite all of these complicating factors, we might still be able to glean some useful information from analyzing this data.

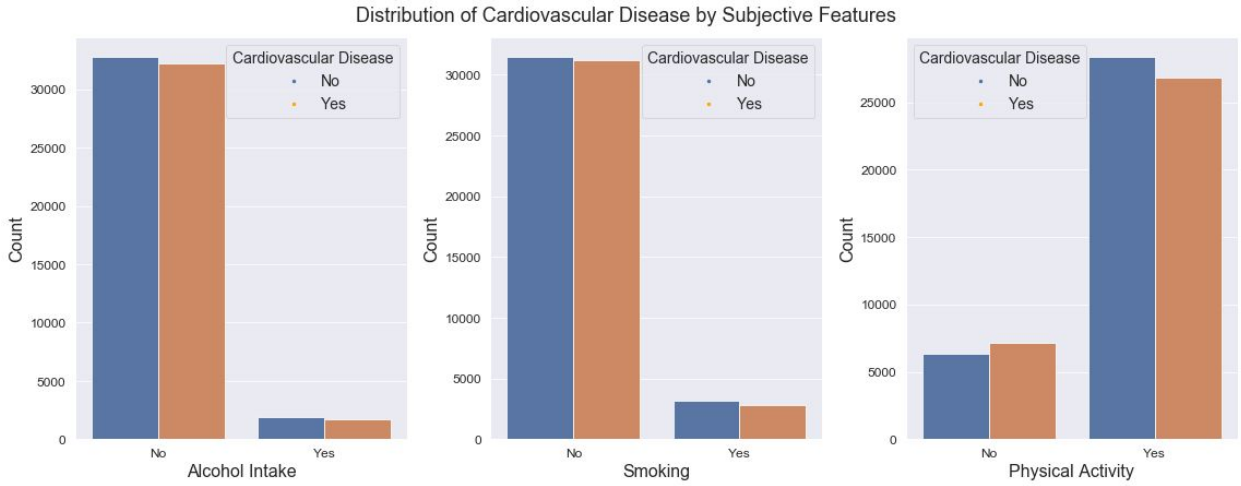


Figure 17: Count plots for smoking, alcohol intake, and physical activity by cardiovascular disease.

Chi-square tests for independence between the subjective features and cardiovascular disease all indicate that cardiovascular disease is in fact associated with all of these features (Table 3). However, the correlations are not necessarily in the expected

	Smoking		Alcohol Intake		Physical Activity	
	No	**Yes**	**No**	**Yes**	**No**	**Yes**
% with CVD	49.74%	46.86%	49.58%	47.80%	53.27%	48.56%
% without CVD	50.26%	53.14%	50.42%	52.20%	46.73%	51.44%
χ^2	18.25		4.36		96.22	
p-value	1.9×10^{-5}		0.037		1.0×10^{-22}	
χ^2 **correlation**	0.0163		0.008		0.0374	

Table 3: Effects of smoking, alcohol intake, and physical activity on cardiovascular disease.

direction. Plotting the percentage of patients with or without CVD rather than simply the number of patients normalizes the data between the positive and negative categories for each feature so that the differences between those with CVD and those without are much easier to see (Figure 18). From these plots, it appears that patients who reported

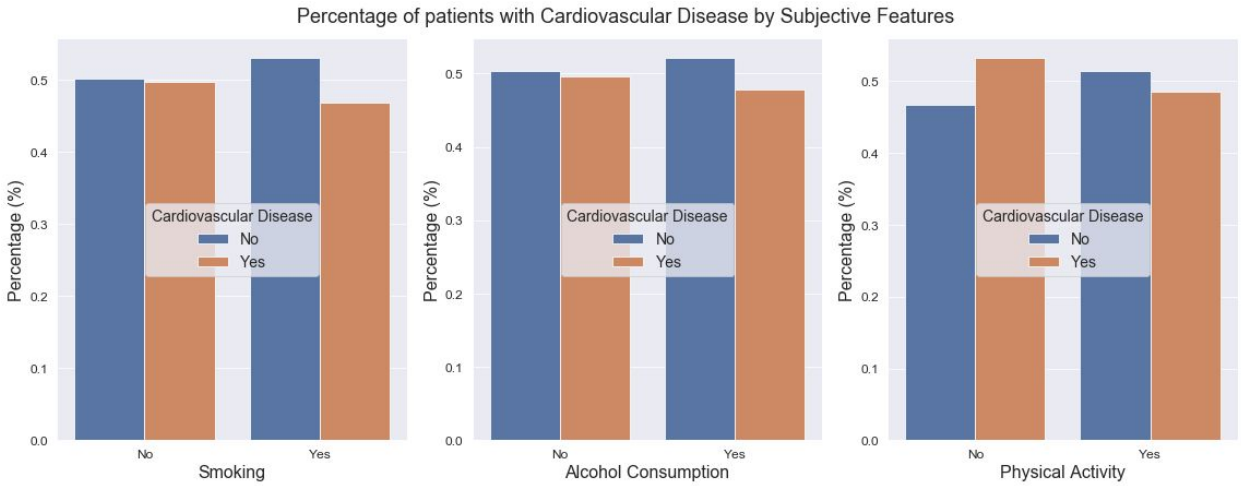


Figure 18: Percentage of patients with or without CVD by smoking, alcohol intake, and physical activity.

'yes' to either smoking or alcohol intake are slightly *less* likely to have cardiovascular disease than patients who responded 'No' to smoking and alcohol intake. Physical activity, on the other hand, shows the opposite effect - patients who reported no physical activity are more likely to have heart disease than those who do exercise. Although the chi-square tests indicate that these differences are statistically significant, the real-world practical significance of these results is much lower when you consider both the magnitude of the difference - only a few percentage points in either direction - and the reliability of the data used in the analysis.

VI. EDA Summary

Overall, this initial analysis of the dataset indicate that age, weight, BMI, systolic and diastolic blood pressure, and cholesterol and glucose levels are all influential factors when it comes to cardiovascular health. Although there is a statistical difference in average height between healthy individuals and those with cardiovascular disease, the effect is too small to have any meaningful applications on its own. However, it might be advantageous to simply include BMI in place of height and weight as a feature during model building, as BMI is derived from an individual's height and weight. As both BMI and weight were correlated with cardiovascular disease to a similar degree, including both would likely be redundant. While gender on its own does not have a strong association with cardiovascular disease, it may have an effect in combination with other factors such as age. Finally, the data around smoking, alcohol intake, and physical activity were self-reported, which leaves a lot of room for interpretation, misunderstanding, or even intentional misinformation, especially as the exact questions put to the participants was not included in the dataset. Therefore, these features should be given less weight or removed altogether when building predictive models.