Predicting Risk of Cardiovascular Disease

CAITLIN MCDONOLD

Cardiovascular Disease

Cardiovascular diseases are the **leading cause of death** worldwide for both men and women.

- **1 in 4 deaths** are caused by cardiovascular disease in the US.
- **~790,000 Americans** suffer a heart attack every year – that's one heart attack every **40 seconds**.

However, up to **90%** of heart disease cases are **preventable**.

The Problem: Early Identification of At-Risk Patients

- Known risk factors include high cholesterol, high blood pressure, inactivity, poor diet, and smoking.
- Many risk factors can be prevented or treated through lifestyle changes or appropriate medication.

Project Goal: Train and build models that can identify patients who are at higher risk of developing cardiovascular disease using data obtained during a typical yearly physical examination.

Client: Primary care physicians and the general public

The Dataset

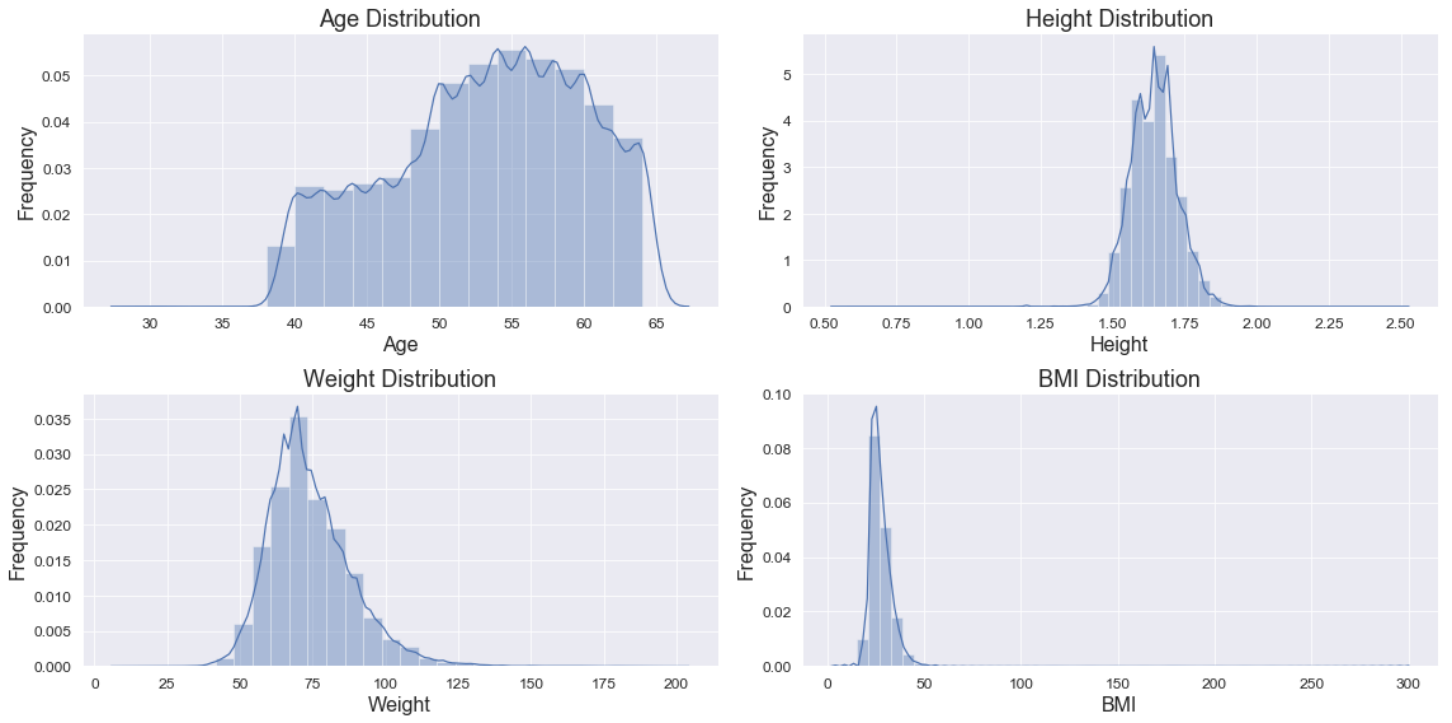
- Data was obtained during medical examinations.
- 70,000 observations and 11 features
- Objective/Examination Features:
    - Age ◦ Height ◦ Weight ◦ Gender
    - Systolic blood pressure ◦ Diastolic blood pressure
    - Cholesterol level ◦ Glucose level
- Subjective Features (self-reported by patients):
    - Smoking ◦ Alcohol Intake ◦ Physical Activity

Data Wrangling

Unit Conversions and Calculation

- Unit of Age were converted from days to years
- Unit of height was converted from centimeters (cm) to meters (m).
- The Body Mass Index (BMI) was calculated for each individual and added as a feature.

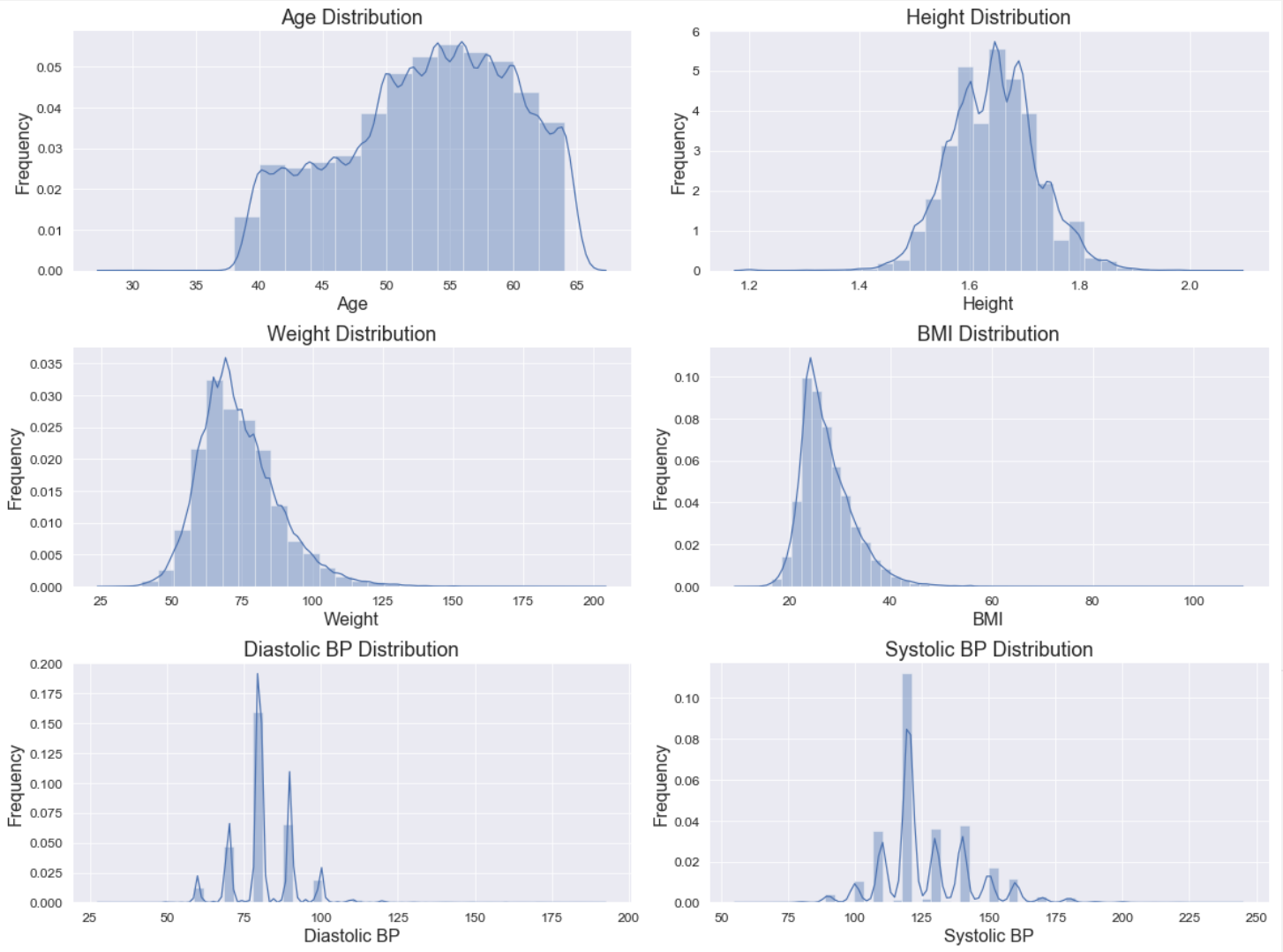
$$BMI = \frac{weight\ (kg)}{height^2 (m^2)}$$



Removing Outliers

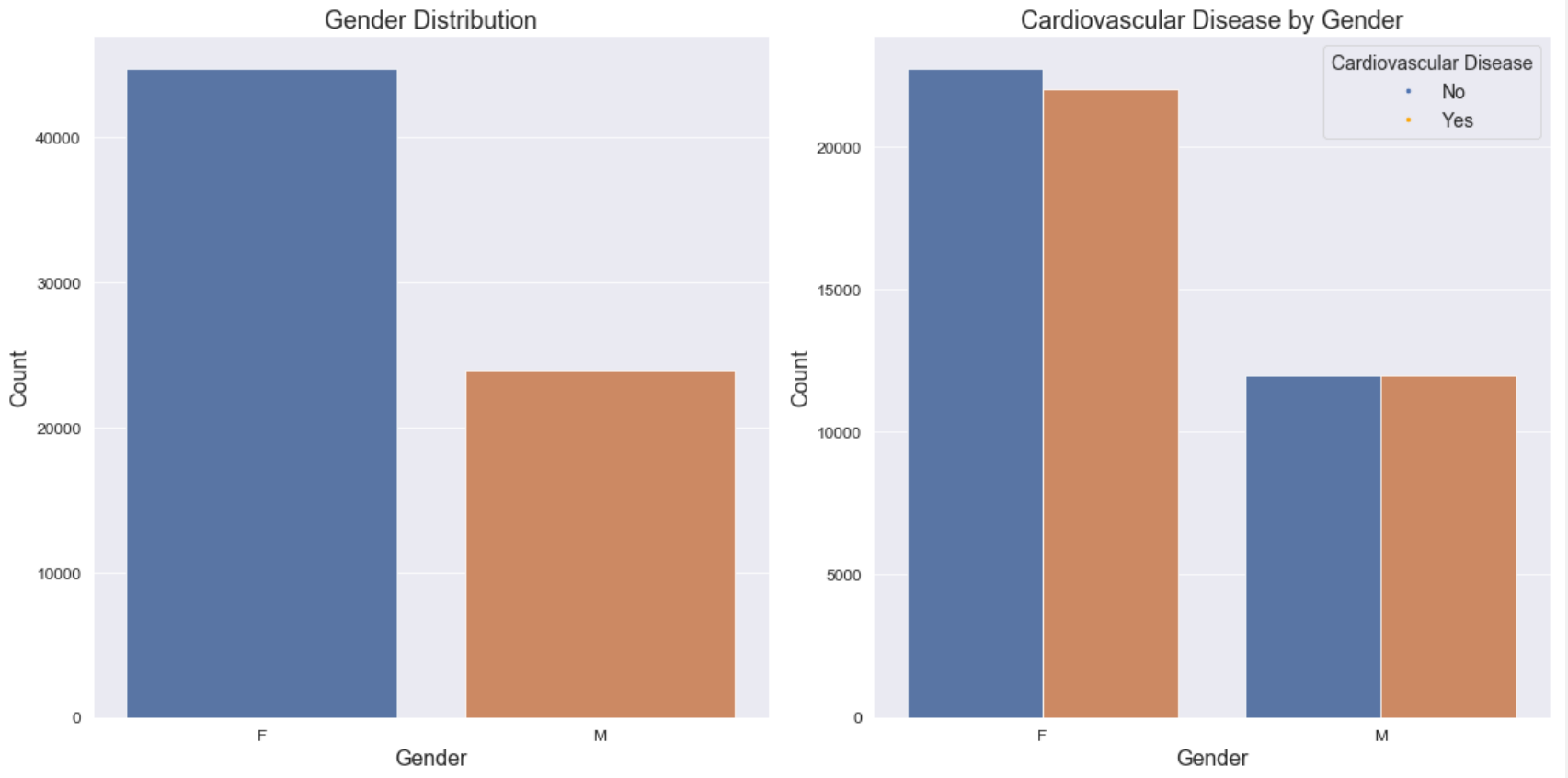
- Individuals that were either very short (52 under 3.9 feet) or very tall (1 over 8 feet).
- 1216 observations with nonsensical blood pressure readings (i.e. 16020 or -150 mmHG).
- 5 individuals with BMI < 10
- Final dataset with all outliers removed contained 68,726 observations.

Exploratory Data Analysis

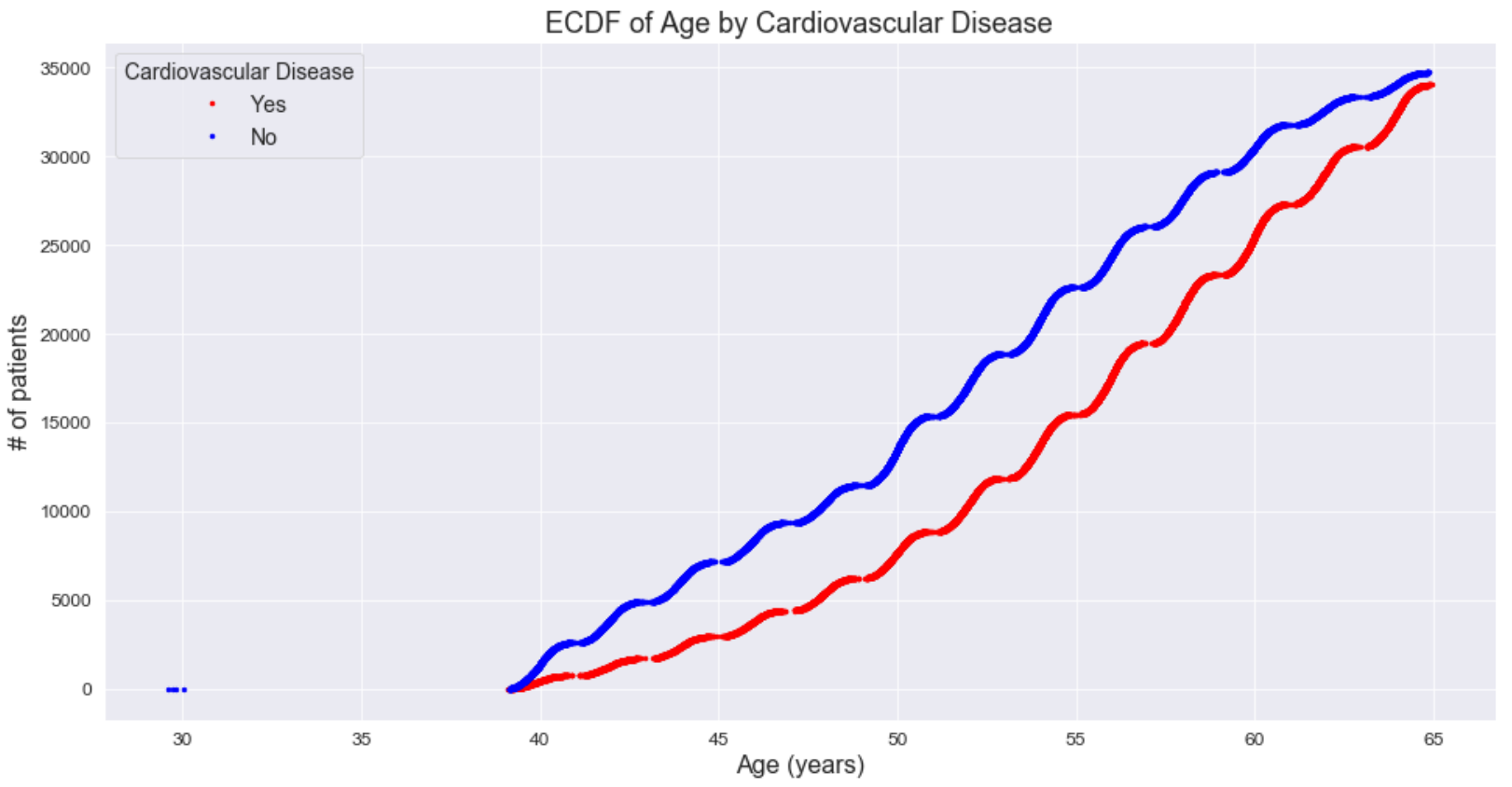


Distributions of Continuous Features

Gender & Cardiovascular Disease



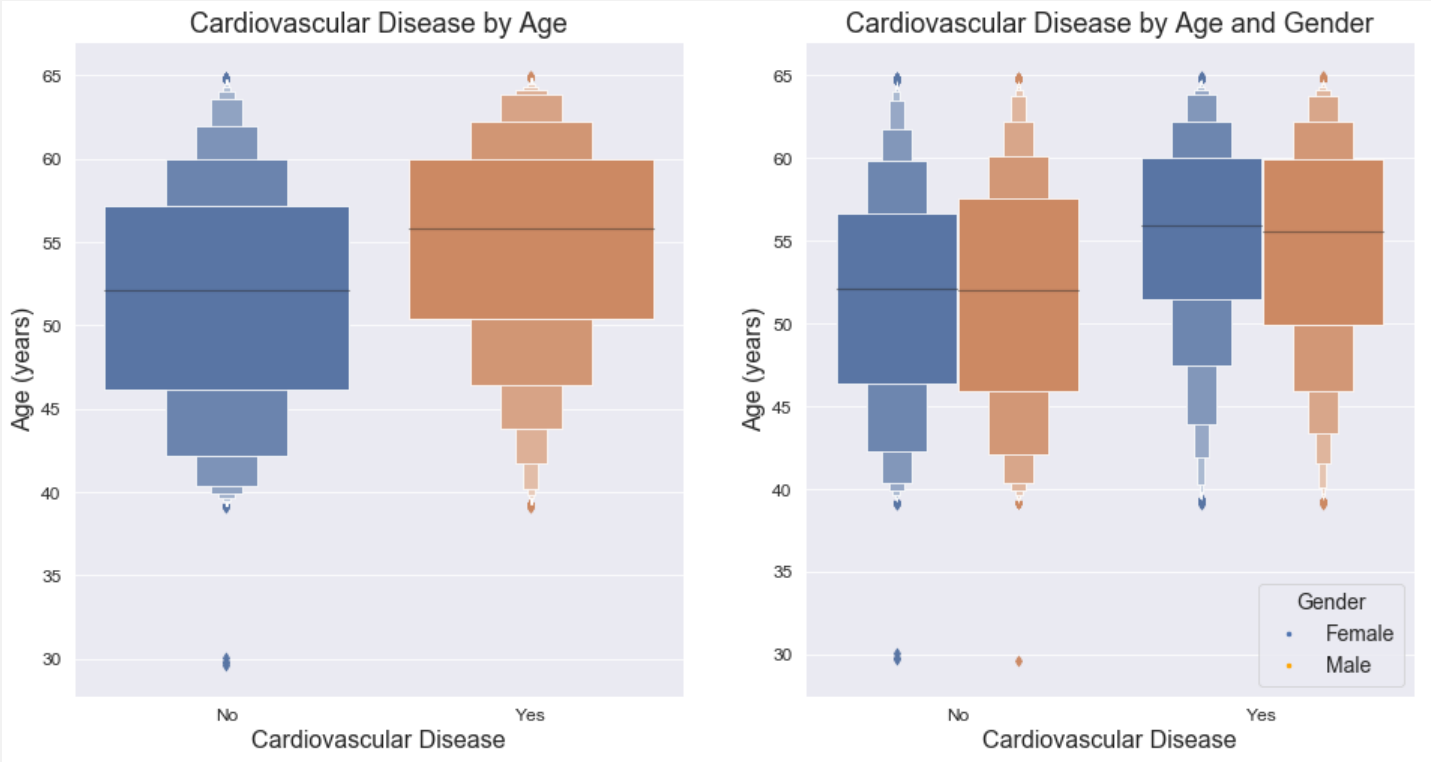
Age & Cardiovascular Disease



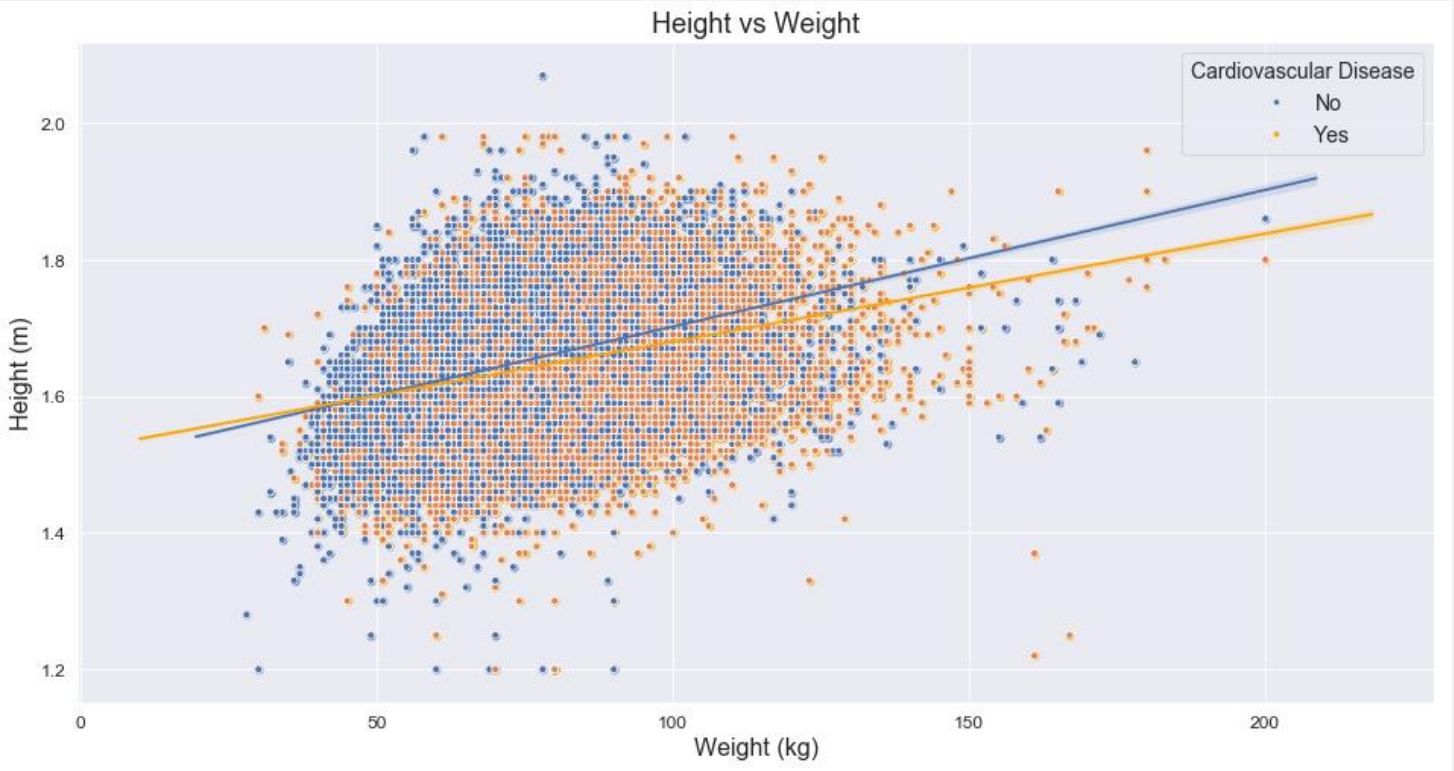
Boxen Plots

- Ideal for visualizing the distributions of large datasets
- Enable more precise estimates of quantiles beyond quartiles
- Provide more details in the tails

Age & Gender Effects

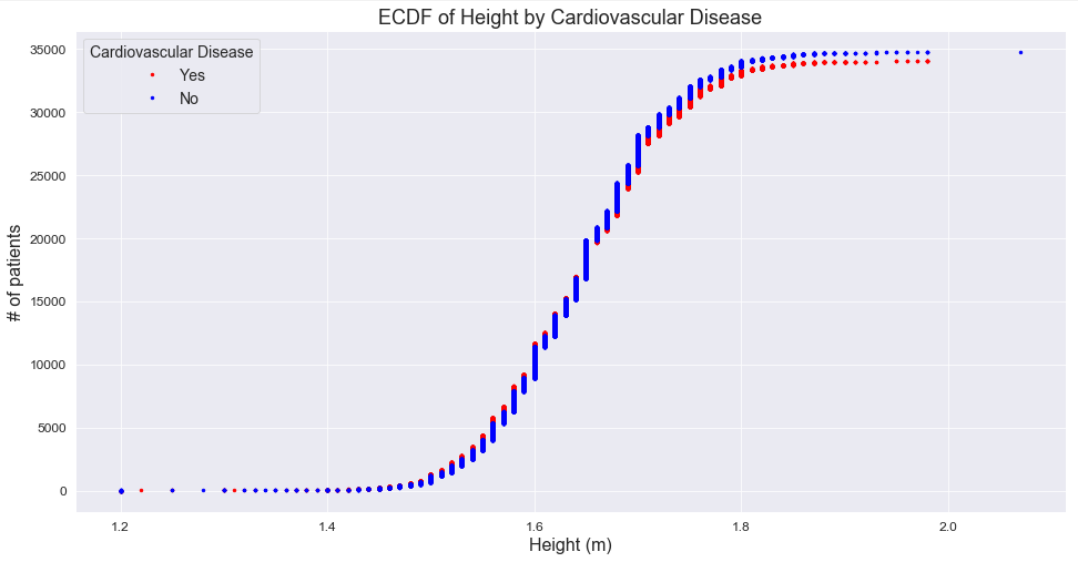


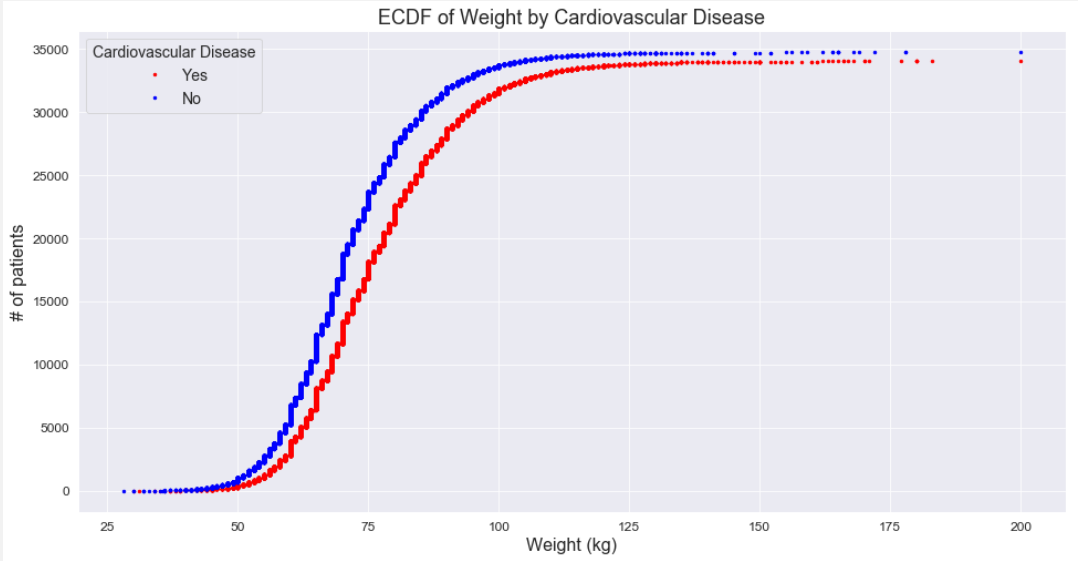
- Risk of cardiovascular disease increases with age (p-value = 0.0)
- Women develop cardiovascular disease later than men (p-value = 0.0)



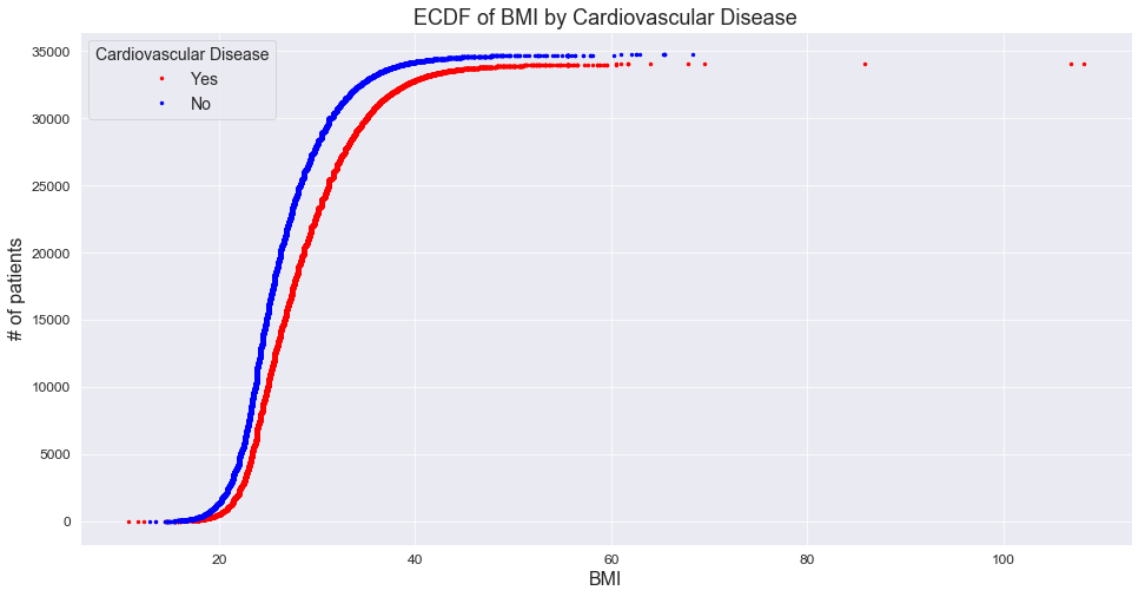
Effect of Height, Weight, and BMI on Cardiovascular Disease

Without Cardiovascular Disease: y = 0.0020x + 1.50
With Cardiovascular Disease: y = 0.0016x + 1.52





ECDFs for Height and Weight

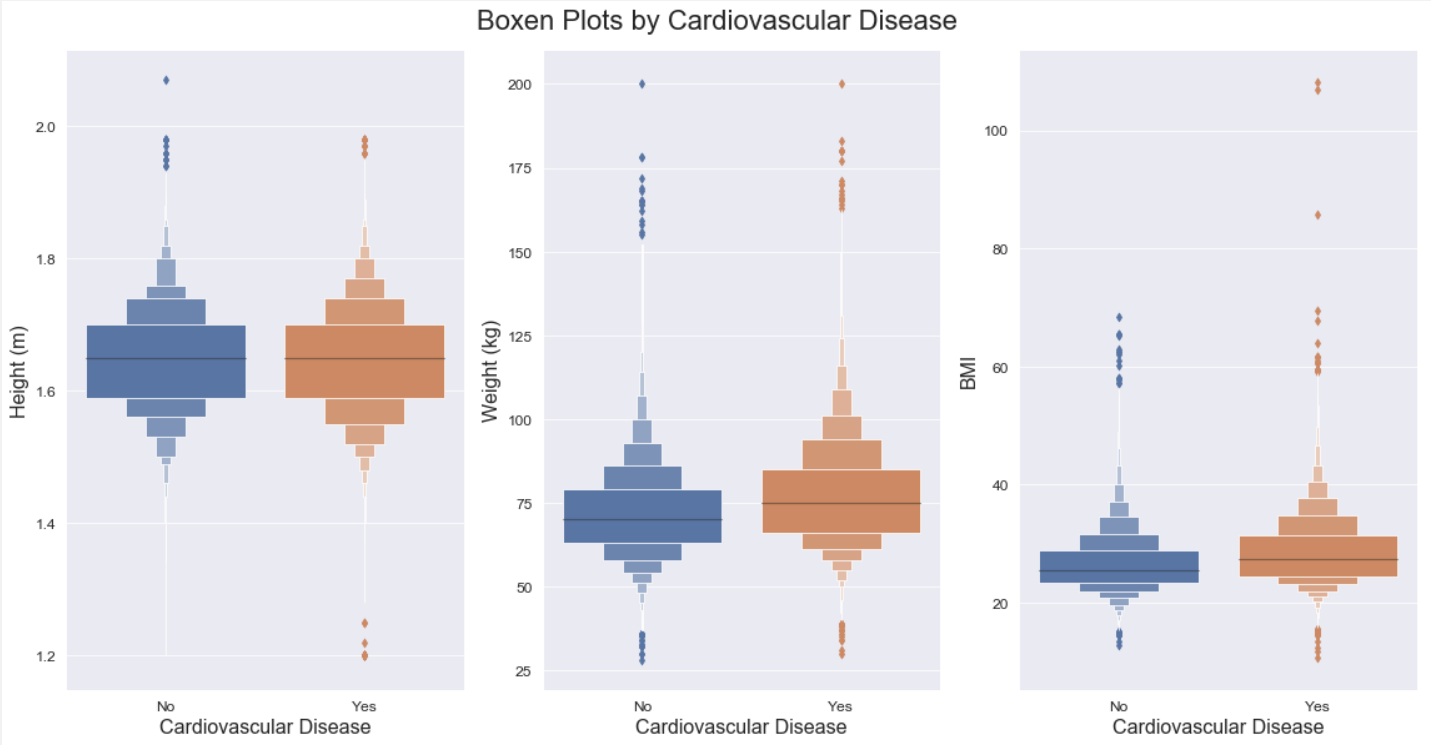


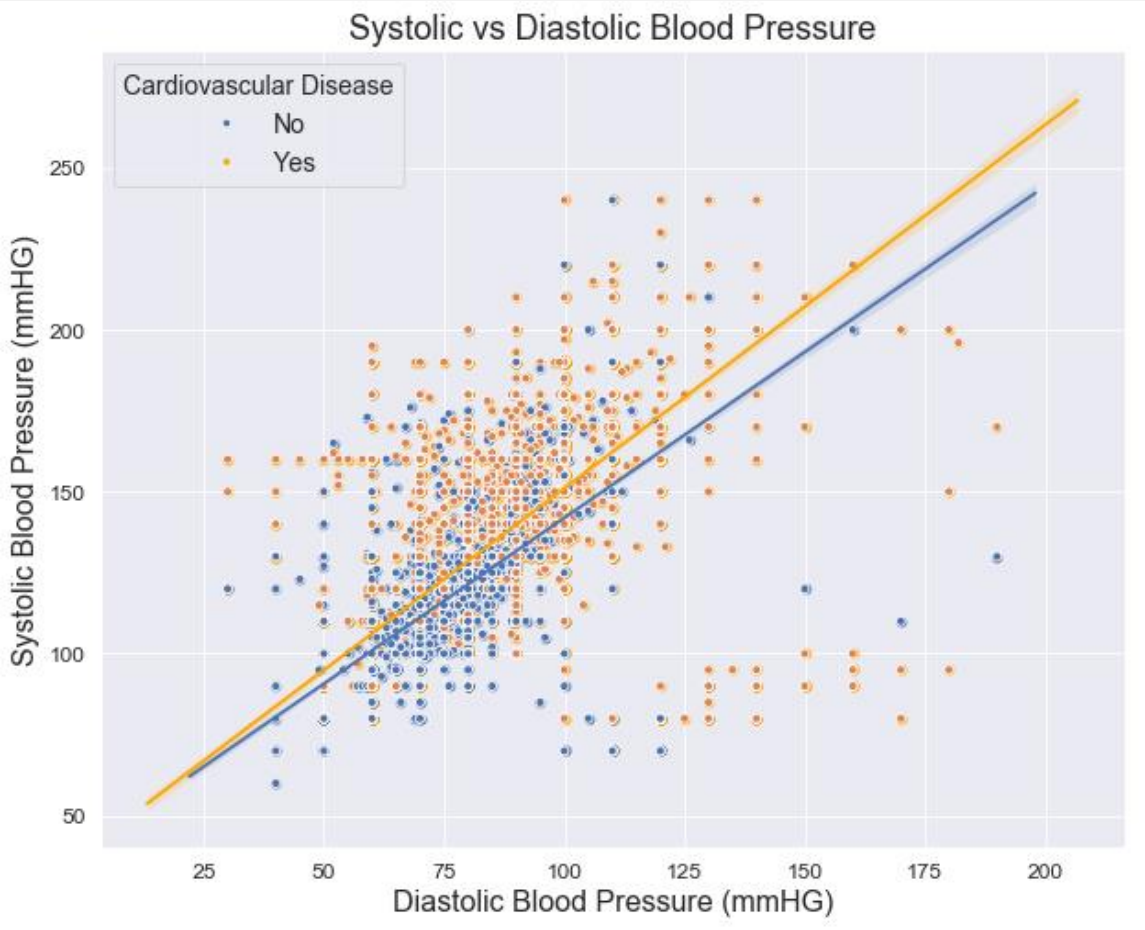
ECDF for BMI

- Both the weight and BMI ECDFs for those with cardiovascular disease are shifted to the right – i.e., towards higher weight/BMI values.
- Based on the ECDFs, height does not appear to have an obvious correlation with cardiovascular disease.

T-tests for height, weight, and BMI

- Average weight and BMI between those with and without CVD are statistically significant.
- The average height, while statistically significant, is not *practically* significant – the difference is only 2 cm.

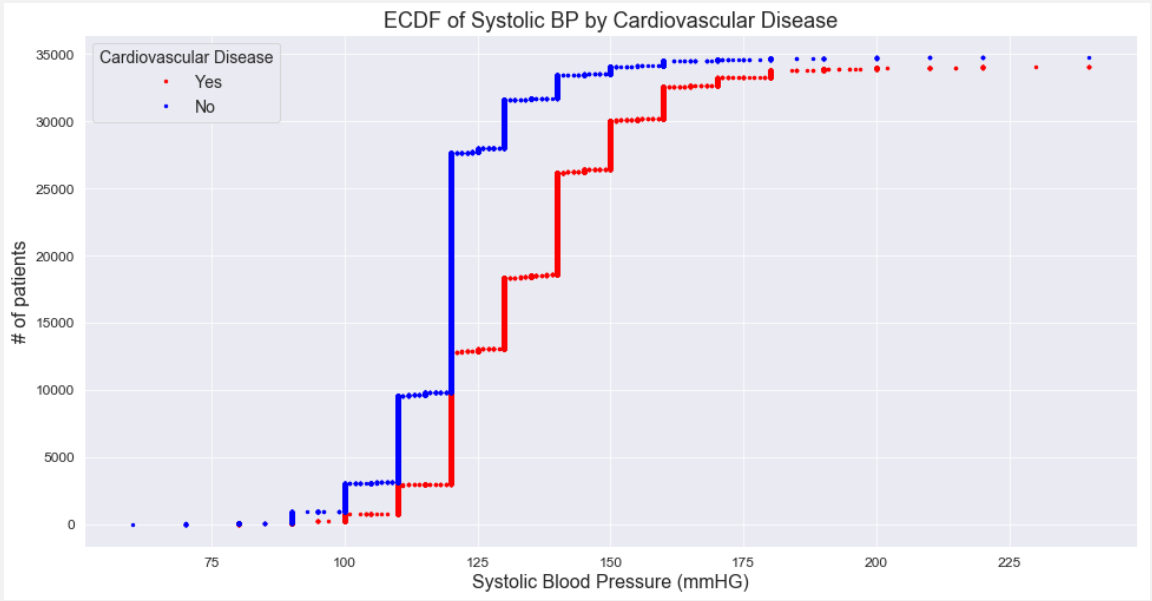


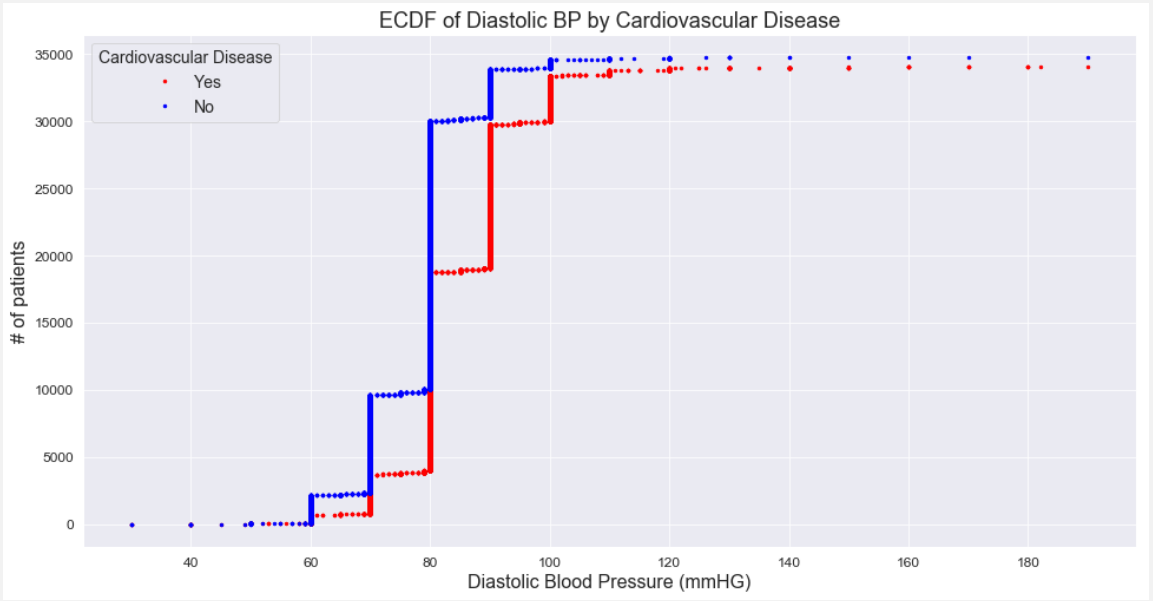


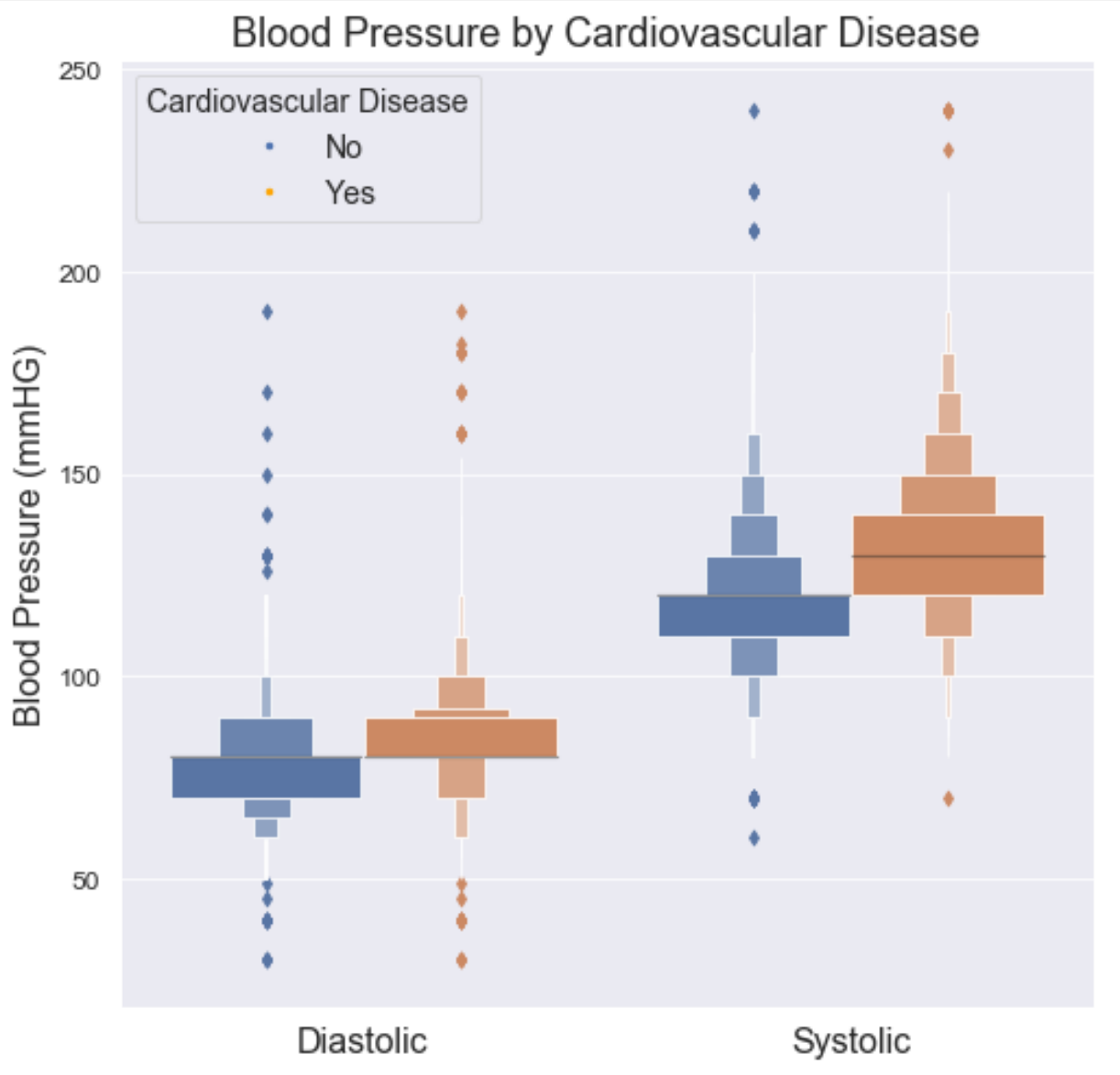
Systolic and Diastolic Blood Pressure

- Systolic blood pressure is the amount of pressure when your heart beats
- Diastolic blood pressure is the amount of pressure when your heart is resting between beats.
- The Pearson correlation coefficient between systolic and diastolic blood pressure is 0.697 (p-value = 0.0)
- Linear regression, no CVD: y = 1.0233x + 39.57
- Linear regression, with CVD: y = 1.1215x + 38.87

Patients with Heart Disease have Higher Blood Pressure







T-tests for Mean Difference in Systolic & Diastolic Blood Pressure

Systolic BP:

T-statistic = 123.14

p-value = 0.0

Diastolic Blood Pressure:

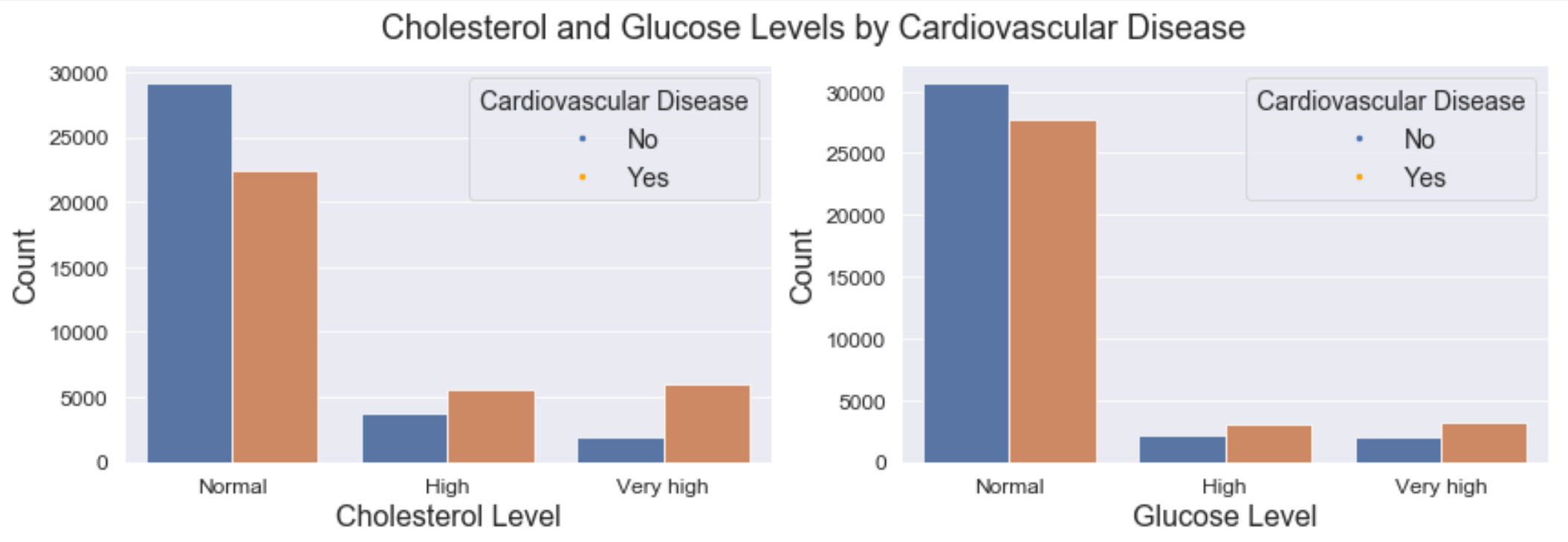
T-statistic = 93.11

p-value = 0.0

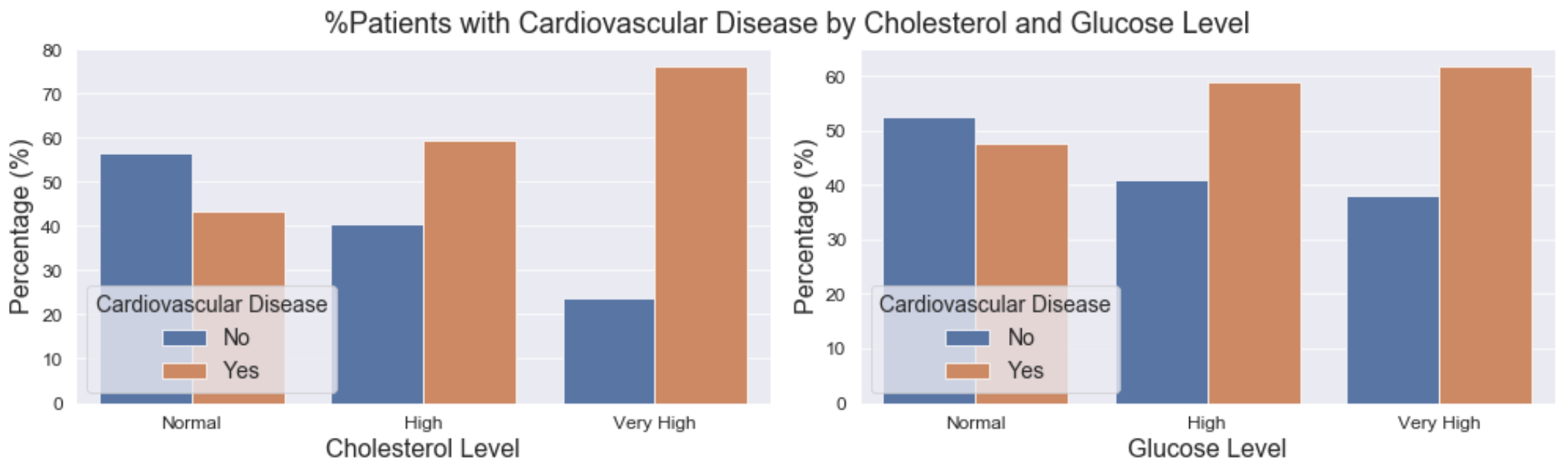
Comparison of Continuous Features

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

Distribution of Patients with Heat Disease by Cholesterol & Glucose Level



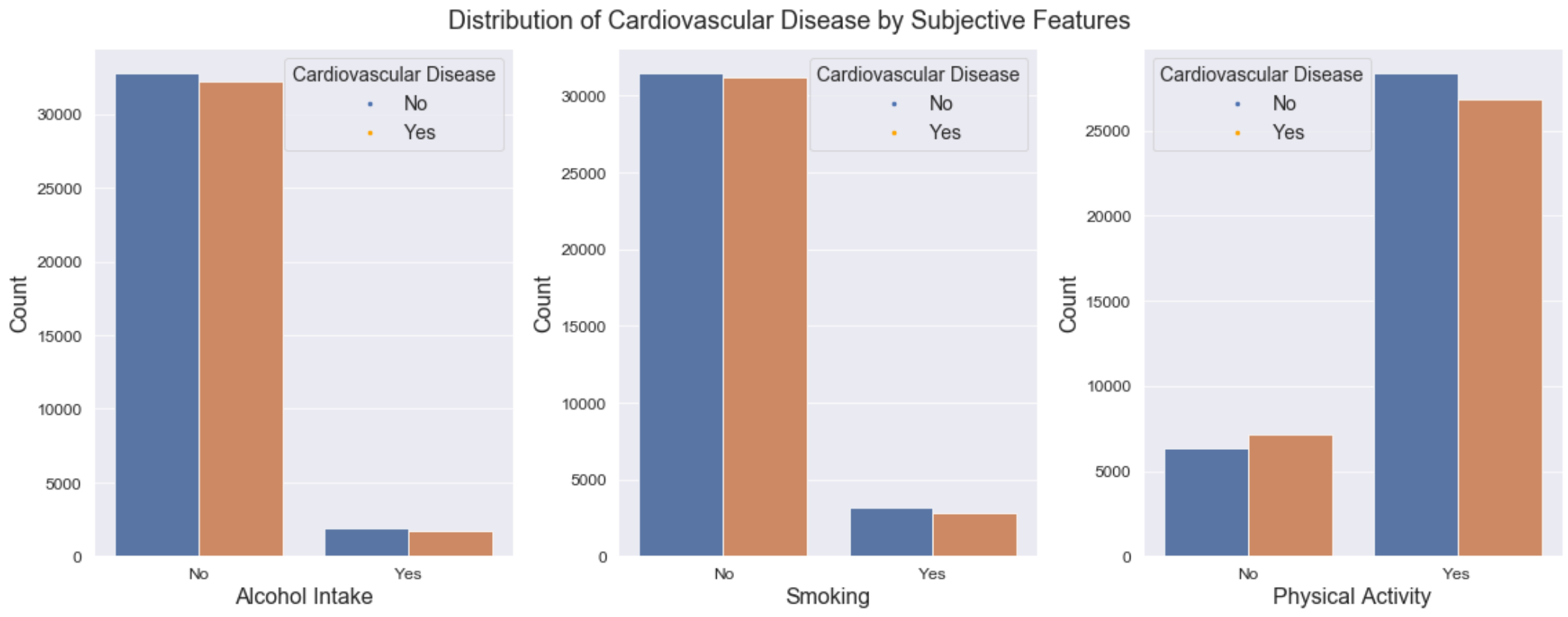
% of Patients with Heart Disease by Cholesterol & Glucose Level



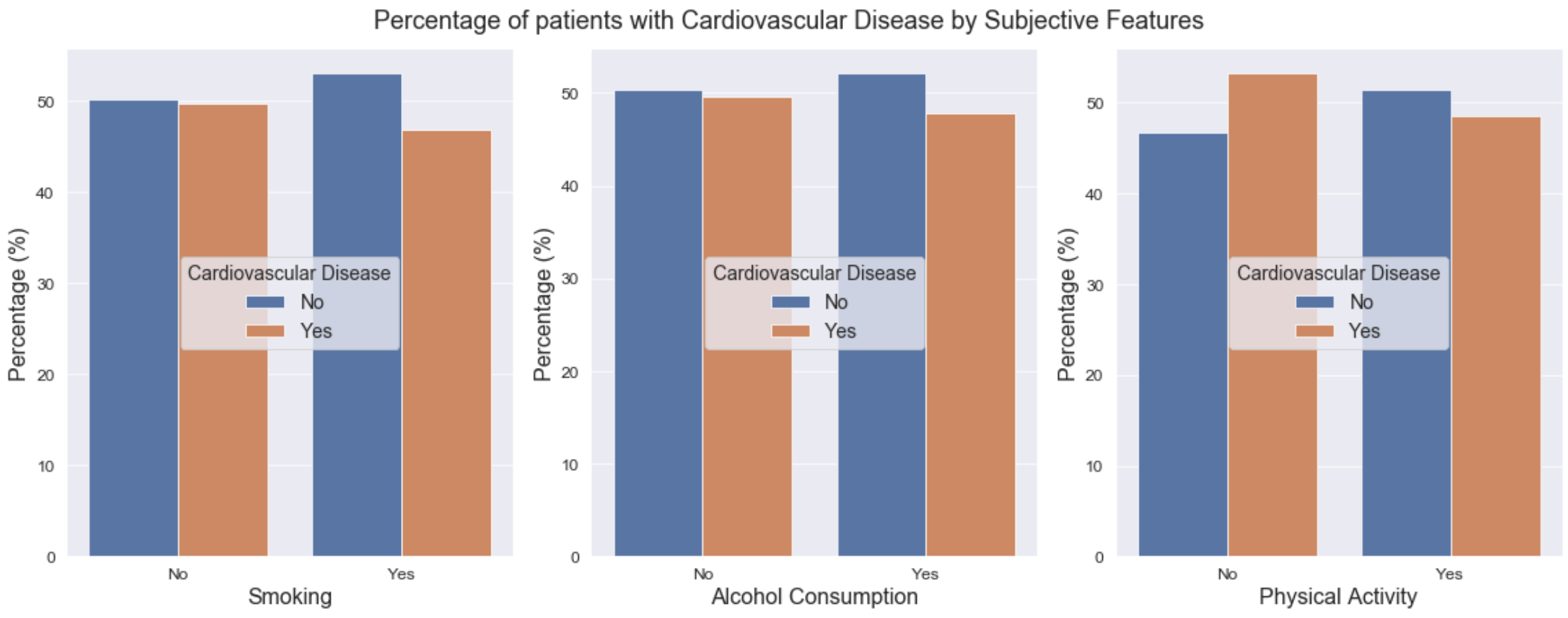
Chi-square Tests for Independence

	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		

Smoking, Alcohol Intake, & Physical Activity: Distributions



Smoking, Alcohol Intake, & Physical Activity: Percentages (%)



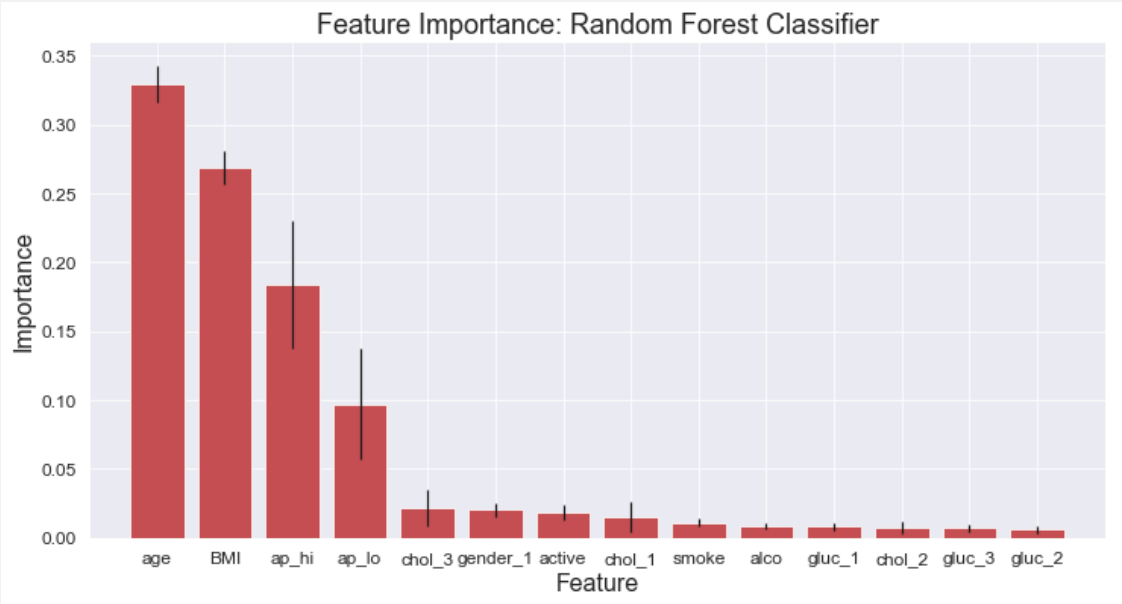
Chi-square Tests for Independence

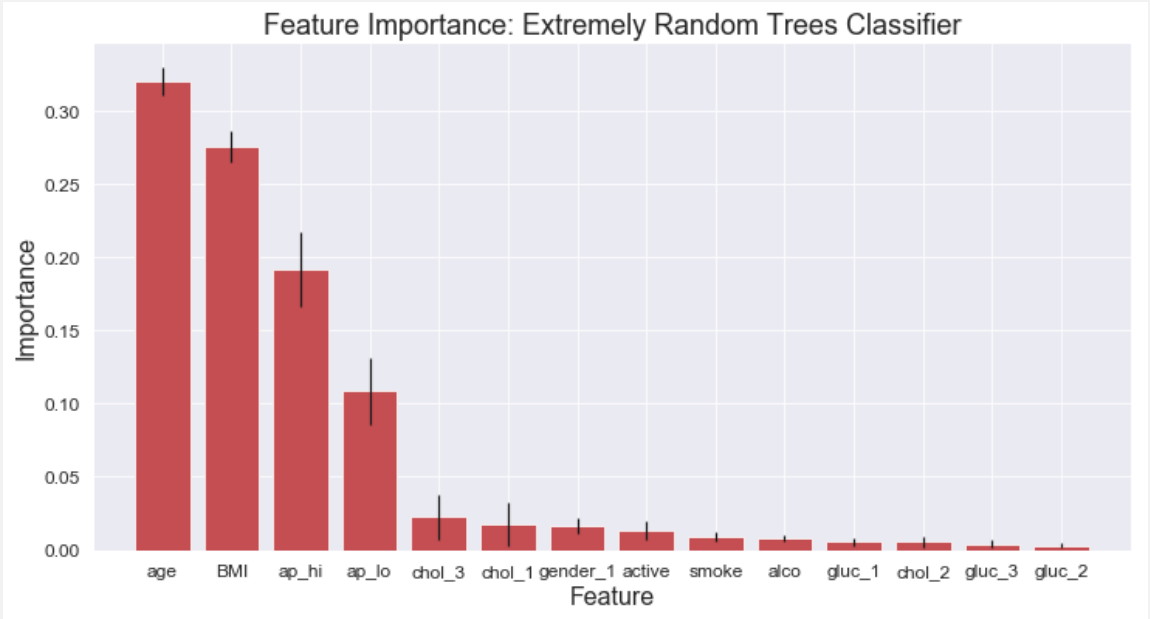
	Smoking		Alcohol Intake		Physical Activity	
	No	**Yes**	**No**	**Yes**	**No**	**Yes**
% with CVD	49.74%	46.86%	49.58%	47.80%	53.27%	48.56%
% without CVD	50.26%	53.14%	50.42%	52.20%	46.73%	51.44%
χ^2	18.25		4.36		96.22	
p-value	1.9×10^{-5}		0.037		1.0×10^{-22}	
χ^2 **correlation**	0.0163		0.008		0.0374	

EDA Summary

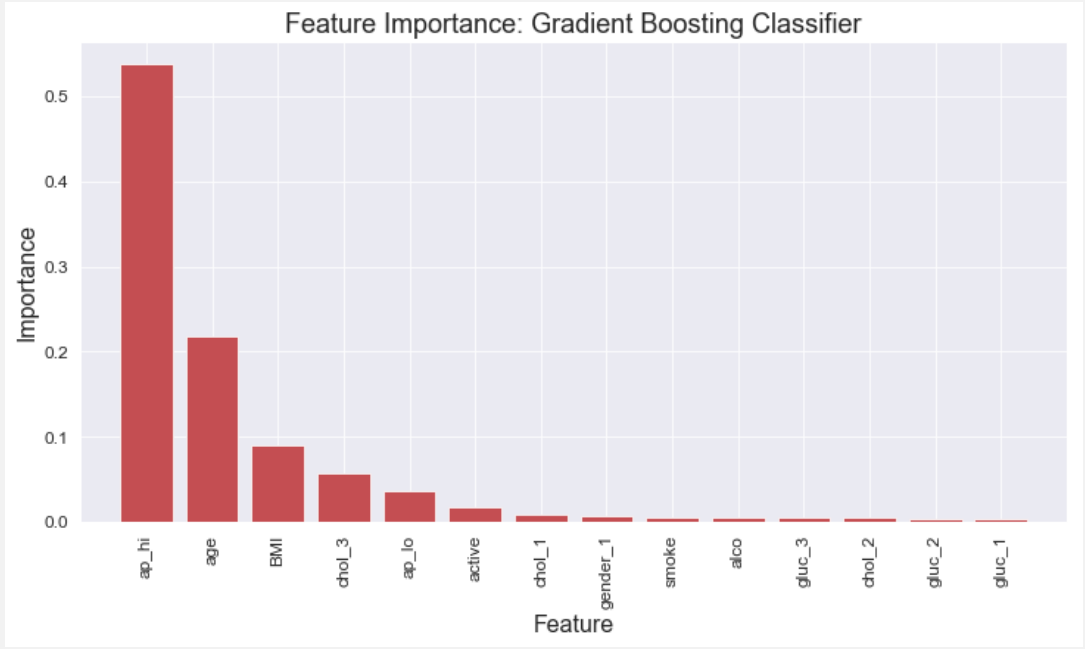
- Age, weight, BMI, systolic & diastolic blood pressure, and cholesterol level are all strongly correlated with CVD.
- Height is likely not relevant to CVD as the mean difference is very small, though statistically significant.
- Gender, while not significantly associate with CVD on its own, may have an effect in combination with other factors such as age.
- Data on smoking and alcohol intake are likely unreliable and should either be weighted less heavily or be removed from the dataset.

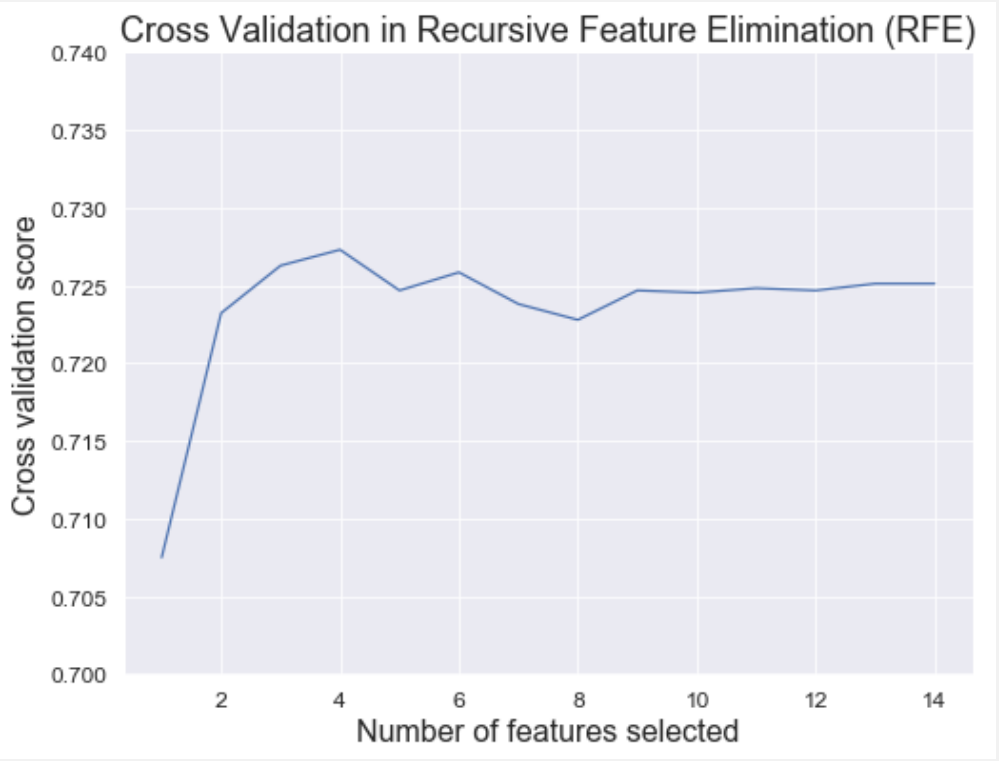
In-Depth Analysis & Machine Learning





Tree-Based Feature Importance





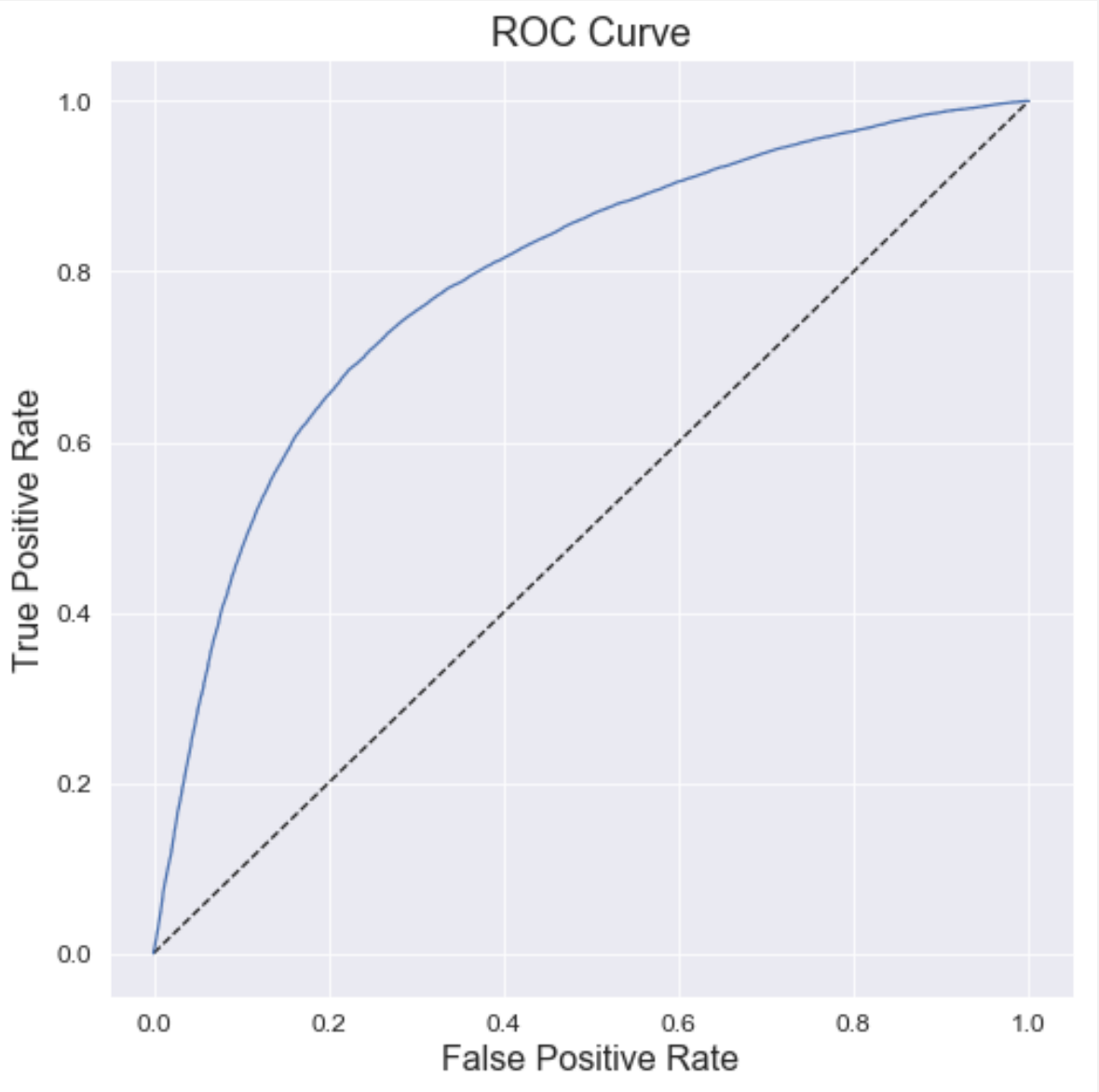
Feature Selection

Logistic Regression Classifier

Accuracy: 73.02%

AUC: 0.79

Precision	Recall	F1-score
0.71	0.78	0.75
0.75	0.67	0.71

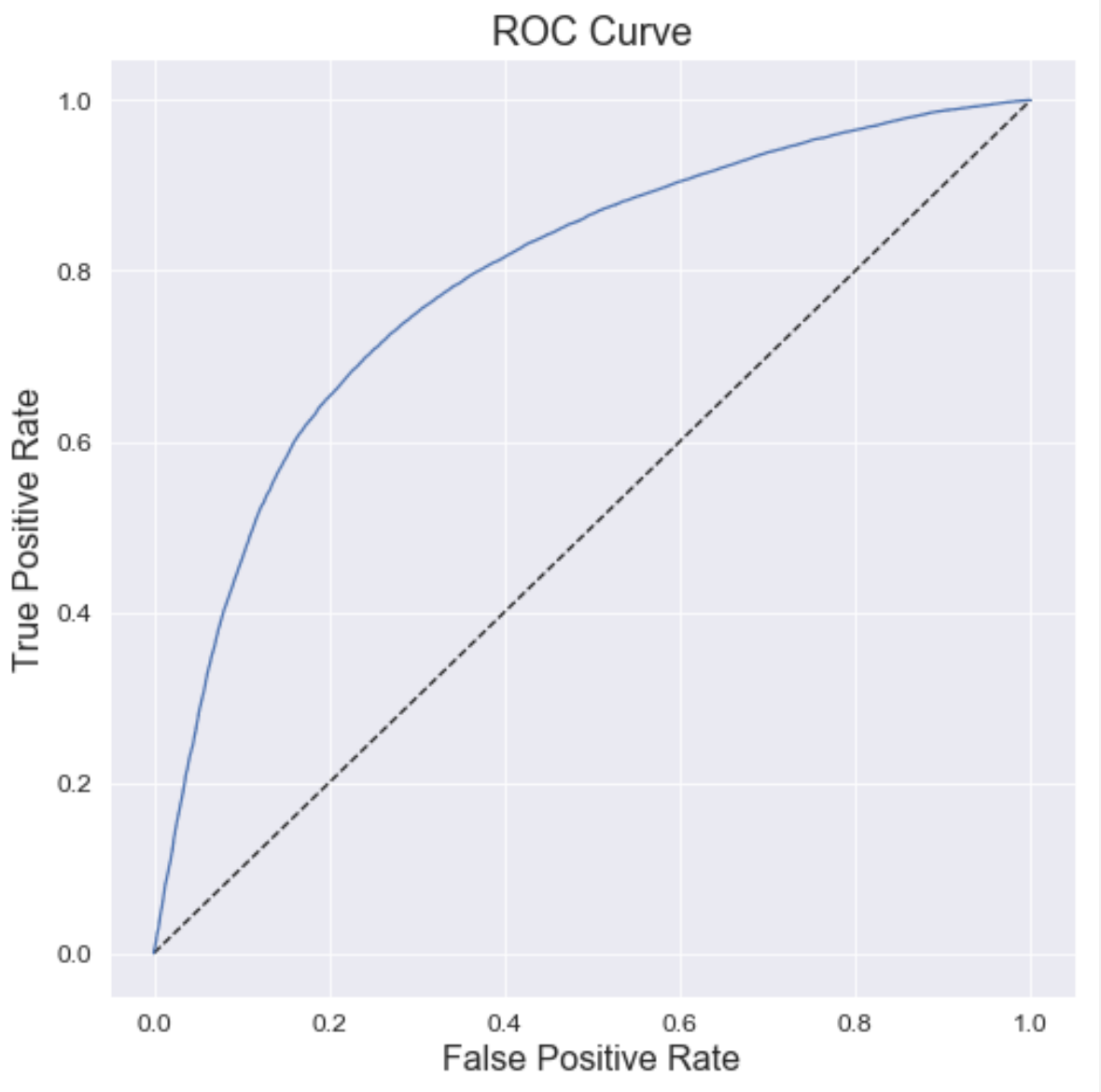


Support Vector Classifier

Accuracy: 72.48%

AUC: 0.79

Precision	Recall	F1-score
0.69	0.82	0.75
0.77	0.63	0.69

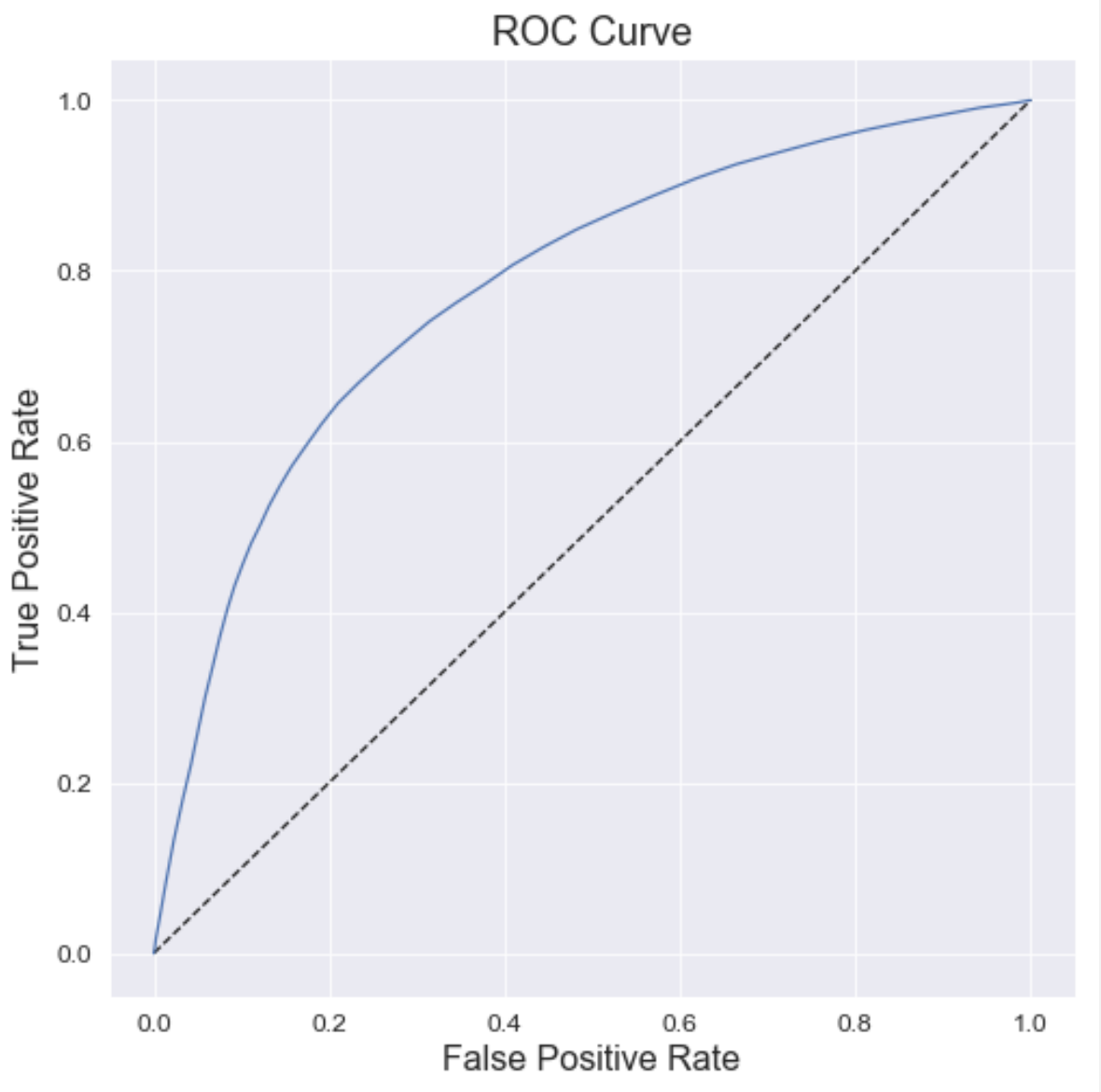


K-Nearest Neighbors Classifier

Accuracy: 71.80%

AUC: 0.78

Precision	Recall	F1-score
0.69	0.79	0.74
0.75	0.64	0.69

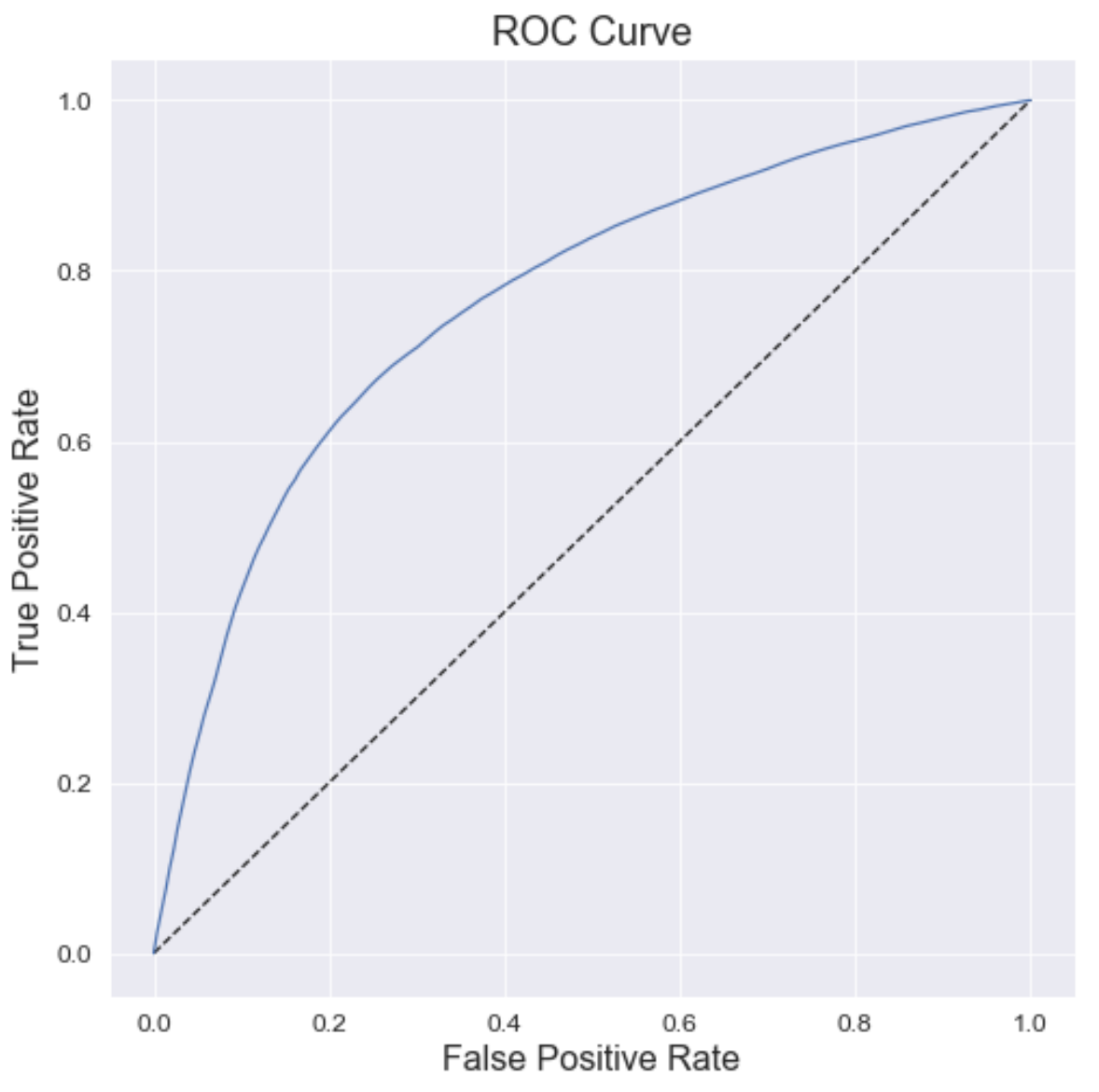


Random Forest Classifier

Accuracy: 70.75%

AUC: 0.77

Precision	Recall	F1-score
0.71	0.72	0.71
0.71	0.70	0.70



Machine Learning Summary

- Most important features are age, BMI, systolic and diastolic blood pressure, and cholesterol level
- Logistic regression classification performed the best, yielding an accuracy of 73.02% and good precision and recall scores while remaining fast to fit and predict.
- Neither different classifiers or different feature subsets had very large effects on the prediction accuracy

Recommendations

- Deploy model as a routine screening tool for healthcare providers to identify patients who should receive more thorough cardiovascular assessment
- Design public health campaigns to promote cardiovascular health and educate the public on the importance of monitoring their weight, blood pressure, and cholesterol levels, especially for older individuals

Future Directions

- Try more complex machine learning algorithms such as AdaBoost or Neural Networks
- Use PySpark to increase computational power to make more time-consuming analyses feasible.