Introduction

Problem Identification

Heart diseases are the second leading cause of deaths in Canada. In 2012, the disease claimed 48,000 lives. A lot of the causes attributed to the spread and prevalence of the disease is as a result of poor lifestyle choices, stress, high blood pressure and cholesterol levels. In order to be able to cater to this growing need for accurate monitoring and evaluation of patients struggling with chest pains or any heart-related symptoms, there needs to be an efficient system in place that can accurately identify patients that are most likely to suffer from heart attacks.

Client: Health Care Professionals

The health system in Canada is a universal one. This means that every Canadian is able to access health care needs without incurring a personal cost. As a result, this system suffers from an efficiency problem and mismanagement of resources. This could mean that patients with a high probability of suffering from a heart attack are being turned away because hospitals do not have the resources or capacity to be able to cope with the growing demand.

Objective

In order to maximize efficiency and promote the optimal allocation of resources, I want to build a model that will predict which patients are most likely to suffer from a heart attack based on several health-related attributes as well as lifestyle indicators that I am provided with.

Data

I have access to data on a patient's **age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression, ST segment slope, fluoroscopy colored vessels, and thalassemia,** which I will be using as my independent variables (descriptive variables) in the analysis. My dependent variable (target variable) will be whether one of the patients' major coronary vessels has a blockage that is greater or smaller than 50%.

Dataset: UCI Machine Learning Repository/Heart Disease Dataset

Variable	Variable Description
Age	[integer]
Sex	[1,0] - Male/Female
Chest Pain Type	1- Typical Angina 2- Atypical Angina 3- Non-anginal Pain 4- Asymptomatic
Resting Blood Pressure	[integer mm Hg]
Serum Cholesterol Level	[integer mg/dl]
Fasting Blood Sugar	True | False - Criteria: >120mg/dl
Resting Electrocardiographic Results	0: Normal | 1: Having ST-T Wave Abnormality | 2: Showing Left Ventricular Hypertrophy
Maximum Heart Rate Achieved	[integer]
Exercise Induced Angina	Yes | No [0,1]
ST Depression	Depression of the ST induced by exercise relative to rest.
ST Segment Slope	Slope of the peak exercise ST segment - 1: Upsloping | 2: Flat | 3: Downsloping
Fluoroscopy Colored Vessels	Number of blood vessels colored by Flourosopy [1-3]
Thalassemia	3=Normal | 6=Fixed Defect | 7=Reversible Defect
Diagnosis of Heart Disease	>50% | <50% narrowing of major vessel [0,1]

Data Wrangling, Storytelling and Inferential Statistics

Part 1: Data Wrangling

Steps to Creating the Dataset

We started by downloading 4 processed datasets from the UCI Machine Learning Repository consisting of data from Cleveland, Long Beach, Switzerland and Hungary. This data is cross sectional and from the year 1988. There are 15 variables with 916 observations. We compiled and combined these datasets into a single working dataset in order to be processed further. We proceeded to inspect the dataset and clean it to make sure it was ready for analysis.

Dealing with Missing Values (Nans)

The accompanying documentation to the data informs us that values are considered missing if they have '-9' entries as well as '?' entries. We did not delete the rows of this dataset that had missing values. Instead, we chose to replace these values with a 'NaN' value (blank value) and keep the remaining row information. We do this in order to avoid taking away variation in our data which would reduce the power and accuracy of our analysis. However, it is important to note that we will eventually drop these rows when conducting supervised machine learning, due to the fact that supervised learning does not take these values into account. It is unwise to interpolate, back-fill or front-fill the data as we risk reducing its quality by altering it.

Correlations

We proceed to inspect whether there might be any high correlations between the variables in our data. We want to avoid the latter due to the fact that this might bias or skew our analysis. We observe no correlations above 50% within our data and therefore, do not have to drop any variables as a result. A full correlation table is exhibited later in this report.

Dropping Variables with Majority Missing Values

In the dataset, we dropped four variables due to the fact that the majority of its data was missing and therefore adds little to no value to our analysis. We now have 11 explanatory variables in our dataset.

Compiling the Cleaned Dataset

Now that we have cleaned the dataset and made it ready for analysis, we saved it as a new file for future uses. Here is a look at what our dataset looks like.

	age	sex	cp	rbp	scl	fbs	rer	mhra	eia	std	new_dhd
0	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	1
1	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	1
2	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	0
3	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	0
4	56.0	1.0	2.0	120.0	236.0	0.0	0.0	178.0	0.0	0.8	0

Part 2: Data Storytelling

In this section, we study our data further by analysing graphical representations. This will give us a deeper insight into what our data can tell us and will help us make important interpretations in our analysis.

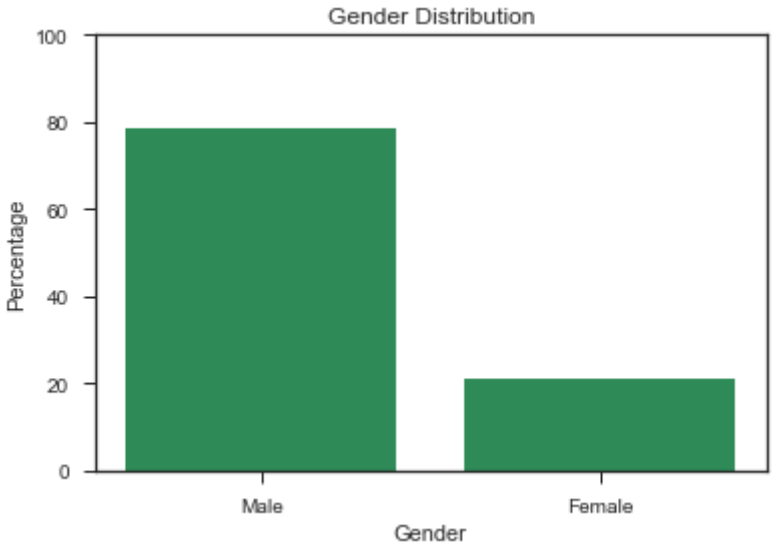
Patient Age Distribution

The average age of the patients in our dataset is between the range of 55-60 years old.

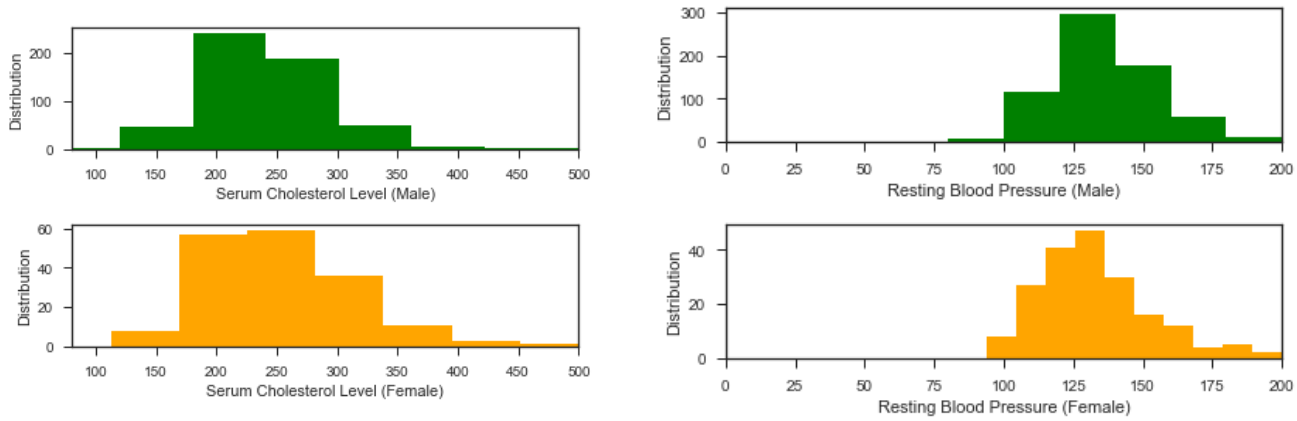
Analysis by Gender

Gender Distribution

The overwhelming percentage of male patients in this dataset is 78.8% as compared to the 21.2% of patients that are female.



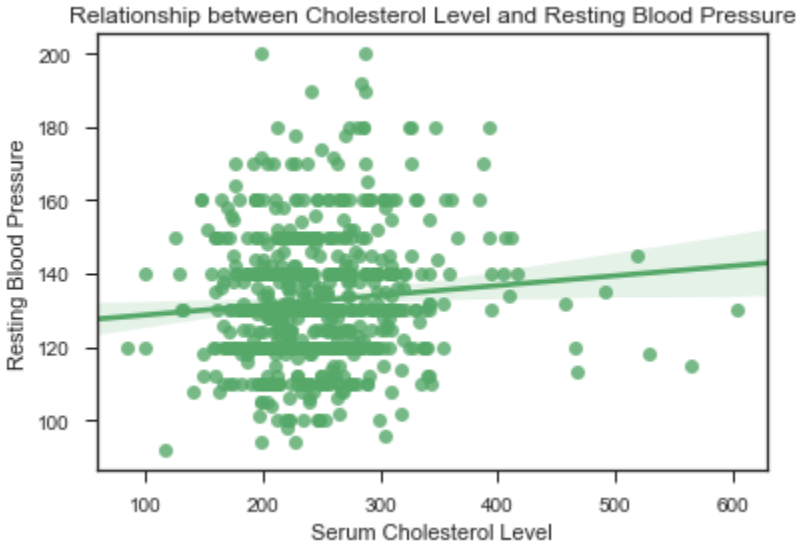
Resting Blood Pressure and Serum Cholesterol Levels by Gender



An analysis of the resting blood pressure of patients shows us that males and females exhibited similar resting blood pressures and serum cholesterol levels.

Analysis of Resting Blood Pressure and Serum Cholesterol Levels

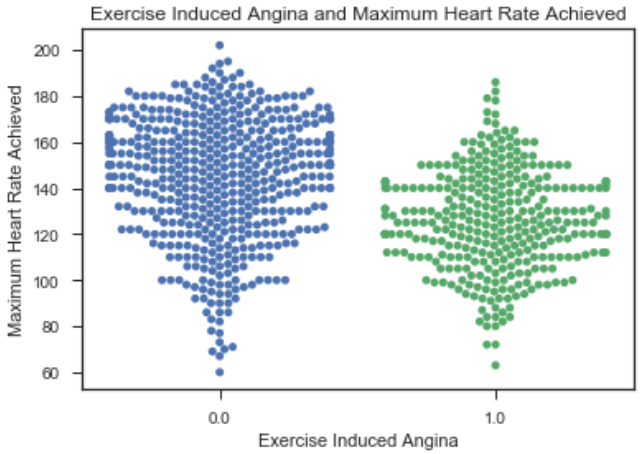
The mean resting blood pressure in our dataset is 132 mm Hg and the mean serum cholesterol level is 199.1 mg/dl.



We wanted to look further to see whether we could establish a relationship between a patient's resting blood pressure and serum cholesterol levels. We observe, based on the above plot, that there a slight positive relationship between both variables.

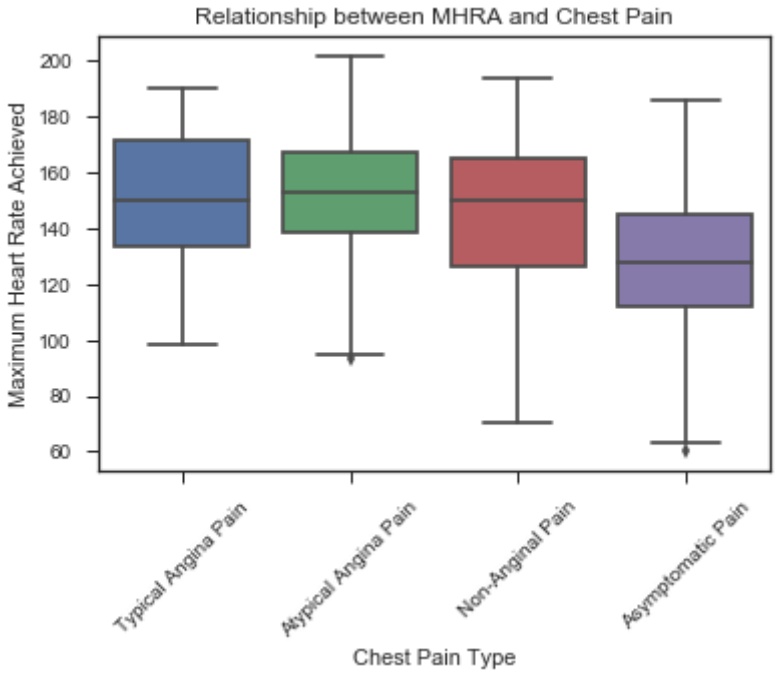
Analysis of Exercise Induced Angina and Maximum Heart Rate Achieved

Exercise Induced Angina (EIA) is pain caused in the chest region as a result of putting the patient under stress. We wanted to identify how the recorded Maximum Heart Rate Achieved (MHRA) during exercise changed with patients that suffered with EIA.



By studying the swarm plot above, we note that the ‘thickness’ of the swarm varies between whether a patient suffered with EIA or not. This means that, patients that suffered with EIA were less than those that did not. Also, observe the height of the swarm plot. The plot studying patients that did not suffer with EIA is ‘higher and thicker’ at the top, which tells us that the same patients were more likely to reach higher heart rates during exercise and were a lot more than the patients that did suffer with EIA. This makes sense, since the latter would not suffer chest pains during exercise and were able to put their bodies under more stress.

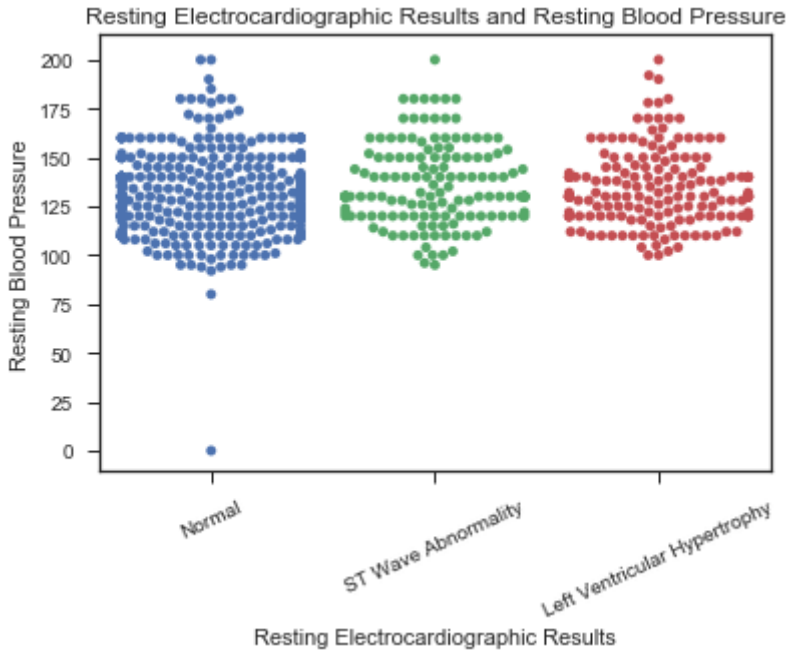
Analysis of Chest Pain and MHRA



We observe above that patients that suffered from Asymptomatic chest pain were less likely to reach higher heart rates when exercising. The remaining chest pain types seems to exhibit similar average heart rates.

Analysis of Electrocardiographic Results and Resting Blood Pressure

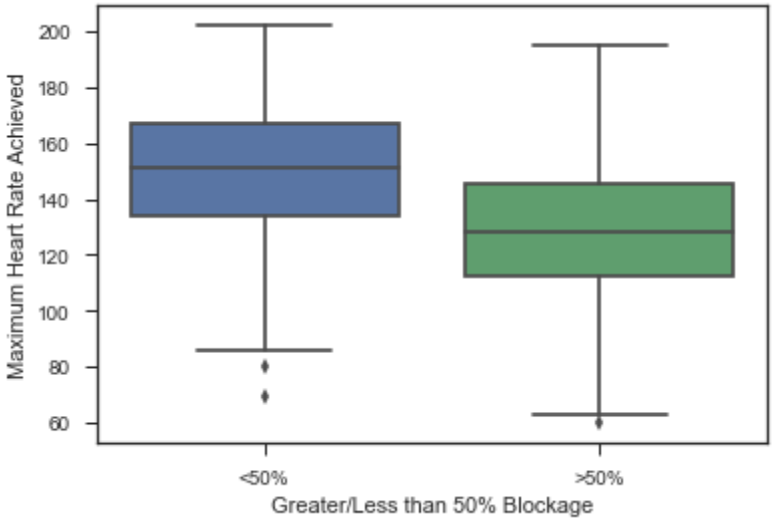
Electrocardiography is a method used to measure electrical impulses from the heart. We analyse these results and the respective patient's resting blood pressure.



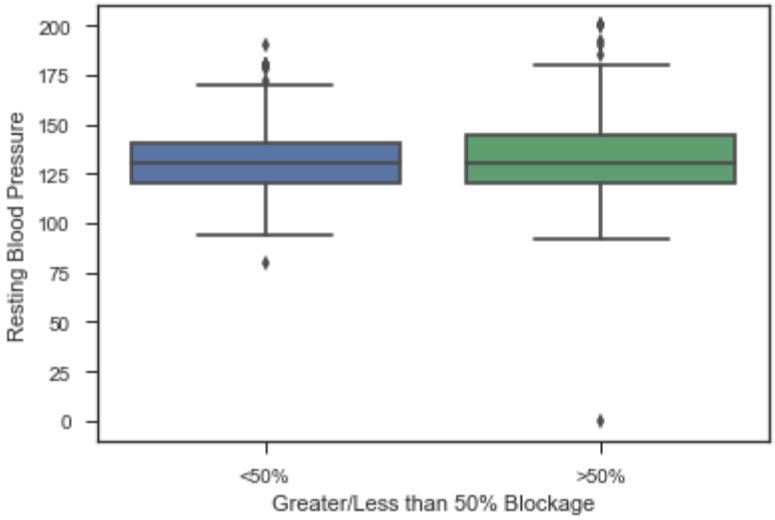
By studying the swarm plot above, we observe that most patients had normal electrocardiographic results. Also, we see that the height and the distribution of the swarm plot for each criterion is fairly similar for both ST wave abnormality and left ventricular hypertrophy.

Analysis of the target variable (DHD) – the presence of >50% or <50% blockage in major coronary vessels

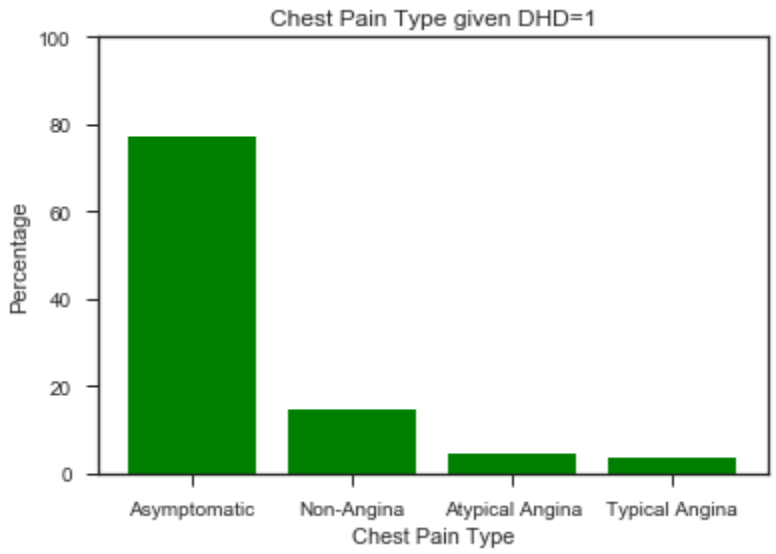
In our current dataset, 55% of patients have coronary vessels with more than 50% blockage.



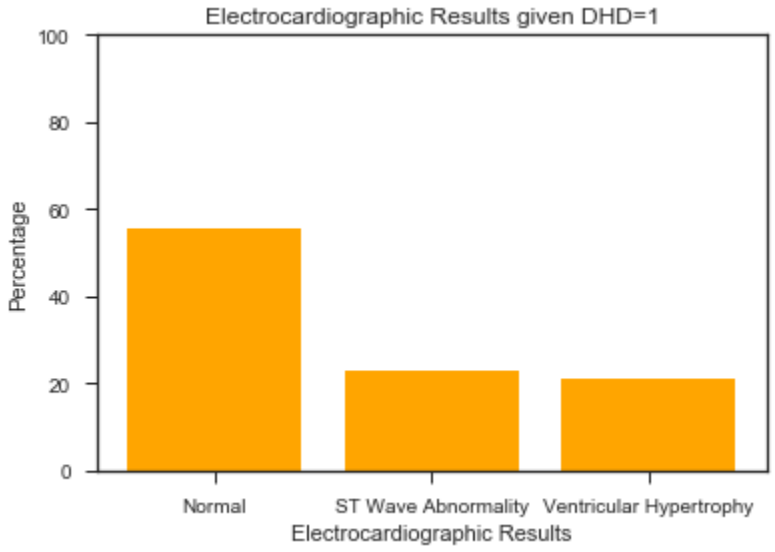
In our first plot, we see that patients with greater than 50% blockage were less likely to reach a MHRA when the patient's body was put under stress.



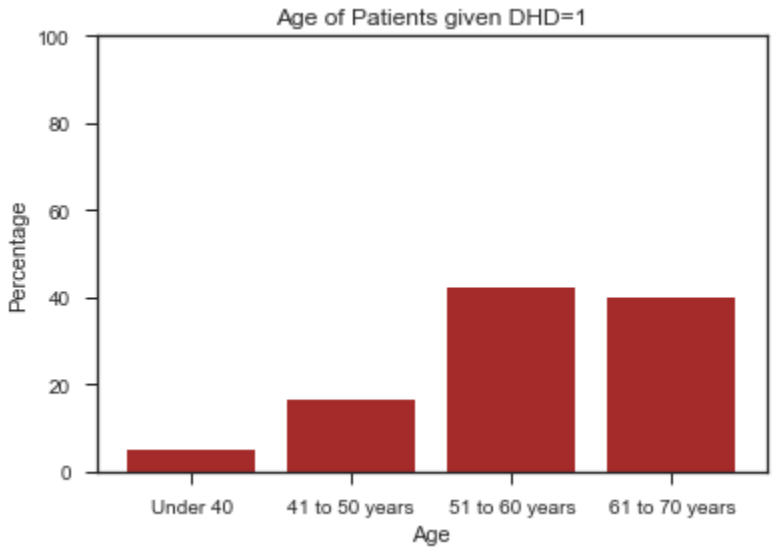
In this plot, we observe that patients with greater or less than 50% blockage were likely to have the same average blood pressure of approximately 130 to 135 mm Hg.



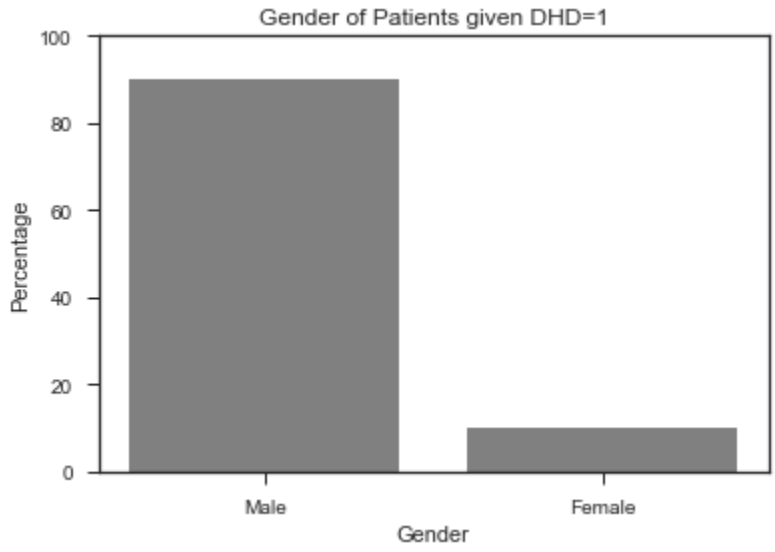
We separated the dataset such that we focused only on patients that suffer from a greater than 50% blocked coronary vessel (DHD=1). We then studied their respective chest pain types. According to the plot above, we observe that most patients suffered from an *asymptomatic* chest pains.



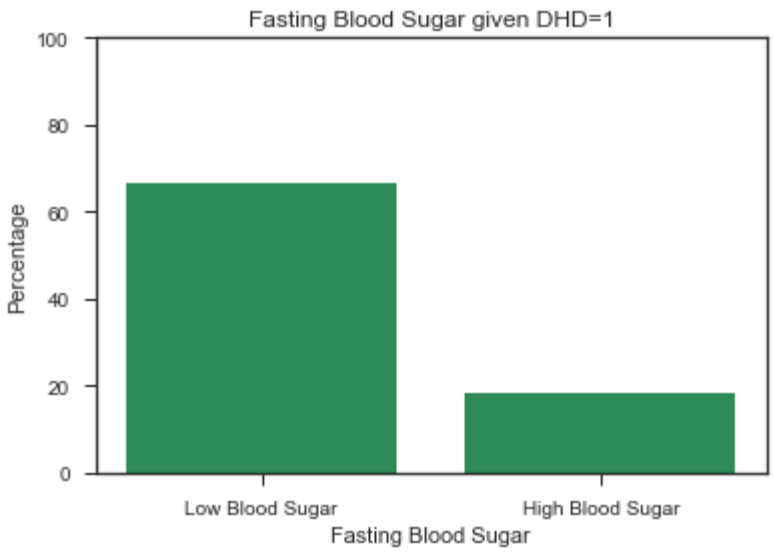
Here we observe that the overwhelming majority of patients have a normal electrocardiographic reading. This is quite alarming given that this is a an important indicator to the physician of any subtle abnormalities in the patient's heart function.



This plot provides insight into the age of the patients that suffer from a blocked coronary vessel. As shown in the plot, the majority of patients are between the ages of 51-70 years old.



The overwhelming majority of patients are male occupying more than 90% of the patient list. This is what we predicted, given that the original distribution of the data also showed a majority of male patients.



We observe that the patients are more likely to have a low blood sugar reading.

Now that we have had better graphical insight into the data and what it tells us about the relationship between variables as well as the characteristics of the patients that suffer from a a greater than 50% blocked vessel, we now start testing the data and challenge some of our hypothesis.

Part 3: Inferential Statistics

Grouping our data by our target variable and analysing means

We start by grouping the dataset by our target variable and study the means of our explanatory variables. It is worth noting that the majority of our explanatory variables are binary variables, so the interpretation of the mean is little less straightforward. However, for binary variables with '0' and '1' entries, the higher the mean, the more likely there were more '1' entries.

new_dhd	age	sex	cp	rbp	scl	fbs	rer	mhra	eia	std
0	50.542787	0.647922	2.767726	129.848329	227.735897	0.108861	0.545232	148.737789	0.141388	0.414433
1	55.936884	0.901381	3.648915	134.049145	176.667339	0.217593	0.651485	128.298729	0.595339	1.260086

We can observe that patients with greater than 50% blocked vessels were more likely to have higher resting blood pressures, however, the same patients averaged lower serum cholesterol levels. Also, the latter group of patients had a lower maximum heart rate reading, and were more likely to have exercise induced angina. Also, we note that these patients were older, with an average of 55 years in contrast to 50 years.

Two Sample t Tests

I separated the dataset into patients with above average and below average blood pressures. I wanted to study whether patients with different blood pressure levels were less likely to differ in their likelihood of whether they would have a greater than 50% or less than 50% blocked vessel. In other words, I hypothesized that both samples would have the same mean for the target variable.

test statistic	p-value
3.43	0.0006

This information provided tells us that we can confidently claim that the mean of the two samples are different and therefore, we can state that resting blood pressure is a significant factor to whether the patient has a major or minor vessel blockage.

Moreover, we proceeded to separate the dataset into two samples of patients based on their electrocardiographic results (normal and abnormal). Using the same test above we were provided with the following statistics.

test statistic	p-value
-2.98	0.002

In this case, we can also state, with confidence, that the means of the two samples with respect to whether the patient had major or minor vessel blockages were the same. Therefore, electrocardiographic were also significant in this case.

We proceed to test whether gender is related to a presence of a blockage in a coronary heart vessel.

test statistic	**p-value**
9.8	0

As in our previous exercises, we conclude that a patient's gender plays a role in whether they will have a major or minor blocked vessel.

The purpose of conducting this series of tests is for us to be able to state that the variables in our study add value to the analysis and will help us make accurate predictions and interpretations.