Report of patients with heart diseases based on SAS Visual

Author: Jiaqi Ling

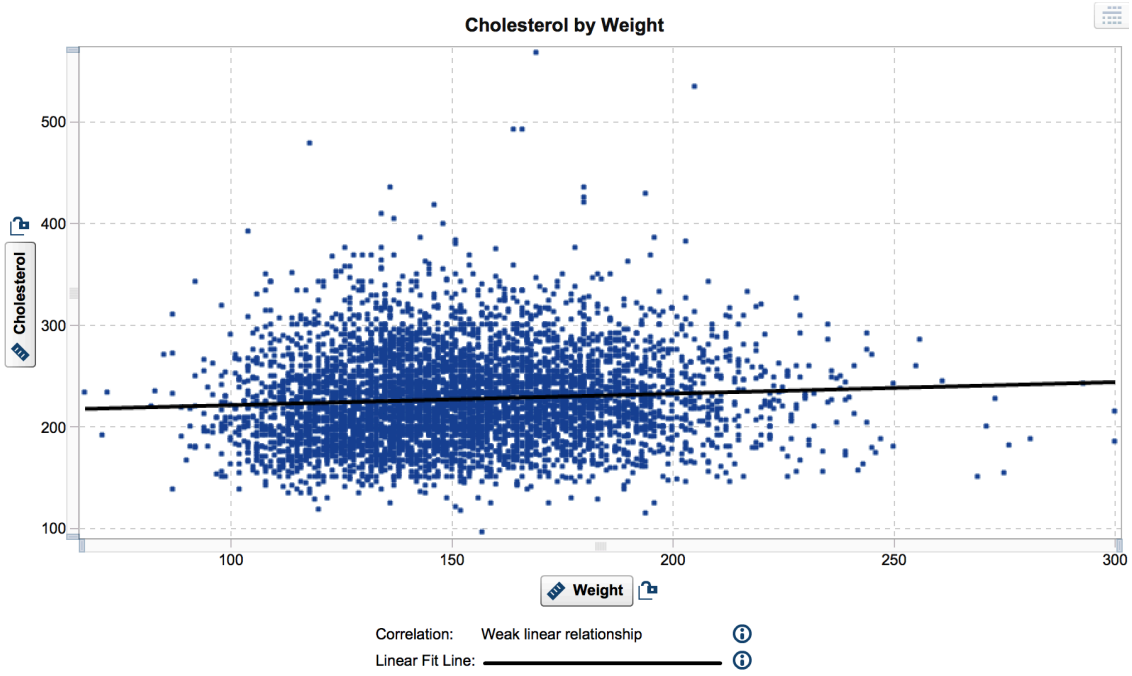
Date: Oct 21, 2018

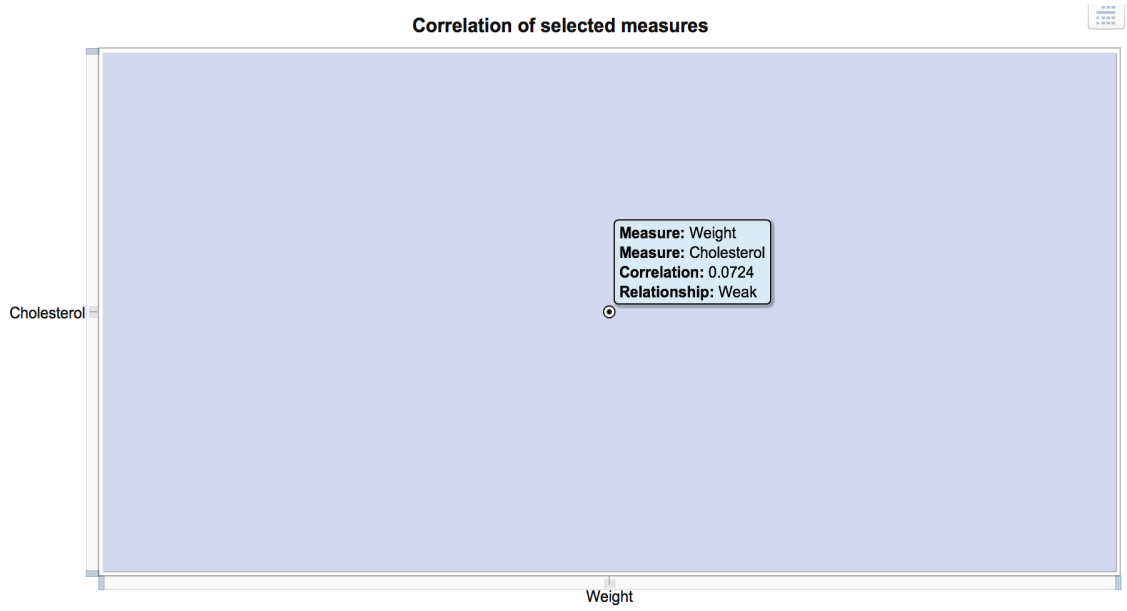
1. Background

This report aims to assist the scientists in analyzing the data about the patients that received heart disease treatment in one of a large hospital chain and provide insightful information to their medical research team. By investigating and analyzing the data, a list of hypotheses is tested, and potential underlying causes of coronary heart disease are specifically explored.

2. Hypothesis testing

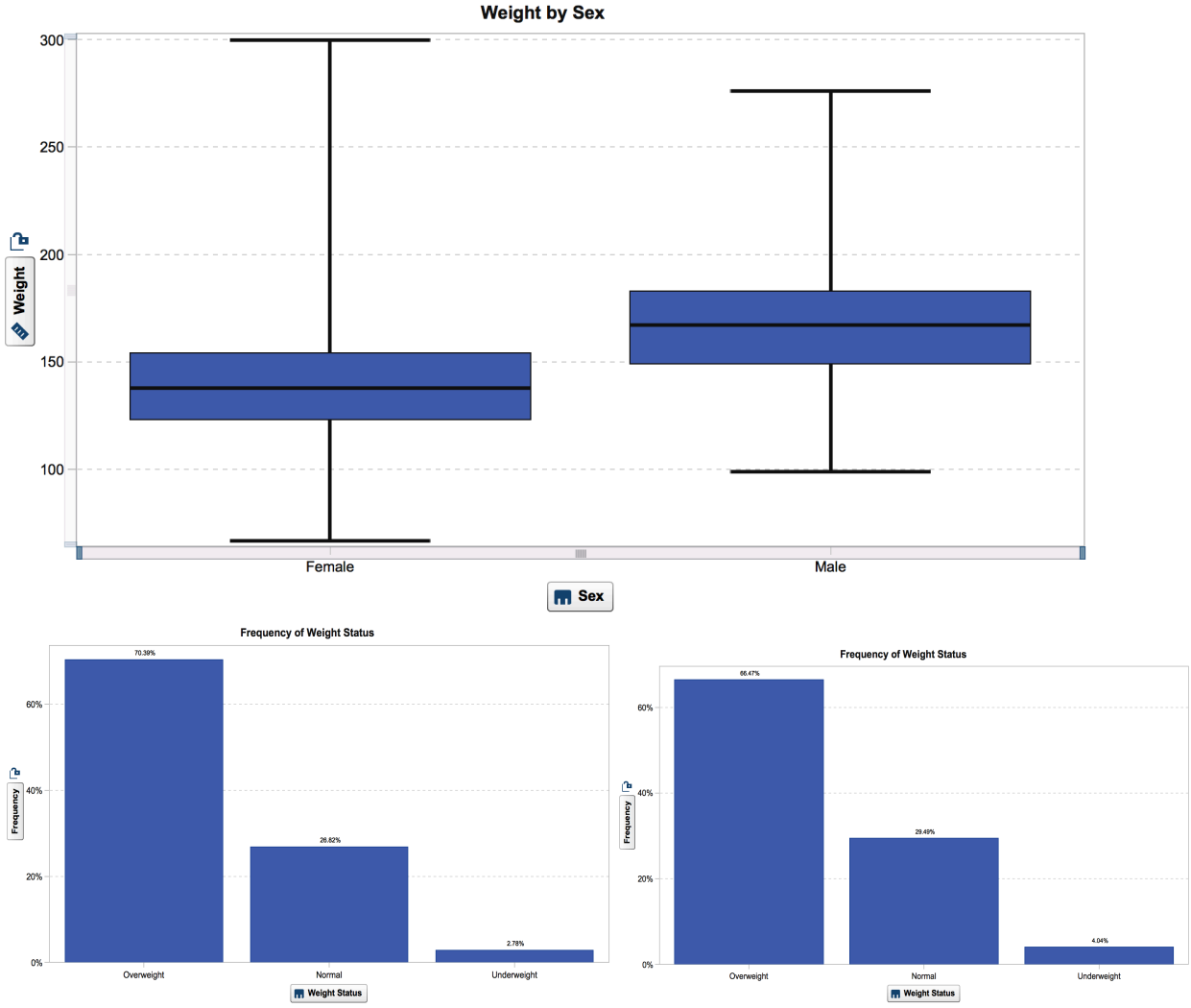
2.1 H1: The weight and cholesterol levels are correlated





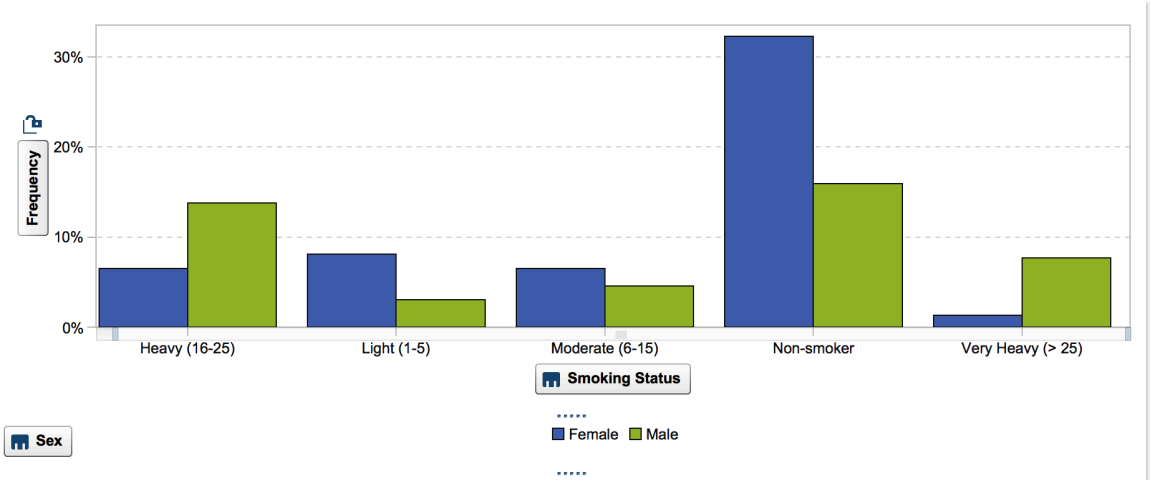
H1 is poorly supported by the dataset. No obvious pattern is found in the scatter plot of cholesterol and weight of patients, and the linear fit line is almost parallel to the x-axis, which means the linear relationship between these two factors are rather weak. By calculating the correlation, 0.07 also shows weak linear relationship between these two factors.

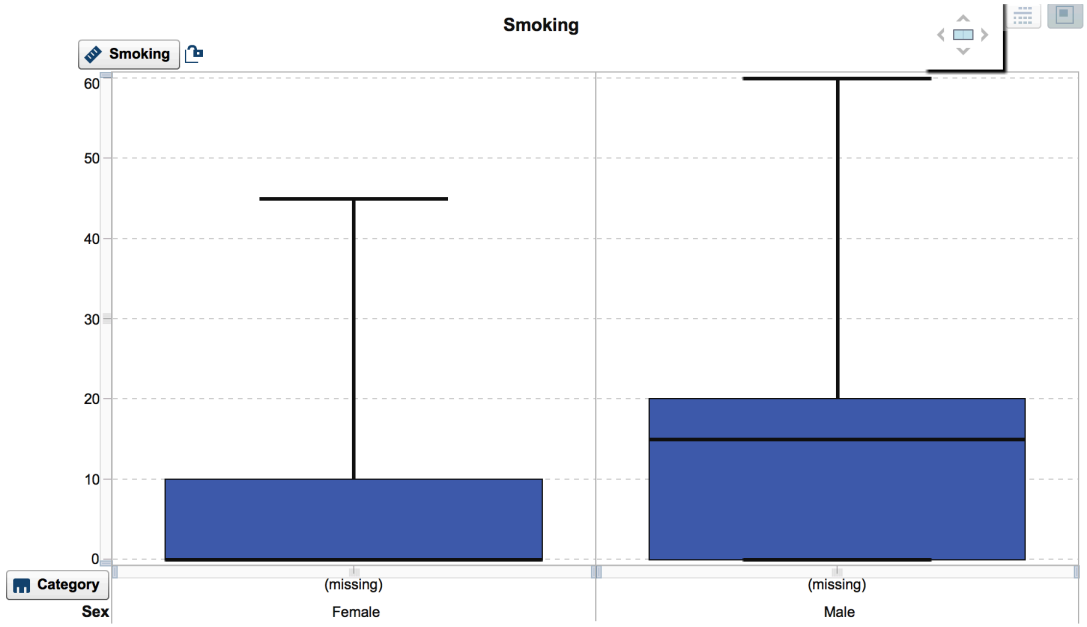
2.2 H2: Men are usually more obese than women



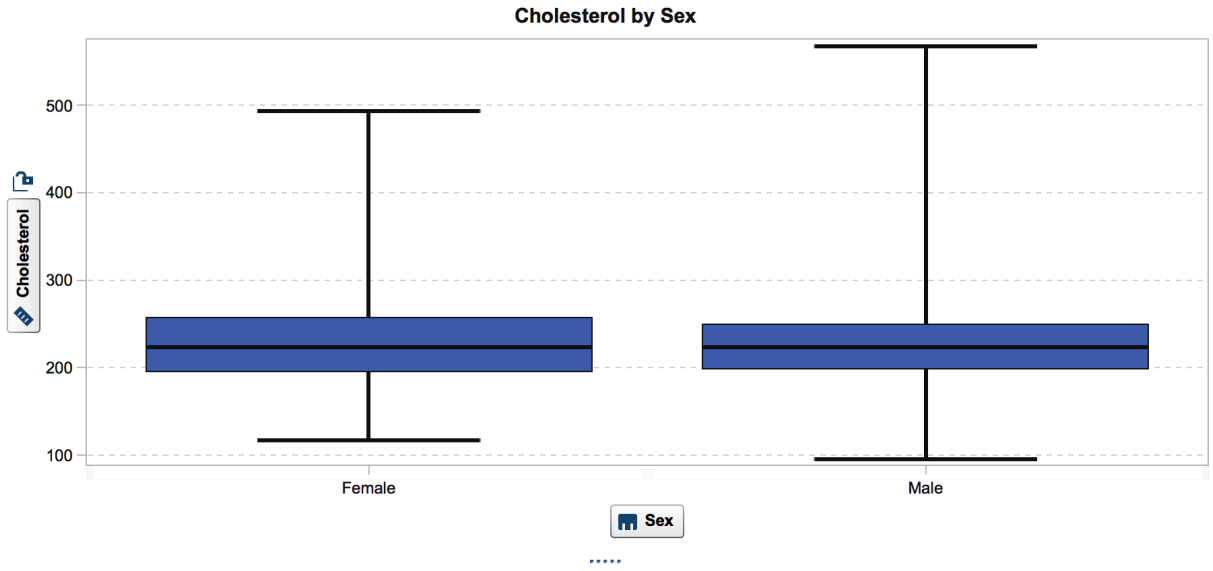
H2 is proved right. Simply from the box plot of weight by sex, the average weight of males is larger that of female. What's more, in the following bar chart, 66.38% of female is overweight, which is less than 70.33% of male. We could come to the conclusion that not only the average weight of males is greater than female due to different physical structures, but men are also more obese than women.

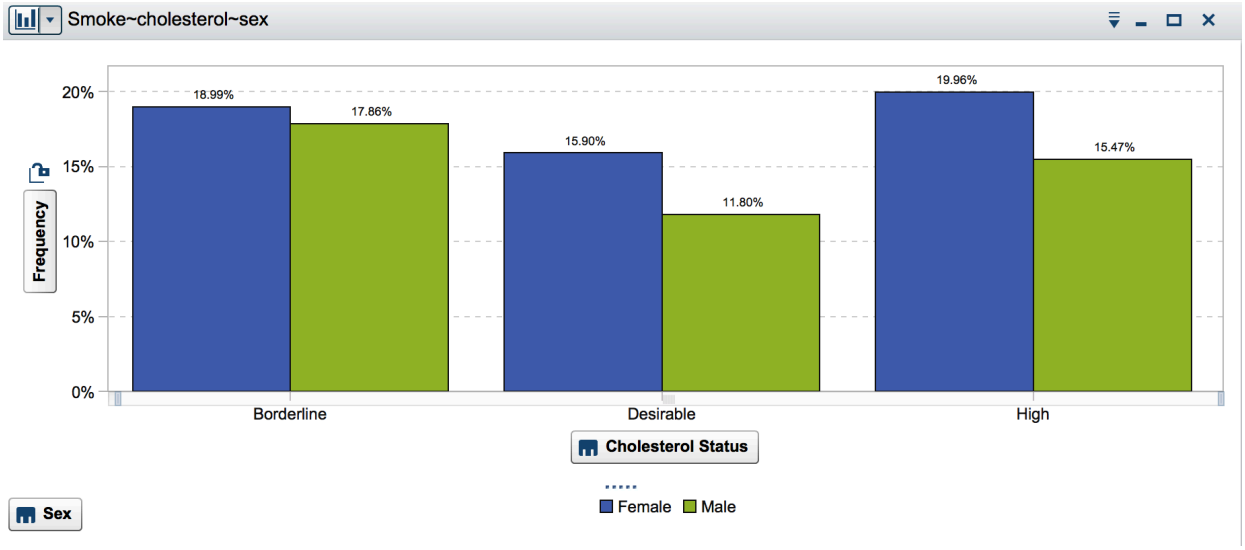
2.3 H3: Women usually smoke less than men, but their cholesterol level is higher





For non-smokers, female (32.28%) outnumbers male (15.96%) in percentage frequency, while the facts are just the opposite for heavy and very heavy smokers. It's not hard to say that women are more often light or non-smokers, while men are more often heavy smokers. Besides, by comparing the average of smoking of two genders, the average smoking of male is much higher than female.

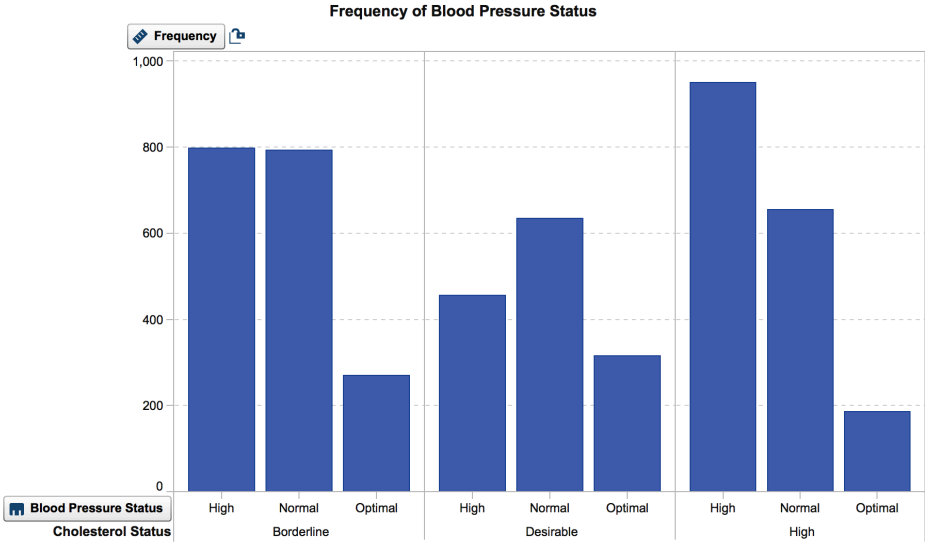


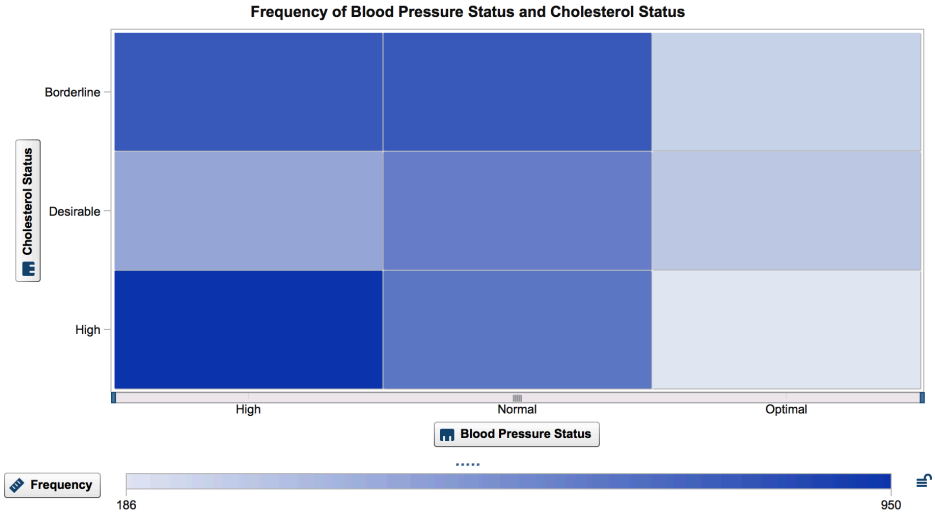


The average cholesterol of female is close to male but in the high cholesterol status group, women appears to be the majority.

Overall, the hypothesis that women usually smoke less than men, but their cholesterol level is higher makes sense.

2.4 H4: The blood pressure is higher for people with higher cholesterol levels





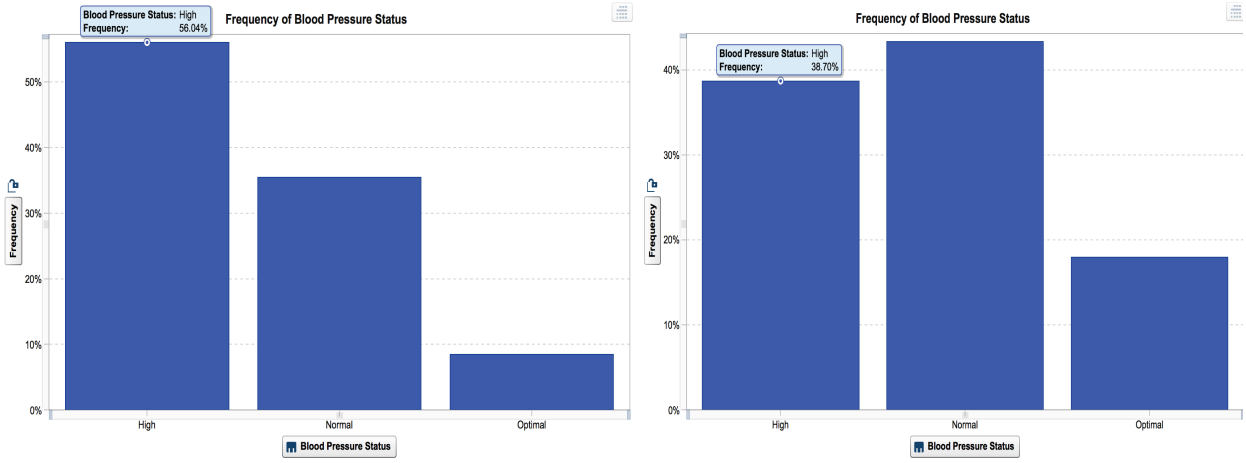
For patients with high cholesterol levels, their blood pressure status is correspondingly higher. And the frequency of high-pressure status vs. cholesterol status appears most frequently in the heat map of other pairs. The hypothesis is correct.

3. Further analysis on coronary heart disease

In our dataset, about 28% (1,449/5,209) patients has been diagnosed coronary heart disease. We first group the patients with/without Age CHD Diagnosed values into two categories and denote them as 1 and 0. Then descriptive analysis and machine learning models are respectively applied to the dataset to explore the distinctive characteristics of this disease.

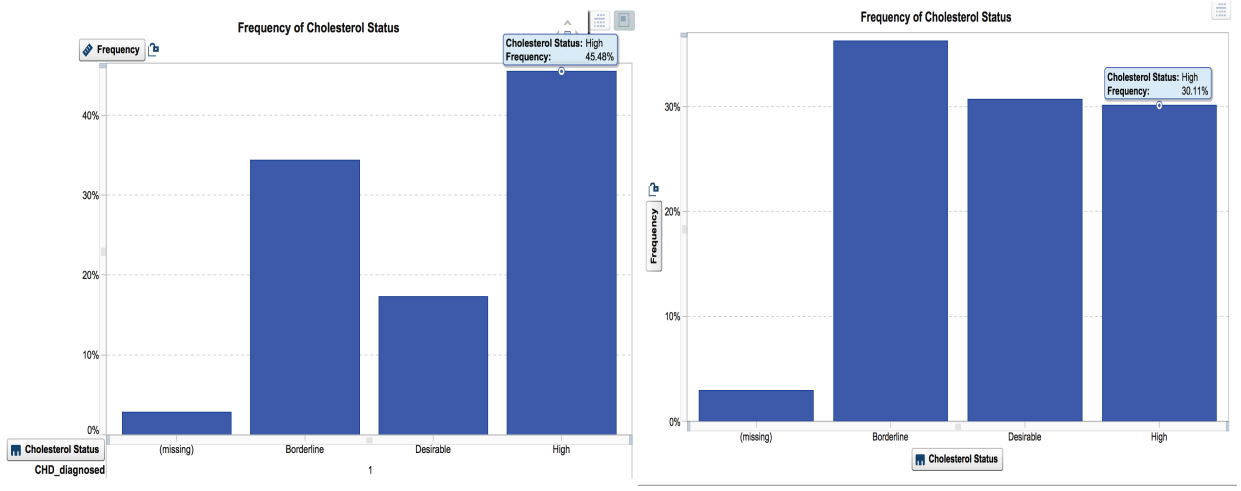
3.1 Descriptive Analysis

3.1.1 CHD occurrence relates with higher blood pressure



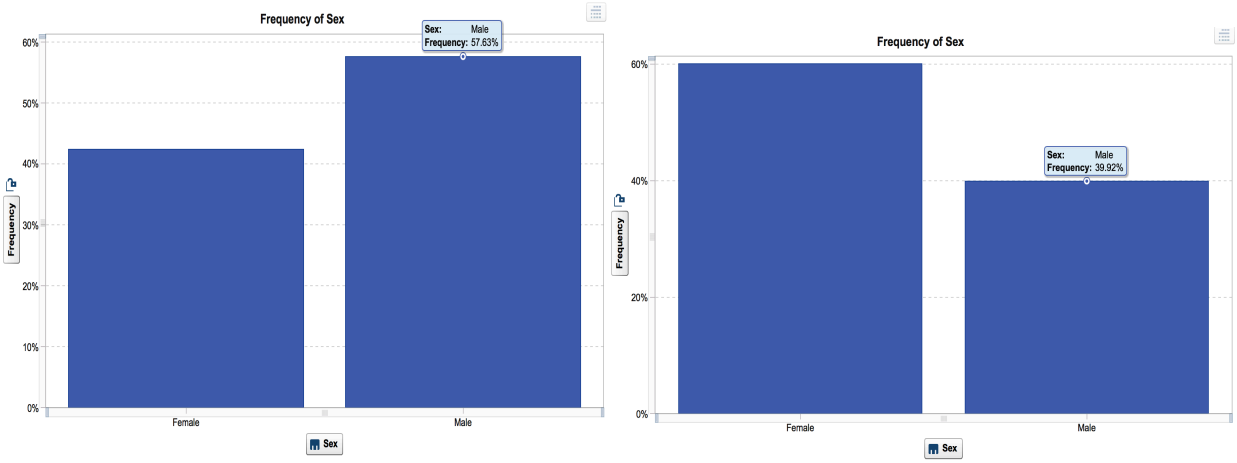
The left bar chart indicates 56.04% of patients with CHD have high blood pressure status, which is much higher than those without CHD (38.70%) in the right plot.

3.1.2 CHD occurrence relates with higher cholesterol pressure



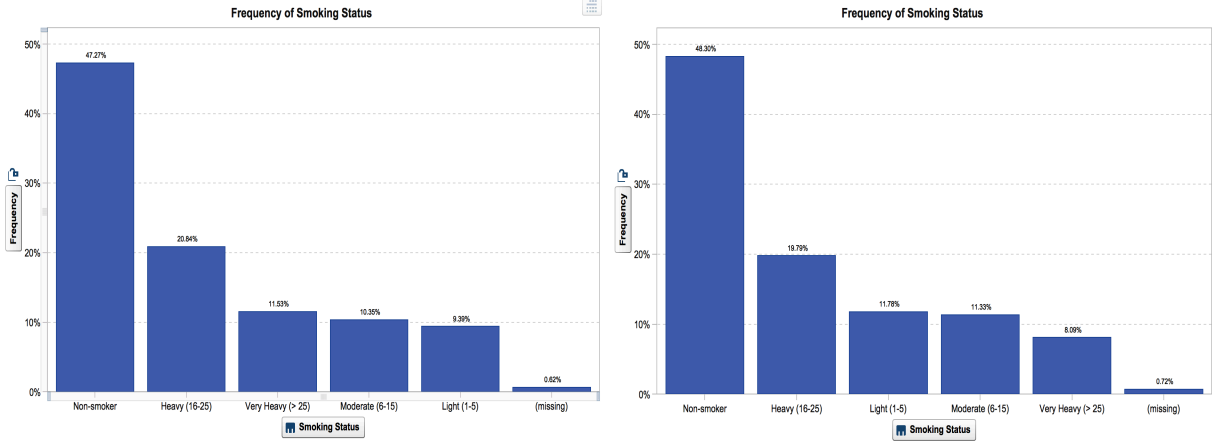
The left bar chart indicates 45.48% of patients with CHD have high cholesterol status, which is much higher than those without CHD (30.11%) in the right graph.

3.1.3 CHD occurrence relates with male gender



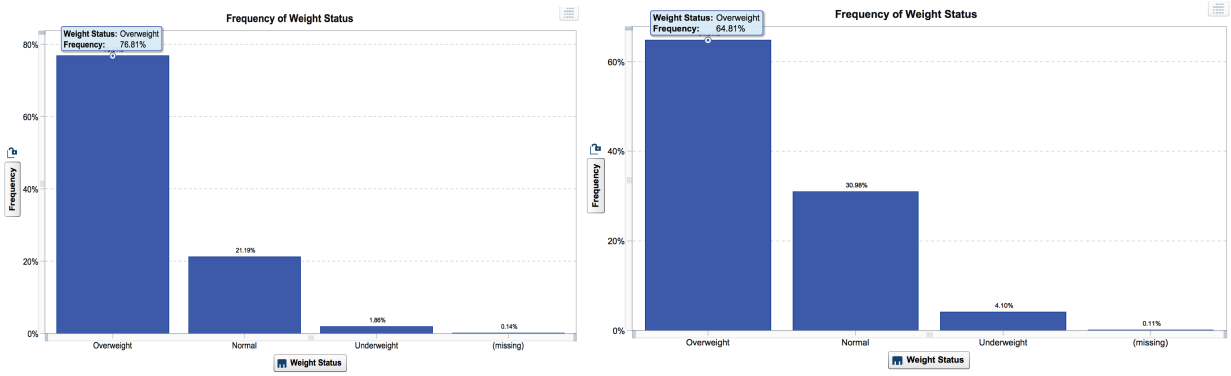
The left bar chart indicates 57.63% of patients with CHD are male, which is much higher than those without CHD (39.92%) in the right graph.

3.1.4 CHD occurrence is irreverent to smoking habits



Those CHD patients tends to have similar frequency distribution with non-CHD patients.

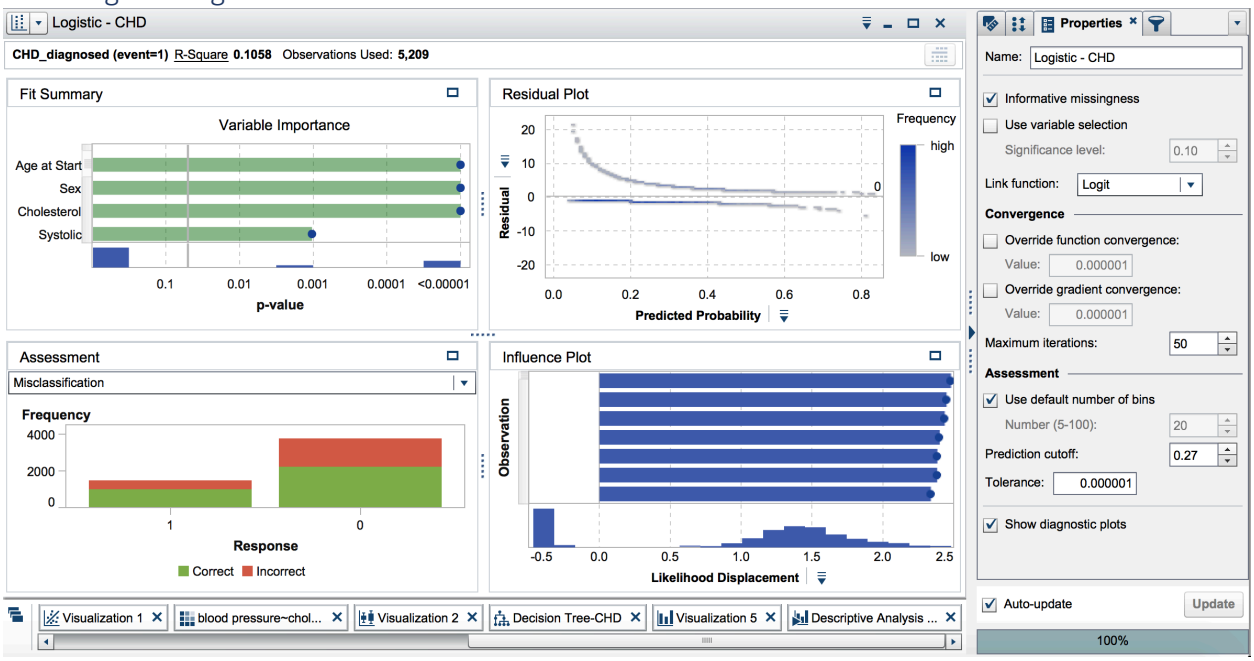
3.1.5 CHD occurrence relates with higher weight status

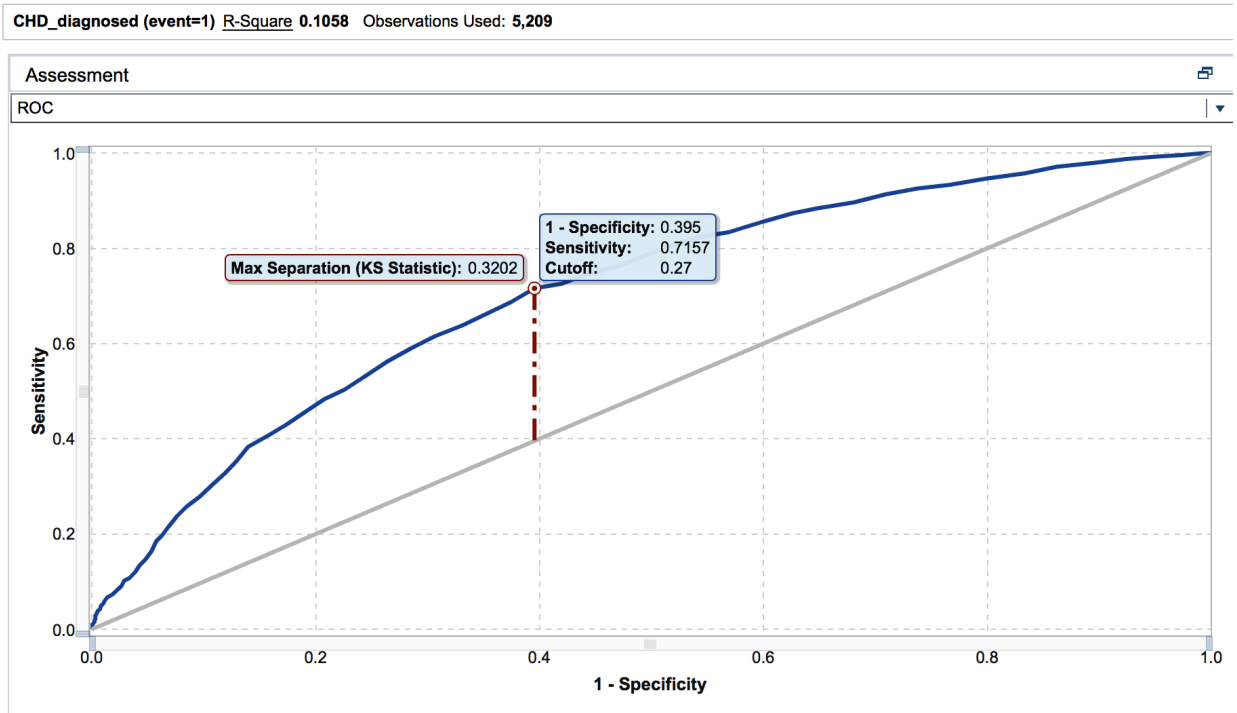


The left bar chart indicates 76.81% of patients with CHD are overweight, which is much higher than those without CHD (64.81%) in the right graph.

3.2 Models building and selection

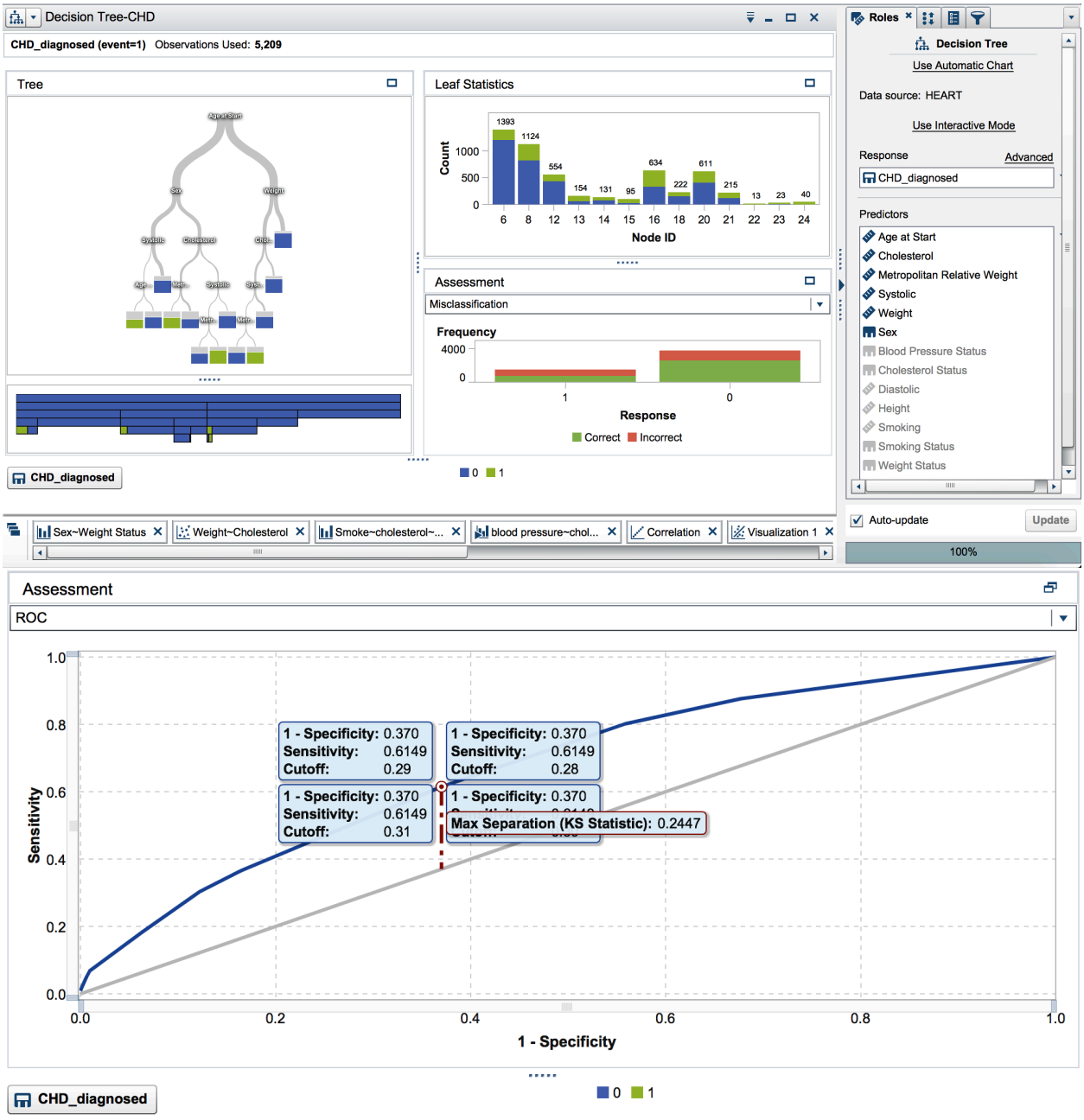
3.2.1 Logistic Regression





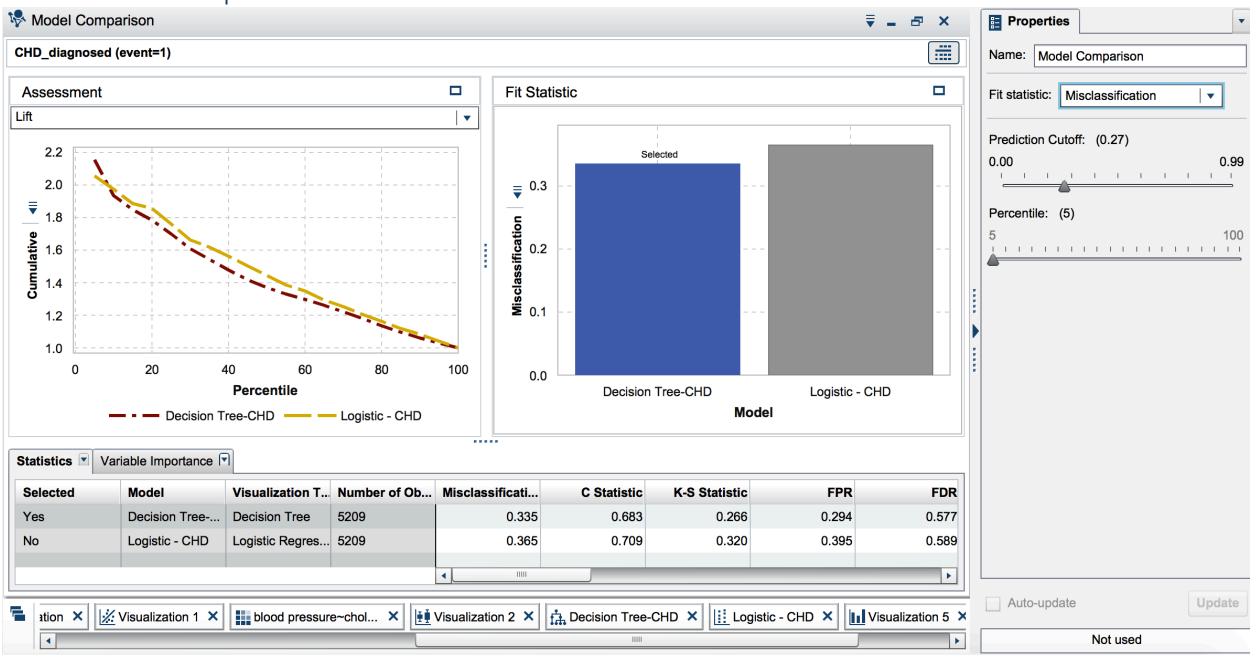
The logistic regression model selects age at start, sex. cholesterol, systolic as the important variable. And by setting the cutoff to 0.27, same as the Max separation point in the ROC plot, we could get a fair misclassification matrix with 1037 true positive, meaning for the 28% of CHD patients, 71.5% will be predicted correctly.

3.2.2 Decision Tree



The decision tree model selects age at start, cholesterol, diastolic, metropolitan relative weight, weight and sex as the important variable. And by setting the cutoff to 0.27, same as the Max separation point in the ROC plot, we could get a misclassification matrix with 811 true positive, meaning for the 28% of CHD patients, 55.97% will be predicted correctly.

3.2.3 Model Comparison



By comparing the of two models, we choose the decision tree as our final model with lower misclassification statistics, meaning decision tree will produce less wrong classification outcomes.

3.3 Main findings

People who suffered from coronary heart diseases are highly related to higher blood pressure, higher cholesterol and heavier weight. Male tends to have higher risk of getting this disease, while smoking habits is not a significant influencing factor for such disease. The best decision tree model trained also selects the corresponding variables including age at start, cholesterol, diastolic, metropolitan relative weight, weight and sex.