Introduction

As part of engagement, this study assists in analyzing data about patients and providing information to medical research teams. We will test the four hypotheses below and try to find assumptions about the potential causes of coronary heart disease from coronary heart disease data.

H1: The weight and cholesterol levels are correlated

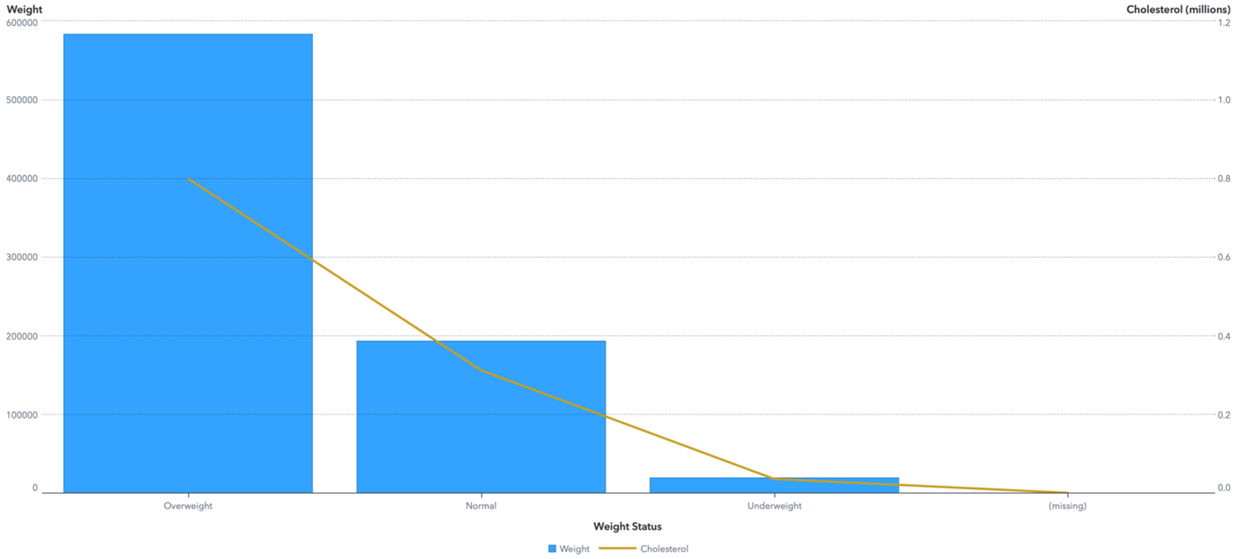
H2: Men are usually more obese than women

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

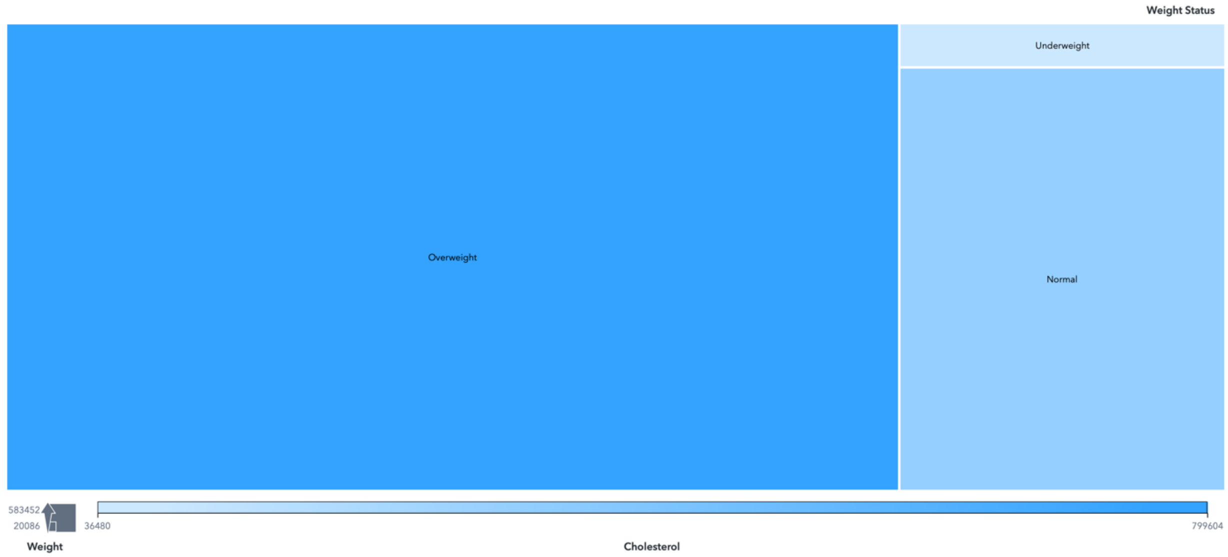
H4: The blood pressure is higher for people with higher cholesterol levels

We will use Heart data, which contains information about patients related to heart disease. Since the data contains missing data, it is necessary to restore or add additional data sets to make more accurate analysis in the future.

H1: The weight and cholesterol levels are correlated

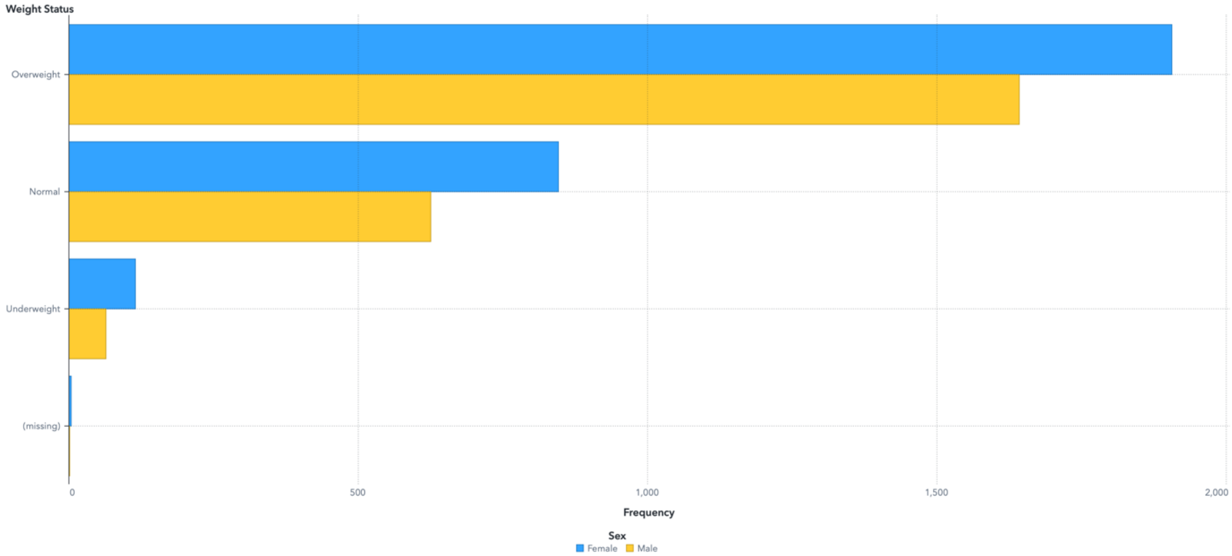


First, I compared the state of the body weight and the cholesterol level. The two data were compared using a bar chart. Looking at the two data according to weight status, weight and cholesterol, respectively, when the weight status was overweight, the weight was 583,452 and cholesterol was 799,604, and the cholesterol level was lower as the weight status was underweight. In other words, we can say that there is a correlation between weight and cholesterol through the bar chart.

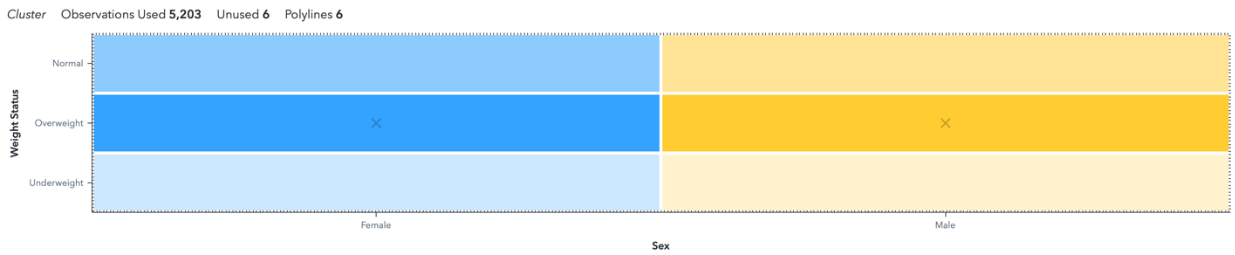


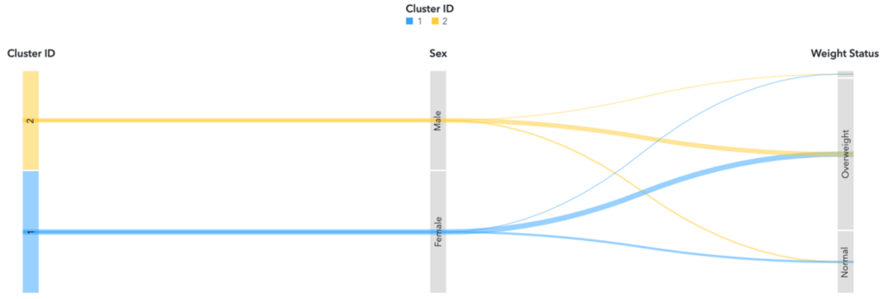
To see more details, I examined how weight and cholesterol are correlated with a Treemap. In the Treemap chart, it was clearly confirmed that the Cholesterol level was higher in the overweight people. Therefore, we could confirm that there was a significant correlation between overweight people and high cholesterol levels. Therefore, H1 is correct.

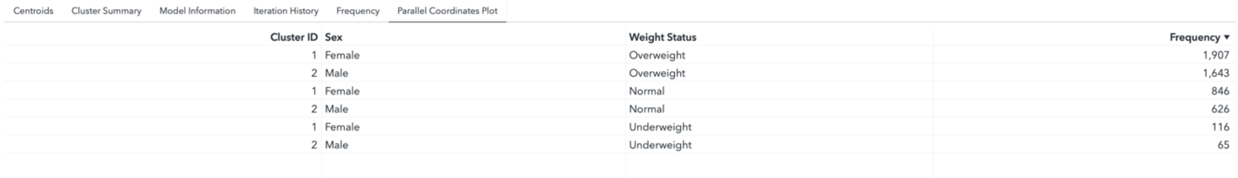
H2: Men are usually more obese than women



First, I looked at weight status according to gender. Overall, it was confirmed that the number of women was higher than that of men in each weight status. However, only this alone could not compare the degree of obesity between women and men, so I compared gender by clustering according to weight status. The largest cluster was female, with 1907 overweight, followed by overweight male with a frequency of 1637. The data showed that both men and women had the highest number of overweight.





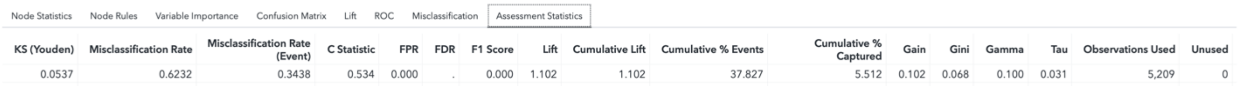


Cluster ID	Sex	Weight Status	Frequency ▾
1	Female	Overweight	1,907
2	Male	Overweight	1,643
1	Female	Normal	846
2	Male	Normal	626
1	Female	Underweight	116
2	Male	Underweight	65

Also, looking at the path of the cluster, the thickest line was an overweight woman. According to the bar chart and cluster graph above, it can be confirmed that the number of overweight women is higher than that of overweight men. Therefore, the second hypothesis, Men are usually more obese than women, is incorrect.

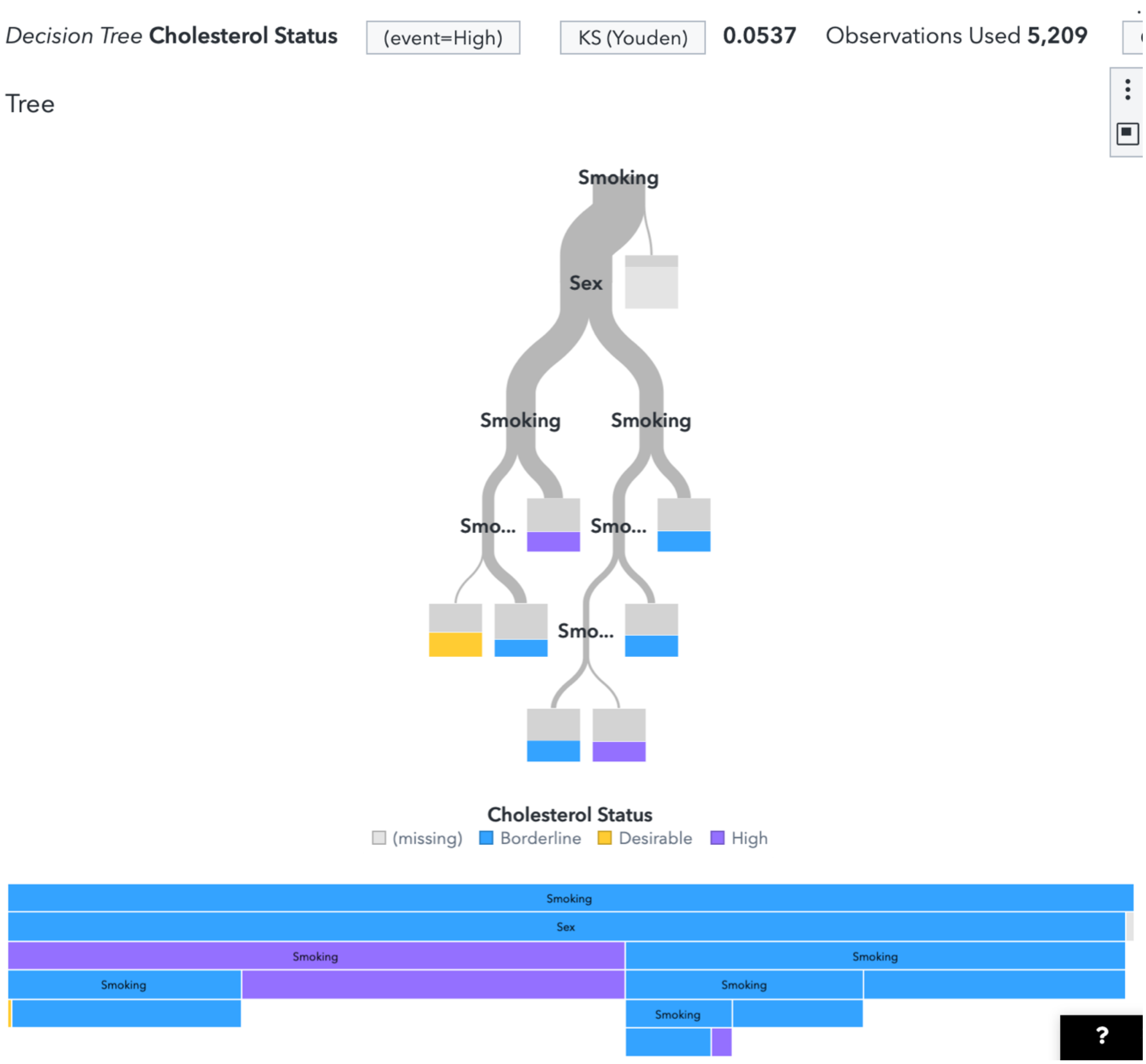
<u>H3: Women usually smoke less than men, but their cholesterol level is higher</u>

To confirm H3, we need to examine whether women smoke less than men and then determine whether women have high cholesterol. Therefore, I thought that the most appropriate visual graph for this hypothesis is a decision tree.



KS (Youden)	Misclassification Rate	Misclassification Rate (Event)	C Statistic	FPR	FDR	F1 Score	Lift	Cumulative Lift	Cumulative % Events	Cumulative % Captured	Gain	Gini	Gamma	Tau	Observations Used	Unused
0.0537	0.6232	0.3438	0.534	0.000	.	0.000	1.102	1.102	37.827	5.512	0.102	0.068	0.100	0.031	5,209	0

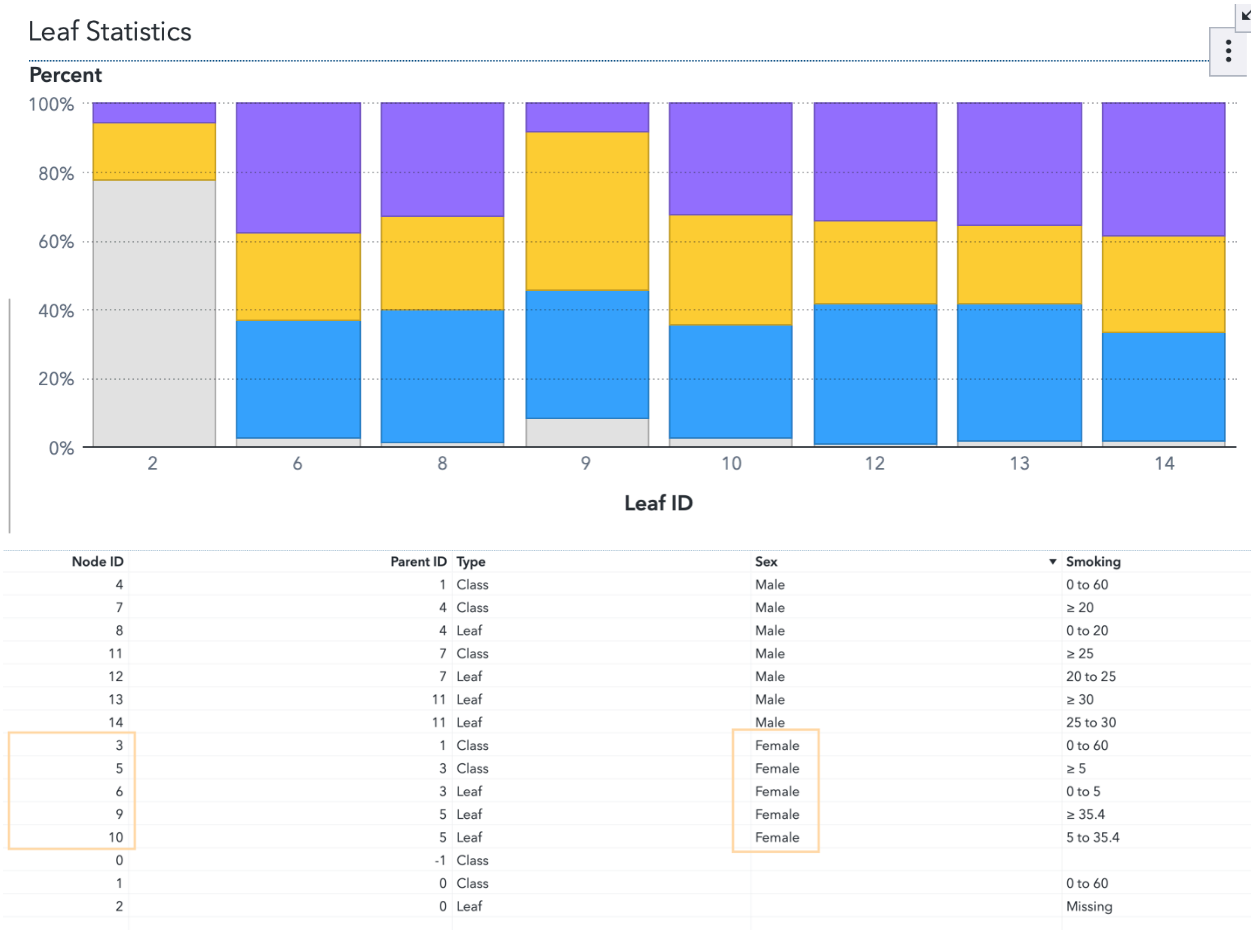
First, since KS (Youden) is 0.0501 and the misclassification rate is 0.632, I judged this decision tree to be appropriate and meaningful.



Decision Tree **Cholesterol Status** (event=High) KS (Youden) **0.0537** Observations Used **5,209**

Tree

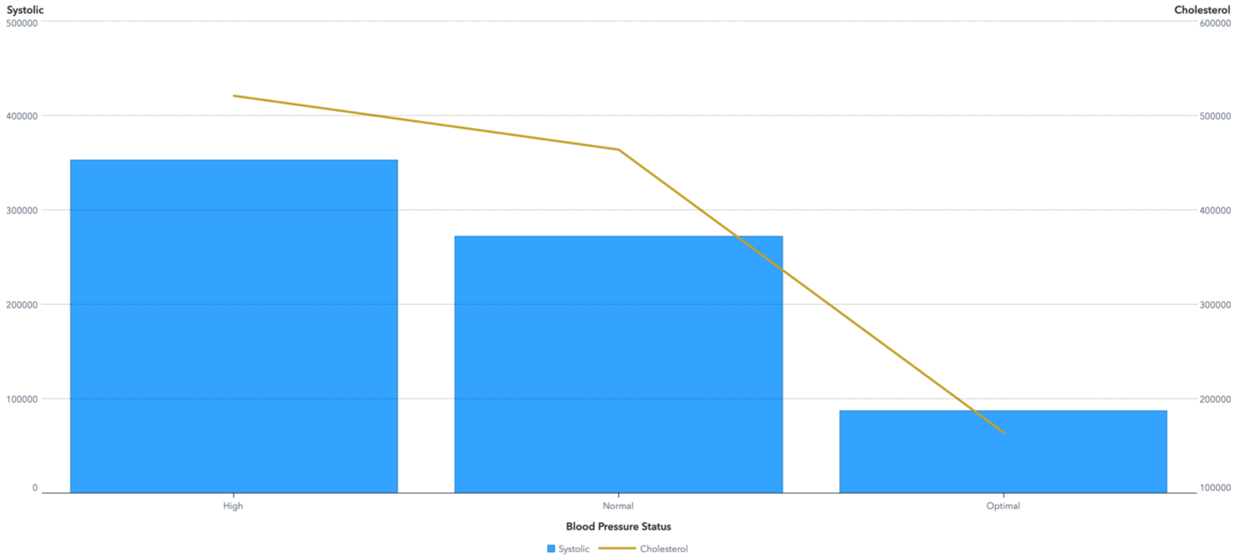
Looking at the decision tree in detail, the first node starts from gender (Node ID 1), and from the second node, it refers to the smoking rate according to gender. When moving to the second node, the left stem is female (Node ID 3) and the right stem is male (Node ID 4), but you can see that the left stem is thicker. In other words, the smoking rate of women is higher. In addition to this, the color displayed on the decision tree shows the cholesterol status. Violet color indicates a high cholesterol state, and the fact that most of the women who smoke show purple color indicates that they have high cholesterol.



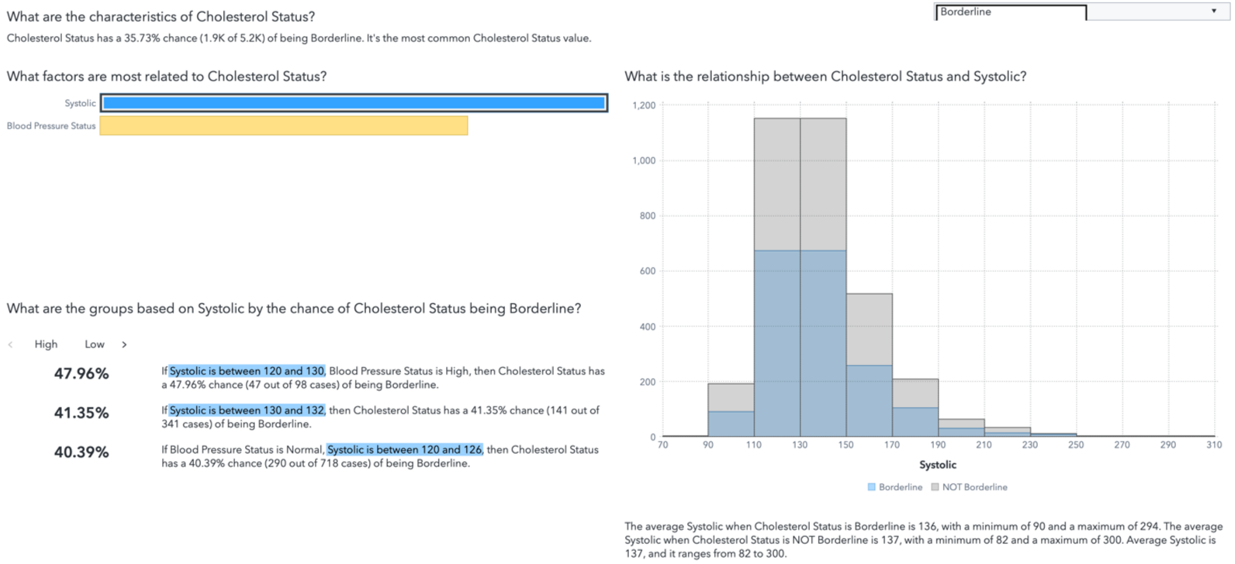
Node ID	Parent ID	Type	Sex	▼ Smoking
4	1	Class	Male	0 to 60
7	4	Class	Male	≥ 20
8	4	Leaf	Male	0 to 20
11	7	Class	Male	≥ 25
12	7	Leaf	Male	20 to 25
13	11	Leaf	Male	≥ 30
14	11	Leaf	Male	25 to 30
3	1	Class	Female	0 to 60
5	3	Class	Female	≥ 5
6	3	Leaf	Female	0 to 5
9	5	Leaf	Female	≥ 35.4
10	5	Leaf	Female	5 to 35.4
0	-1	Class		
1	0	Class		0 to 60
2	0	Leaf		Missing

I also looked at Leaf Statistics to learn more. First, leaf IDs 3, 5, 6, 9, 10 represent females, and leaf IDs 4, 7, 8, 11, 12, 13, 14 represent males. We could confirm that the other IDs except leaf ID 9 had a high percentage of high cholesterol in females. Therefore, although it is true that women have high cholesterol, Hypothesis 3, Women usually smoke less than men, but their cholesterol level is higher, is incorrect because women smoke less than men.

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



First, I looked at the correlation between blood pressure status and cholesterol through a basic graph. Blood pressure status can be divided into high, normal, and optimal categories, and a systolic was used to measure the blood pressure. As a result, I confirmed that there is a correlation between cholesterol and blood pressure. The higher blood pressure status and systolic, the higher the cholesterol level, so it was confirmed that there was a positive correlation between the two.



What are the characteristics of Cholesterol Status?

Cholesterol Status has a 35.73% chance (1.9K of 5.2K) of being Borderline. It's the most common Cholesterol Status value.

What factors are most related to Cholesterol Status?

What are the groups based on Systolic by the chance of Cholesterol Status being Borderline?

47.96%	If Systolic is between 120 and 130, Blood Pressure Status is High, then Cholesterol Status has a 47.96% chance (47 out of 98 cases) of being Borderline.
41.35%	If Systolic is between 130 and 132, then Cholesterol Status has a 41.35% chance (141 out of 341 cases) of being Borderline.
40.39%	If Blood Pressure Status is Normal, Systolic is between 120 and 126, then Cholesterol Status has a 40.39% chance (290 out of 718 cases) of being Borderline.

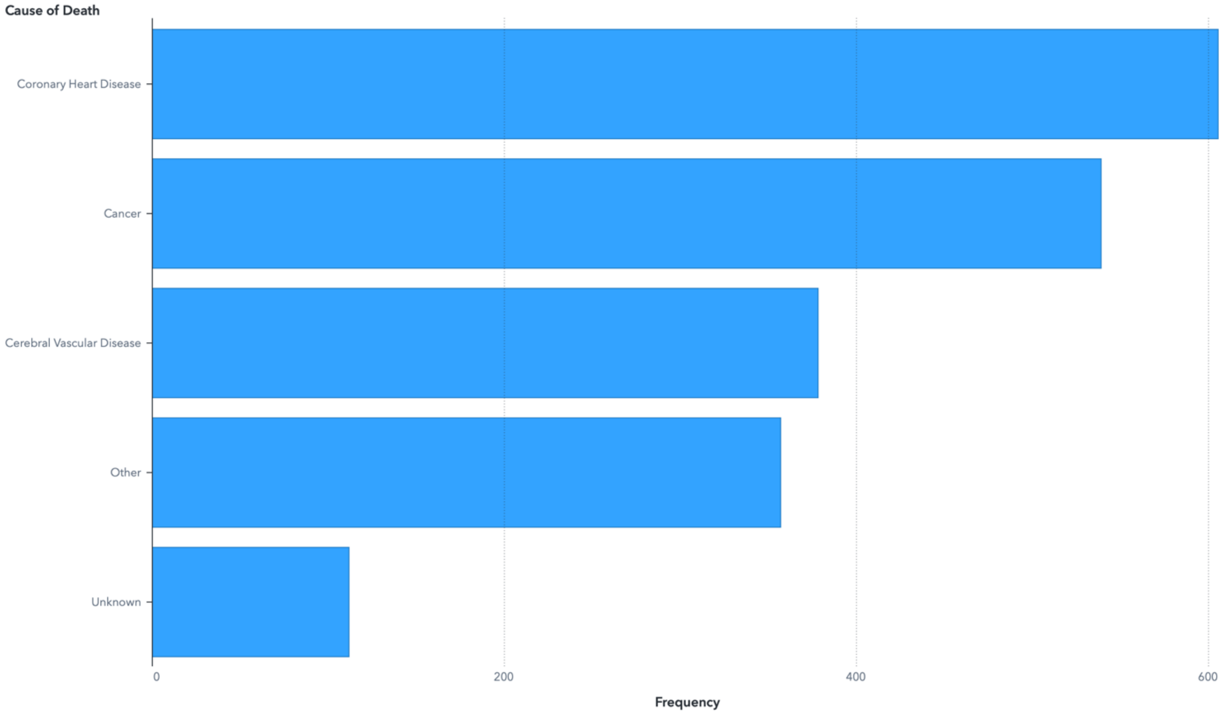
What is the relationship between Cholesterol Status and Systolic?

The average Systolic when Cholesterol Status is Borderline is 136, with a minimum of 90 and a maximum of 294. The average Systolic when Cholesterol Status is NOT Borderline is 137, with a minimum of 82 and a maximum of 300. Average Systolic is 137, and it ranges from 82 to 300.

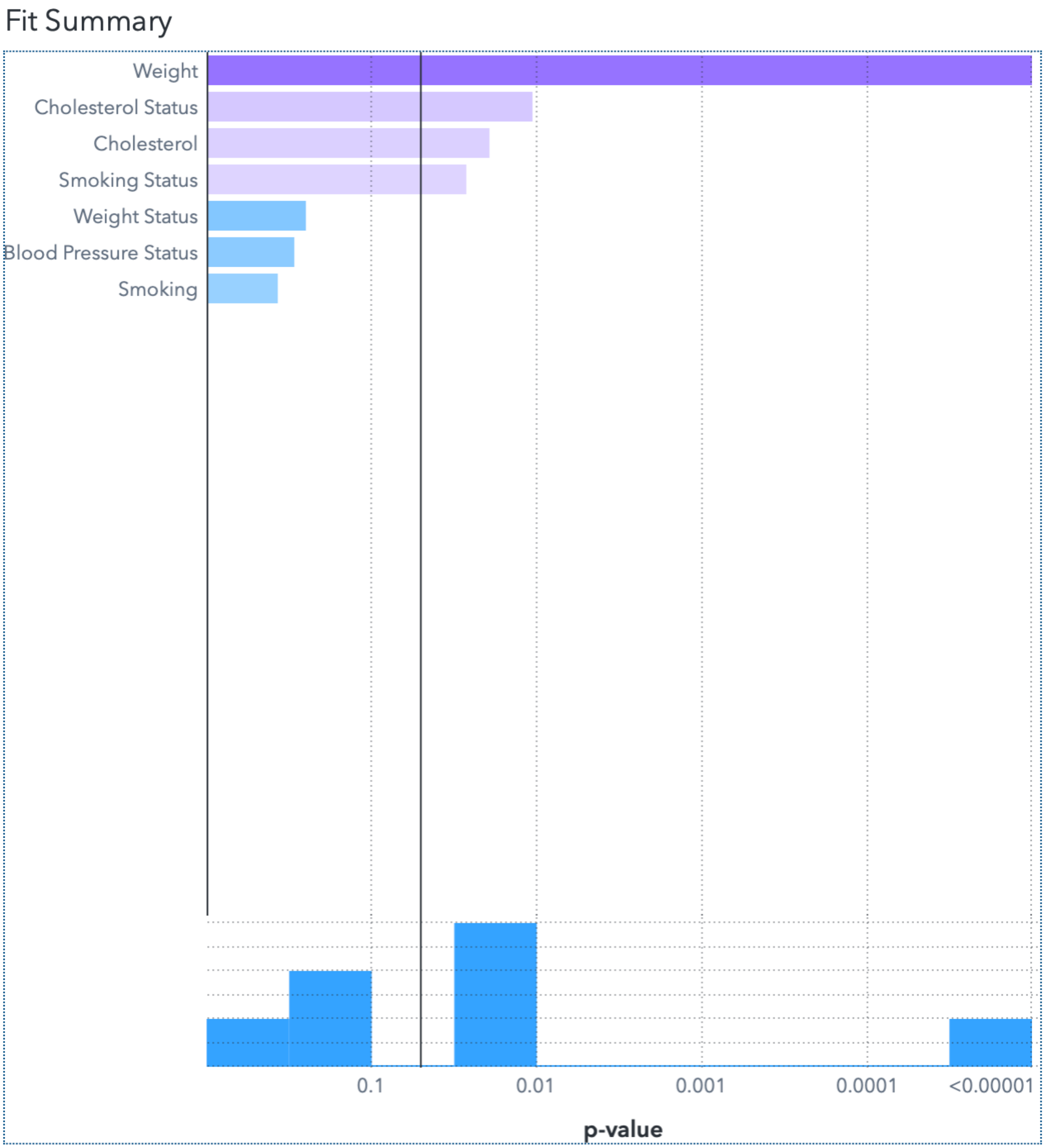
To find out more, explanation analytics using the same data was run. Through the above graphs, I confirmed that there is a high correlation between Blood Pressure and Cholesterol Status. In addition to this, when I compared it with other data (Smoking and Weight) that were previously dealt with by examining several hypotheses, I was able to confirm that Blood

Pressure had the greatest effect on cholesterol status. Therefore, Hypothesis 4, The blood pressure is higher for people with higher cholesterol levels, is correct.

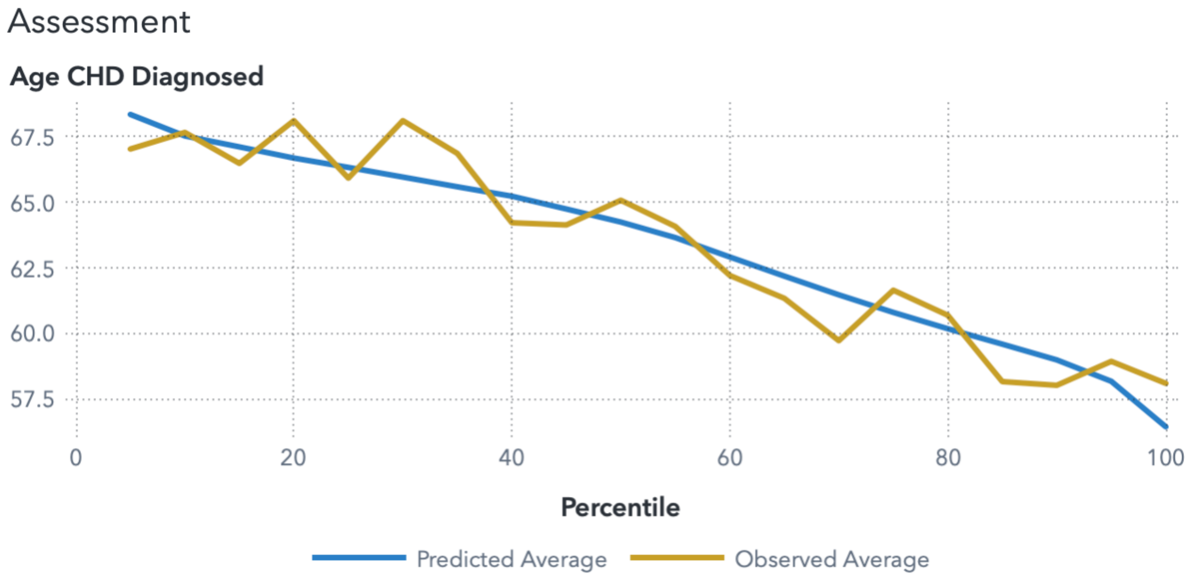
The coronary heart disease and trying to find the assumption about potential cause of the coronary heart disease in the data.



First, I looked at the cause of death in the data. As a result of examining the basic bar chart, coronary heart disease accounted for the highest rate among the causes of death.



Next, I attempted to find the cause of coronary heart disease. I performed Linear Regression to find the factors most influencing CHD Diagnosed. As a result of linear regression analysis, the factors with p-value less than 0.05 were Weight, Cholesterol Status, Cholesterol, and Smoking Statue. Therefore, it can be said that these four factors are the potential causes of coronary heart disease.



Finally, I looked at the Assessment graph. Since the line for the predicted values and the line for the observations are close, we can be sure that the linear regression model is meaningful. And it can be confirmed that the factors that most influence the coronary heart disease are weight and cholesterol level.

Conclusion

Taken together, I found that there was a high positive correlation between weight and cholesterol, and that women were more obese than men. In addition, not only did women smoke more than men, but both women and men had very high cholesterol levels when they are smoker. Also, there is a significant positive correlation between cholesterol and blood pressure. Moreover, to avoid coronary heart disease, control of body weight and cholesterol is necessary, and smoking should be reduced.