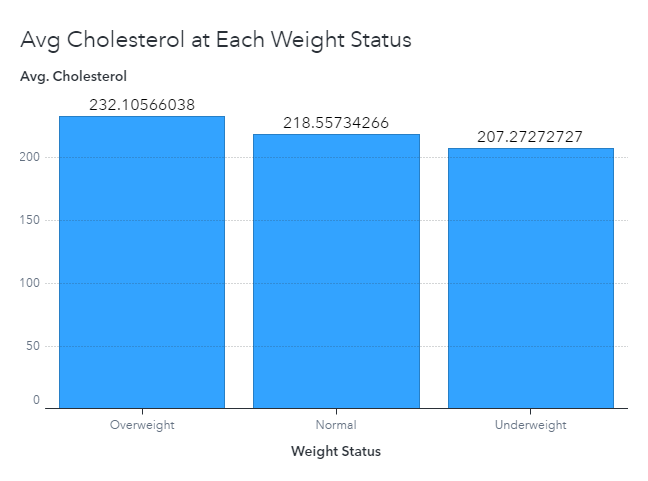
**About the Report:**

This report was made by the dataset of Heart from SAS Viyas for learners. There were total 5209 people were involved in this dataset and about 55% is women with 2873 and 45% of men with 2336. This Report is made for the large hospital chain.

Before started to talk about the relationship between the weight and cholesterol, I would need to clarify that I filtered out all data that is shown as “missing” because it would increase the accuracy of the data rather than putting mean value on “missing” part.

**Hypothesis 1 : The weight and cholesterol levels are correlated**

Mostly we predict that weight and cholesterol have strong relationship and knowledgeable people about healthy always highlight that overweight will cause many diseases. Yes, overweight may be the main cause of many diseases and illness, and we can avoid these if we diet and take careful about ingredients. Although the relationship between weight and diseases has been prove, how about the relationship between weight and cholesterol levels?

 Figure 1

From the bar chart that have the average weight by each weight status, we may misunderstand that the overweight has higher cholesterol level than normal and underweight so it have the relationship between Cholesterol level and Weight Status. How about the another bar chart below.

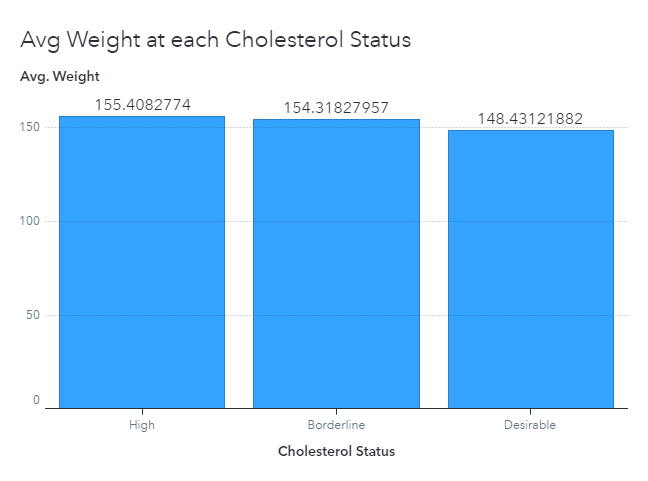


Figure 2

It is the same dataset only changing the weight status to cholesterol status with changing average cholesterol level to average weight. From the bar chart above, we may think that cholesterol status with the weight may not have relationship and we need to alter our thoughts from figure 1 to figure 2 which is had relationship to not have relationship. So I developed another bar chart with line. The line will describe the Average cholesterol and bar will describe the average weight by each weight status.

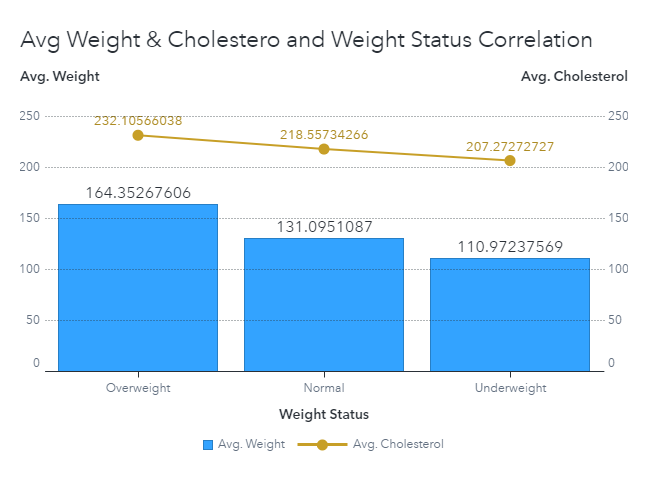


Figure 3

From the figure 3, we could understand that category of the overweight just have little bit higher cholesterol like it has higher value for weight. However, we may still confuse that it may have the relationship, so I decided to use different chart called scatter plot.

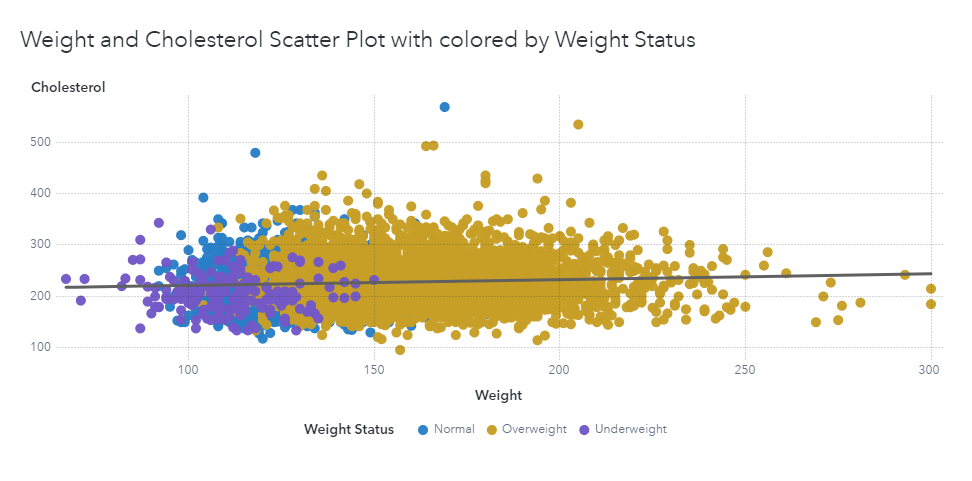
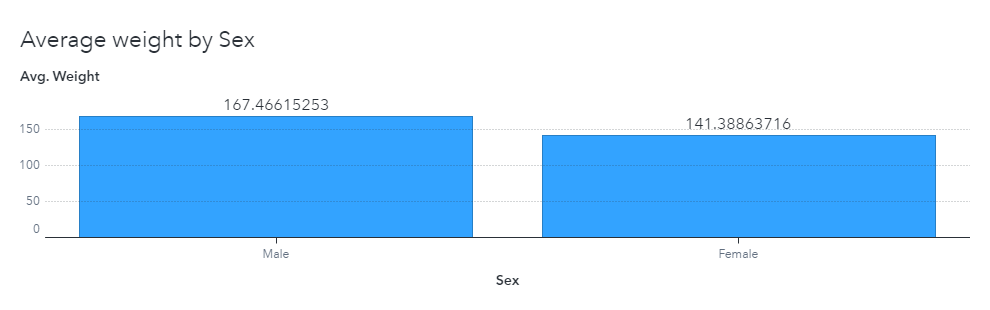


Figure 4

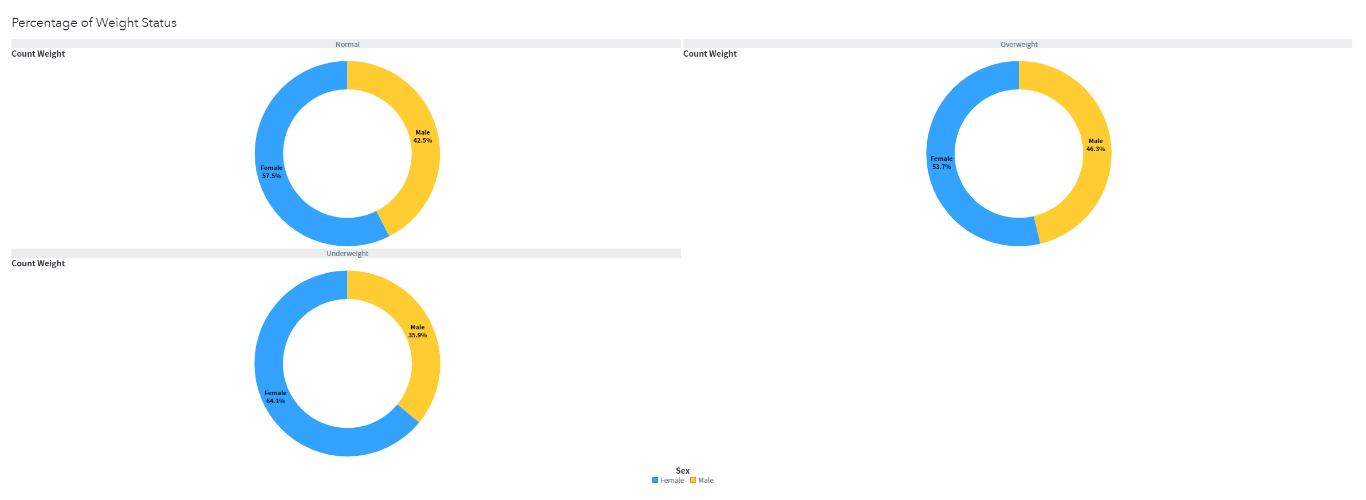
From figure 4, blue circle is for normal weight, purple circle is for underweight, and gold circle for overweight. From the figure 4, we could clarify that there are no relationship between weight and cholesterol unlike our prediction in previous. Although there are many gold circle is located right side of the chart however it does not interpret that it has the relationship between cholesterol and weight.

Lastly, to verify the hypothesis of relationship between weight and cholesterol is incorrect, I wanted to check the correlation rate. The correlation was 0.07 and the correlation of 0.07 is not strong value to explain the relationship between weight and cholesterol. So I want to say that the hypothesis of have a relationship between weight and cholesterol is false.

**Hypothesis 2 : Men are usually more obese than women**

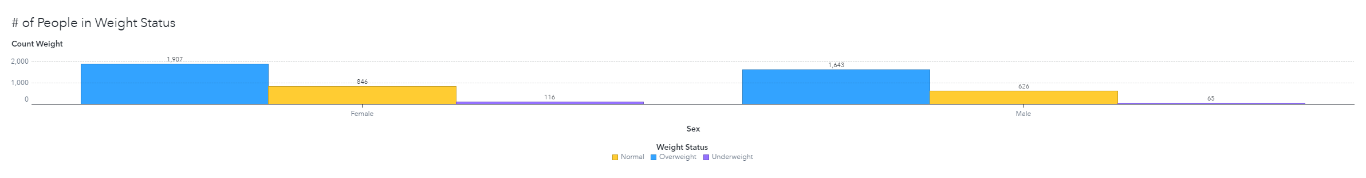
Before start to interpret of the hypothesis, there are total 5209 people were in the dataset with average weight for male was 167.47 pounds and 141.39 pound for women.

Figure

To clarify the hypothesis, I need to understand how many people will cover in each weight status. I 

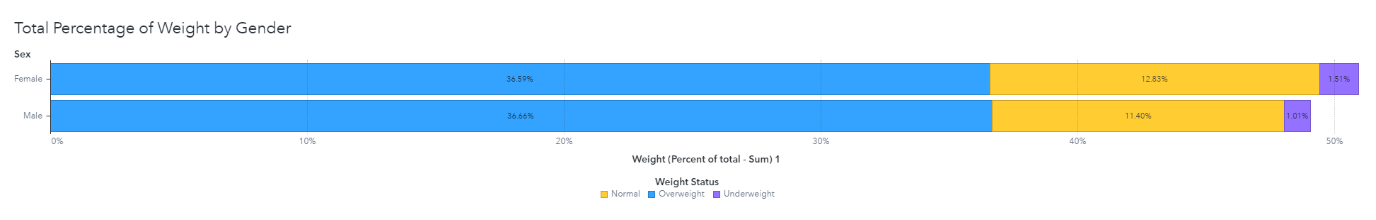
Figure

From the pie chart at figure 6, many women are in each weight category with higher percentage. From this figure, I could understand that each category of weight has higher ratio of women than men. However, this chart is not actually helping to prove the hypothesis so I decided to see the actual number people in each category rather than percentage.



Figure

In fact it is same data that only difference is the percentage or actual number. However, from the figure, I could understand the whole data that there are a lot of overweight people in the dataset and they are the customer of the hospital.

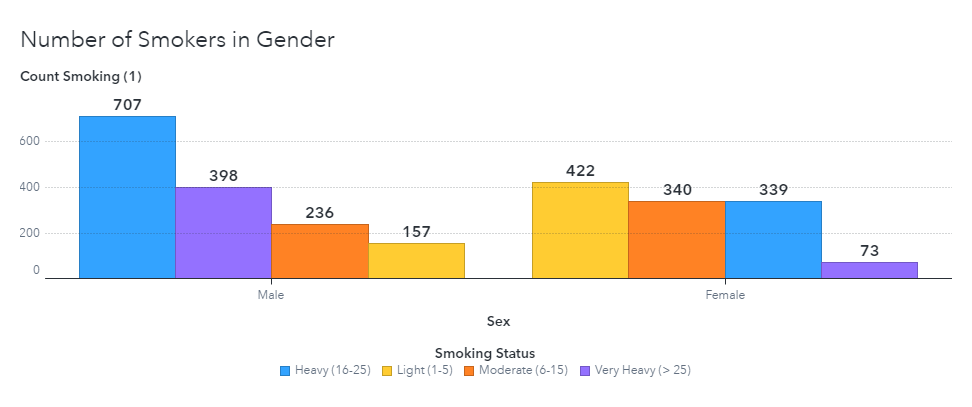


Figure

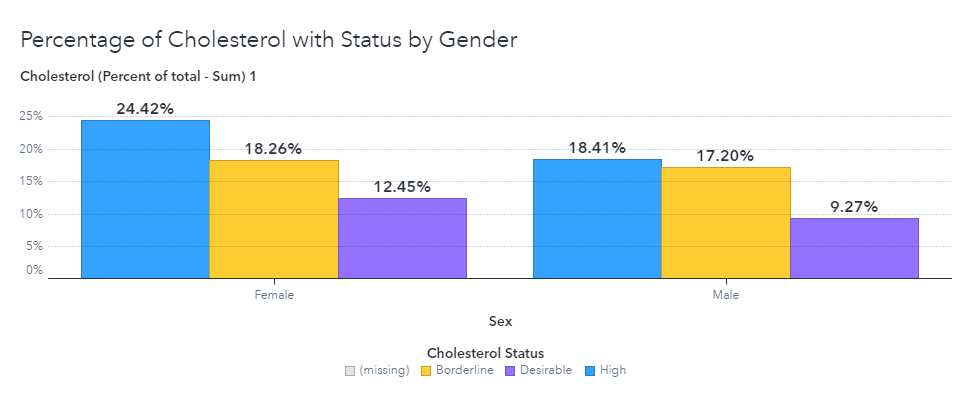
Above figure, we were able to say the incorrection of hypothesis. Although there are more women in each weight category and less weight value for female. The percentage of each weight category shows that the overweight female is 36.59% while male is 36.66%. Although there is a small difference between these percentages, however, this is under expected calculation. Therefore, the hypothesis of men obese more than women would be incorrect hypothesis on this dataset.

**Hypothesis 3 : Women usually smoke less than men, but their cholesterol level is higher**

When we ranked the smoking status by the gender, most of smokers in male is Heavy smoker while female is light smokers.



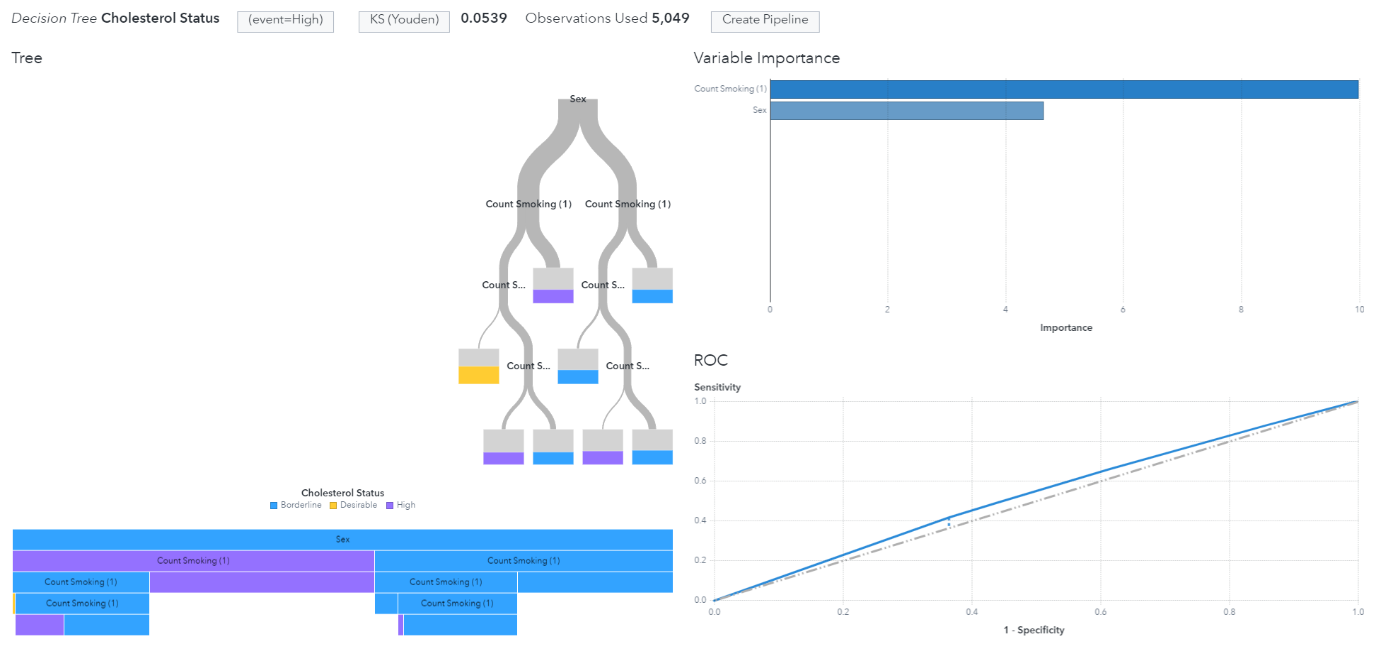
From the above bar chart, we could find the hypothesis that about 73 percent of male smoke a cigarette more than 16 cigarette while 64% of female smoke less than 15 cigarette.



Figure

From the figure 9, we could also see that 24.42% of female and 18.41% of male is under status of High cholesterol. When we see this figure, we also can agree about the hypothesis that number of women have higher cholesterol than male.

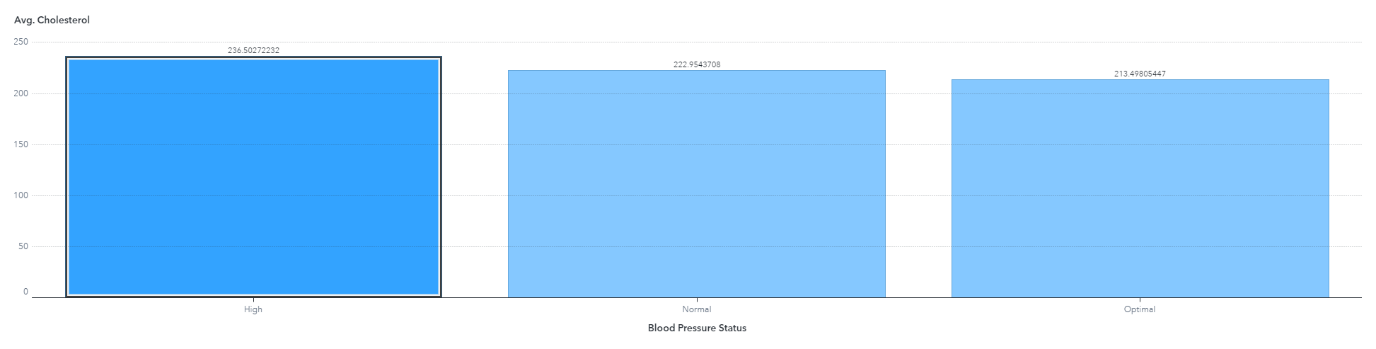
However, this is not the perfect tool to compare our hypothesis, so I decided to use decision tree.



Figure

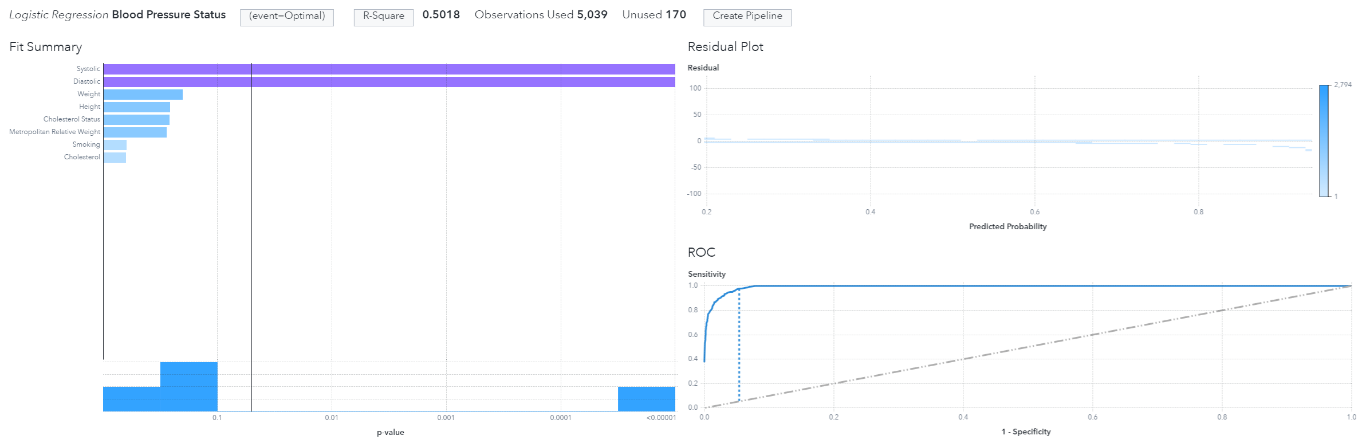
From the figure 10, I was able to follow the decision tree to see the hypothesis. Although we could see the evidence of women usually smoke less than men, we couldn’t find the specific values for women has higher cholesterol level than men.

**Hypothesis 4 : The blood pressure is higher for people with higher cholesterol levels**

****

Figure

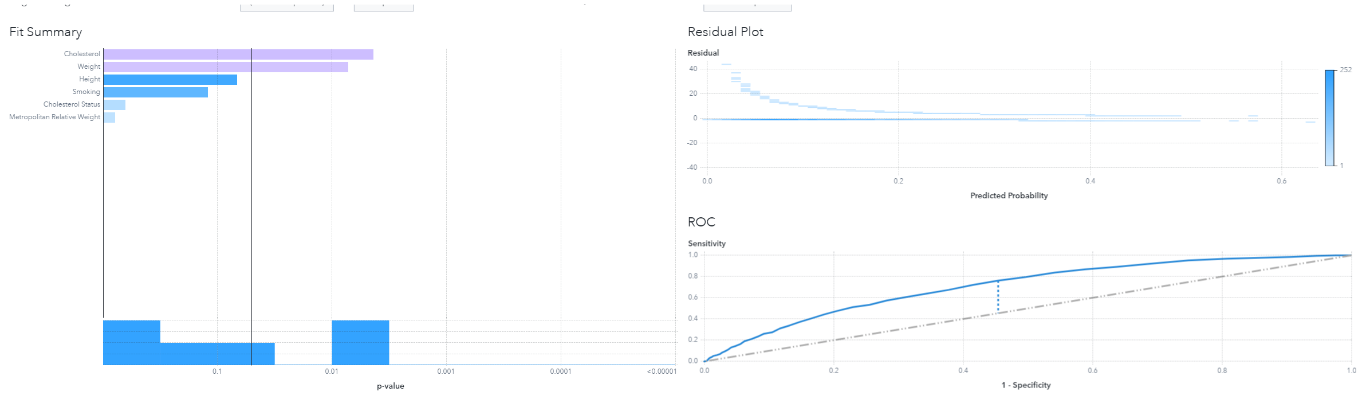
When we check the average cholesterol value for blood pressure status, I could see the little bit higher value for high blood pressure had higher cholesterol value. However, it does not interpret the whole hypothesis to prove. To clarify and prove the hypothesis, I decided to use logistic regression model. From my basic knowledge, I knew that systolic and diastolic is the measurement of blood pressure so I decided to use it for regression model and I also created another model without these two.



Figure

Before explain the about the model, I would like to highlight the thick line at Fit Summary. Simply when the line is long, it means that it has strong relationship and when it is short, it doesn’t.

From the figure 12, I made the regression model with Systolic and Diastolic. Obviously, Systolic and Diastolic has the strongest relationship with Blood Pressure Status.

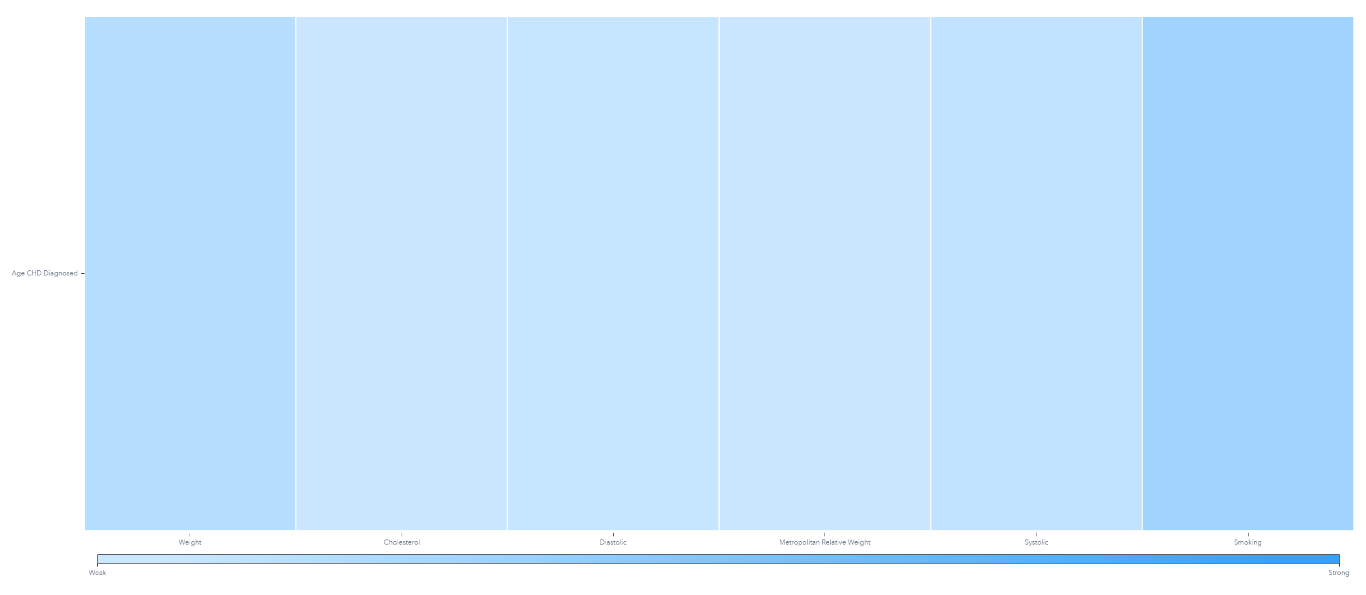


Figure

From the Figure 13, this is the regression model without Systolic and Diastolic. From this model, we could check the line of Cholesterol that is long compared to others such as weight, height, smoking, and metropolitan relative weight. On top of that, Cholesterol has p-value of 0.004. when the p-value is less than 0.05 means that we accept the null hypothesis which is the blood pressure is higher for people with higher cholesterol levels.

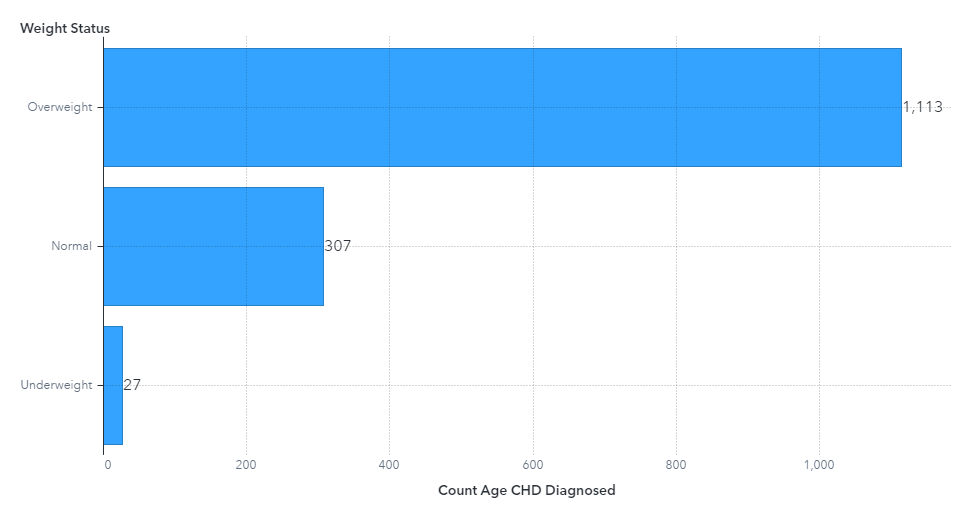
**Hypothesis 5 (Addition) : What causes CHD**

From this hypothesis, I want to see the relationship between CHD vs others such as weight, cholesterol, diastolic, metropolitan relative weight, systolic, and smoking. It will help to find which one or two thing will cause the most out the list for CHD.

To see the relationship between these, I decided to use correlation chart. 

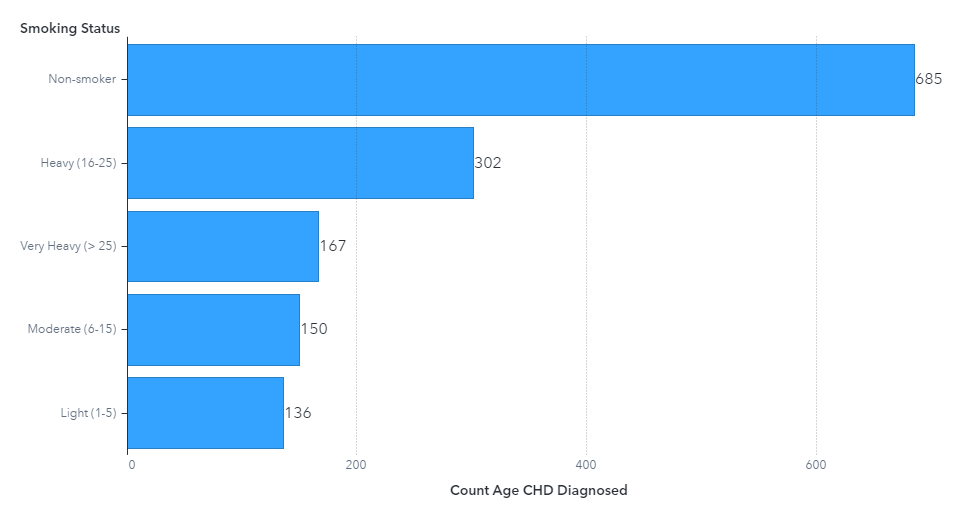
Figure

This is the correlation chart between Age of CHD Diagnosed vs others. To use correlation charts, we need to carefully see the depth of color. For instance, when the color is lighter means that has weak relationship, when it has dark sky blue that has stronger relationship with CHD Diagnosed. From the figure 14, we could find the top two value that may have causes for CHD diagnosed which is Weight and Smoking. Therefore, I decided to see the bar chart to compare between number of CHD diagnosed vs smoking Status and number of CHD diagnosed vs Weight Status.



Figure

From the figure 15, we could be able to see that overweight may be the cause of get CHD diagnosed.



Figure

From the figure 16, interestingly non-smoker had higher value for CHD diagnosed compared to heavy smoker or very heavy smoker. Does it mean that smoking will help to reduce of CHD diagnosed? We need further research about this interesting result to answer the question. It is too early to verify another hypothesis of light smoking will help to reduce of CHD diagnosed.

To sum up of the Hypothesis 5, smoking and weight has the relationship of causing CHD diagnosed.

**Business opinion**

To maximize the analyst result, the hospital needs to take care more about the dataset, it must be collect more data with more specific columns and row to develop the dataset. Also I was able to check many outliers from the dataset such as age was 500 and still a lived. These types of outliers must be organized better.