Heart Disease Diagnosis Analysis

By Melchor Ronquillo, Shrdha Shrestha, and Sophie Srisak

1. Introduction

For our project, we decided to analyze factors that could potentially predict the possibility of having a diagnosis of heart disease. Our dataset includes 14 variables that consist of 303 individuals' demographics such as age, sex, health conditions, and symptoms that may be used to predict whether a patient receives a diagnosis for heart disease. The dataset provides both categorical and numerical variables to work with. Some of the variables pertaining to heart related conditions are exercise-induced angina, number of major vessels, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, resting electrocardiographic results, oldpeak which is the ST depression induced by exercise relative to rest, maximum heart rate achieved, the slope of the peak exercise ST segment, thallic defect level, and the output which indicates whether the chance of diagnosis of heart disease was higher or lower based on how much the diameter of the arteries are narrowing. We decided to primarily focus on seven variables which are age ('age'), sex ('sex'), type of chest pain ('cp'), resting blood pressure ('trtbps'), cholesterol levels ('chol'), fasting blood sugar exceeding 120 mg/dl ('fbs'), and maximum heart rate achieved ('thalach') because we thought these are these variables are ones that anyone could easily read and access. According to MayoClinic, certain risk factors contribute to fatty plaque buildup inside of narrow arteries and can lead to the risk of a heart attack, angina, or stroke. The Centers for Disease Control and Prevention and MayoClinic mention that some of the risk factors include older age, high blood pressure, high blood cholesterol, obesity, and lack of physical activity to name a few. Some symptoms of heart disease are shortness of breath, chest pain, and racing heartbeat. We hope to find which factors can help predict the chance of getting a diagnosis of heart disease. Specifically, we wanted to test to see if the variables cholesterol level and chest pain type are the most important variables in predicting the chance of heart disease.

2. Statistical Test

We first created a new dataset that transcribed the data where sex, chest pain type, and high fasting blood sugar were translated into words in the new dataset called "heartcleaned." We then created a vertical bar graph plot to view the distribution of sex, chest pain, and high fasting blood sugar. In the analysis of sex graph (Figure 1), we see that among men the chance of high and low were similar however, the higher chance of heart disease was more common among women in this study sample. From this graph, we can see that there were more male participants in the study than females which has a possibility to skew results. When looking at the bar graph for chest pain type (Figure 2), we see that most of the individuals in this study had typical chest pain type while the least had asymptomatic chest pain type. We can also see that many of the individuals that had a high chance of heart disease had non-anginal chest pain type. Then, when looking at the vertical bar graph of high fasting blood sugar (Figure 3), we see that many of the individuals in the study did not have high fasting blood sugar. However, the chance of having a diagnosis for heart disease was about even among those that had high fasting blood sugar and those that did not have high fasting blood sugar.

We then created a scatterplot of the numerical variables we are looking at which are age, resting blood pressure, cholesterol level, and maximum heart rate achieved. When looking at Figure 4 of our scatter plots, none of our variables have complete separation from each other meaning that our numerical variables do have some relationship with output. Therefore, our numerical variables do have a relationship with the chance of receiving a diagnosis for heart disease.

Furthermore, we wanted to analyze the categorical variables by output so we sorted the data by output. We then did a proc frequency on the variables sex, chest pain type, fasting blood sugar level, and output. For output by gender (Figure 5), we see again that there were 207 males and 96 females in the study so there was an uneven distribution of sex. We also saw that for those that did have a high chance of having heart disease that 56.36% of them were male while 43.64% of them were female. Therefore, for our models later on we decided to exclude the variable 'sex' since it was heavily skewed due to the uneven distribution of patients taken in. Furthermore, when looking at the frequency table for output by chest pain type (Figure 6), we see that among those who had a high chance of heart disease, 41.82% had non-anginal chest pain, 24.85% had atypical, 23.64% had typical, and 9.7% had asymptomatic chest pain. Also, the frequency table for whether or not the individuals had high fasting blood sugar levels (Figure 7) showed that 86.06% did not have a high fasting blood sugar among those that had a high chance of having heart disease. Lastly, looking at the frequency of output by output (Figure 8), we can clearly see that 138 individuals had a low chance of receiving a diagnosis for heart disease and 165 individuals had a high chance of receiving a diagnosis. This is important to note that since 54.46% of individuals had a high chance and 45.54% of individuals had a low chance of receiving a diagnosis for heart disease, it was not drastically skewed in one direction which is why we can predict the output from the given variables.

Afterwards, we wanted to analyze the numerical variables we had which were age, resting blood pressure, cholesterol level, and maximum heart rate achieved in Figure 9. For those with a low chance of receiving a diagnosis for heart disease, we can see that the mean age was 56.60, the mean resting blood pressure was 134.3985507 mm Hg, and the mean cholesterol level was 251.0869565 mg/dl, and the maximum heart rate achieved was 139.1014493. On the other hand, for those that had a high chance of receiving a diagnosis for heart disease, we saw that the mean age was 52.49, the mean resting blood pressure was 129.3030303 mm Hg, cholesterol level was 242.2303030 mg/dl and the maximum heart rate achieved was 158.4666667. Also, for those with a high chance of a diagnosis, the youngest person was 29 years old and the oldest was 76 years old; for those with a low chance of a diagnosis, the youngest was 35 and the oldest was 77. By looking at the means, we have a better understanding of the averages for each output group based on the numerical variables we wanted to look at.

Given the lack of complete separation with our numerical variables, we then built a logistic model using both our categorical and numerical variables where we set output as the dependent variable with age, type of chest pain, cholesterol, resting blood pressure, fasting blood sugar exceeding 120 mg/dl, and maximum heart rate achieved. When regressed for the output, at a 95% confidence interval, we found that three variables were statistically insignificant: age, cholesterol, and fasting blood sugar. The p-value of age was shown to be 0.5626, the p-value of fasting blood sugar was 0.4922, the p-value of cholesterol was 0.4474 (Figure 10). As a result of these higher p-values, we further analyze and model the data by running a logistic model without the variables we deemed insignificant. Furthermore, this model returned an AIC of 324.534, and an AUC from the ROC plot of 0.8265. Our new model had three variables: type of chest pain, resting blood pressure, and maximum heart rate achieved. After we run the new logistic model with a stepwise selection, we see that the p-values of all the variables have significantly decreased. Type of chest pain, and maximum achieved heart rate both had a p-value of <.0001, and resting blood pressure had a p-value of 0.0034. This new model returned an AIC of 320.199 and an AUC from the ROC plot of 0.8240 (Figure 11). In terms of testing the model itself, we look at 4 different tables. Referring to Figure 11, the first table “Testing Global Null Hypothesis”, we observe that the likelihood p-value is statistically significant at Pr > ChiSq <.0001, meaning that our overall model is statistically significant. From the “Hosmer and Lemeshow Goodness-of-Fit Test”, the p-value of .4068 is insignificant at alpha = 0.05, meaning there is no statistically significant difference between the observed value and the expected values produced by the “Partition for the Hosmer and Lemeshow Test” table. Comparing the two models, we saw only a small change between the AIC and AUC, and the model that has fewer predictors still yielded better scores.

We wanted to compare how the model with variables of our initial interest went against a full model containing all the variables in the data set. We took a similar approach in building this model by running a logistic model of all the variables on output, determining which variables were statistically insignificant at an alpha level of 0.05, and plugging the variables that were statistically significant in a new logistic regression model using a stepwise selection. We first had to analyze the other variables that were not included in our initial model which were oldpeak (ST depression induced by exercise relative to rest), resting electrocardiographic results, exercise induced angina, the slope of the peak exercise ST segment , number of major vessels, and thallic defect level. A scatterplot of oldpeak on output returned no complete separation, meaning we could include this variable in our full model, and the bar graphs of each categorical variable showed some form of interaction with output. We continued to enter all the variables into the model, regressed on output, and determined that the variables age, resting blood pressure, cholesterol level, fasting blood sugar, resting electrocardiographic results, and the slope of the peak exercise ST segment_were not statistically significant at an alpha level of 0.05 (Figure 12). Referring to Figure 13, the new full model returned an AIC of 253.770 and an AUC from the ROC plot of 0.9074. We ran another model including only the statistically significant variables and applied a stepwise selection and it returned a little better of an AIC of 250.236 but a little worse AUC from the ROC plot of .8980. To test this model, we refer to the information from Figure 14. Looking at the "Testing Global Null Hypothesis", we observe that the likelihood p-value is statistically significant with the p-value <.0001, which means that this model overall is statistically significant. From the "Hosmer and Lemeshow Goodness-of-Fit Test", the p-value of 0.3547 is insignificant at alpha = 0.05, meaning there is no statistically significant difference between the observed value and the expected values produced by the "Partition for the Hosmer and Lemeshow Test" table. Overall this model is drastically better than the model we had created earlier with the variables we thought could predict output. The only variables that overlapped between our model and this new model were maximum heart rate and chest pain type, which means that those two variables are very significant in terms of the value of output.

3. Conclusion

Based on the logistic regression of the 13 variables, disregarding sex, and removing statistically insignificant variables in a stepwise manner, our results showed that the variables exercise-induced angina, the number of major blood vessels, old peak, chest pain type, thallic defect level, and maximum heart rate achieved were the best predictors for determining the output of whether or not an individual has a high or low chance of developing heart disease. Therefore, our thesis was incorrect since we initially stated that cholesterol level and chest pain type were the most important variables in predicting heart disease. One issue with our analysis might be due to the fact that we were not able to use the variable sex in our model because the distribution of males and females in this dataset was heavily skewed towards men. This may affect the analysis since we know that "men are generally at a greater risk of heart disease while the risk for women goes up after menopause," so sex does play a role in risk (MayoClinic). Some confounding variables that we thought may play as factors are if any of the individuals are taking medication and if they have any pre-existing conditions because they may affect certain variables as well as contribute to the possible outcomes. Overall, this may help someone in the field since we see that the variables exercise induced angina, number of major blood vessels, and old peak had the most impact in predicting the chances of heart disease. Therefore, medical professionals can pay special attention to these factors and know that if they see an increase or deviation from the norm in these variables, then the patient may be at high risk for having heart disease. They could then focus on preventative measures that will better help lower the patient's chances of developing heart disease. To extend the research, we may want to find a better sample that evenly distributes the population size of males and females to better study all the variables concisely and reduce potential error in the analysis.

4. Appendix

Figure 1.

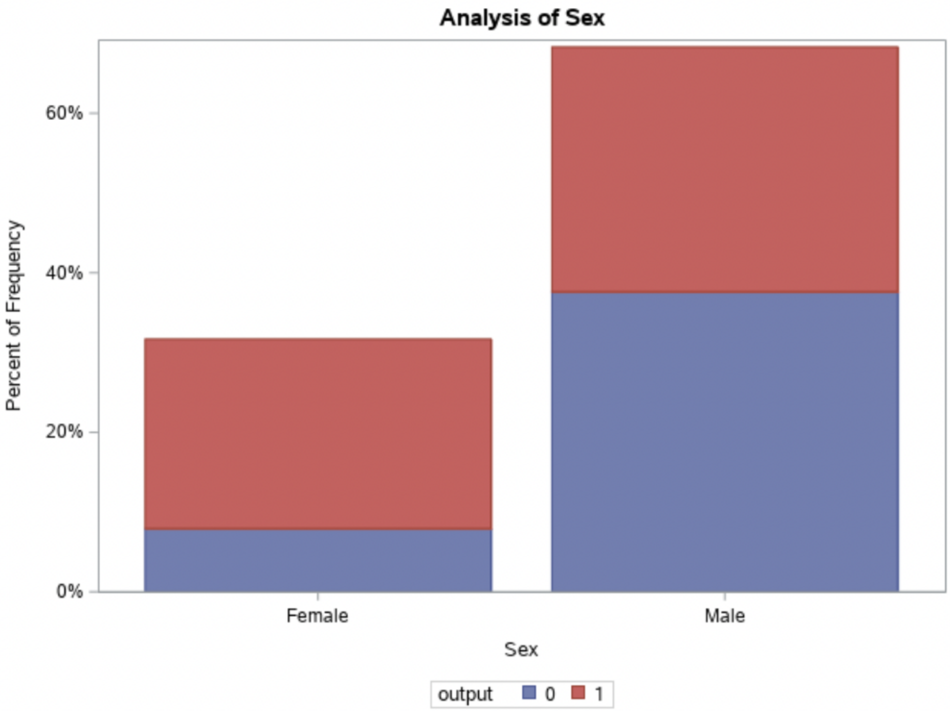


Figure 2.

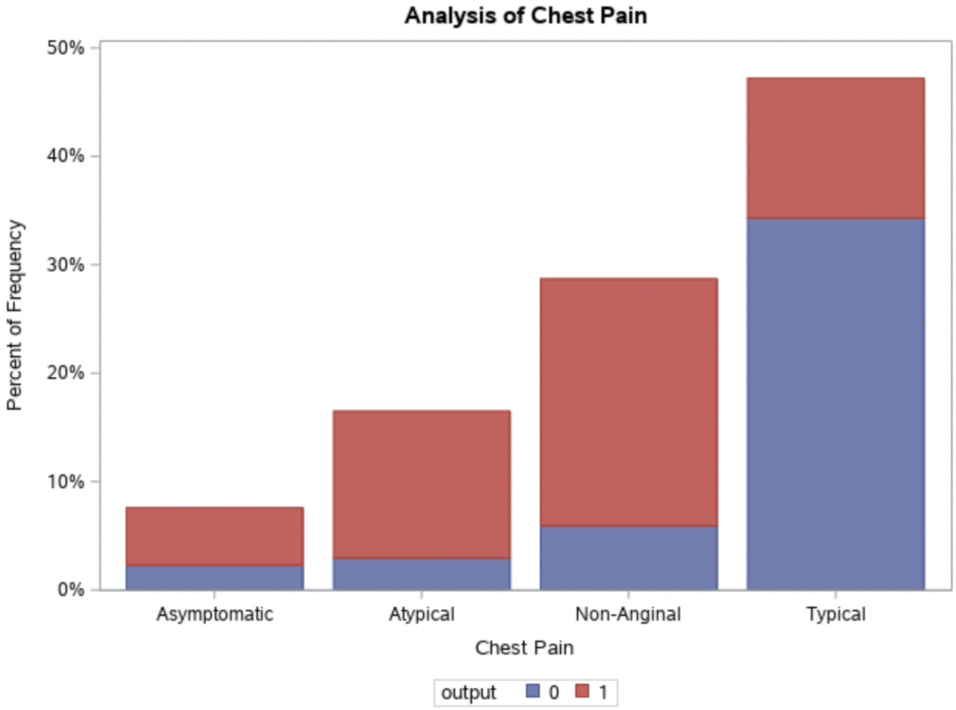


Figure 3.

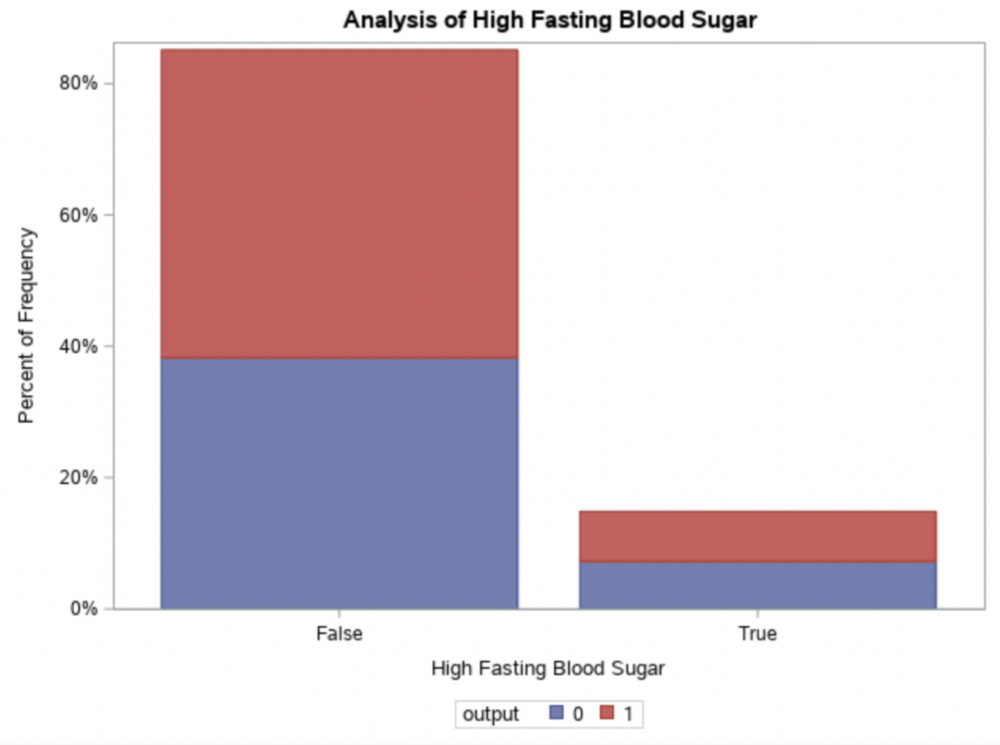
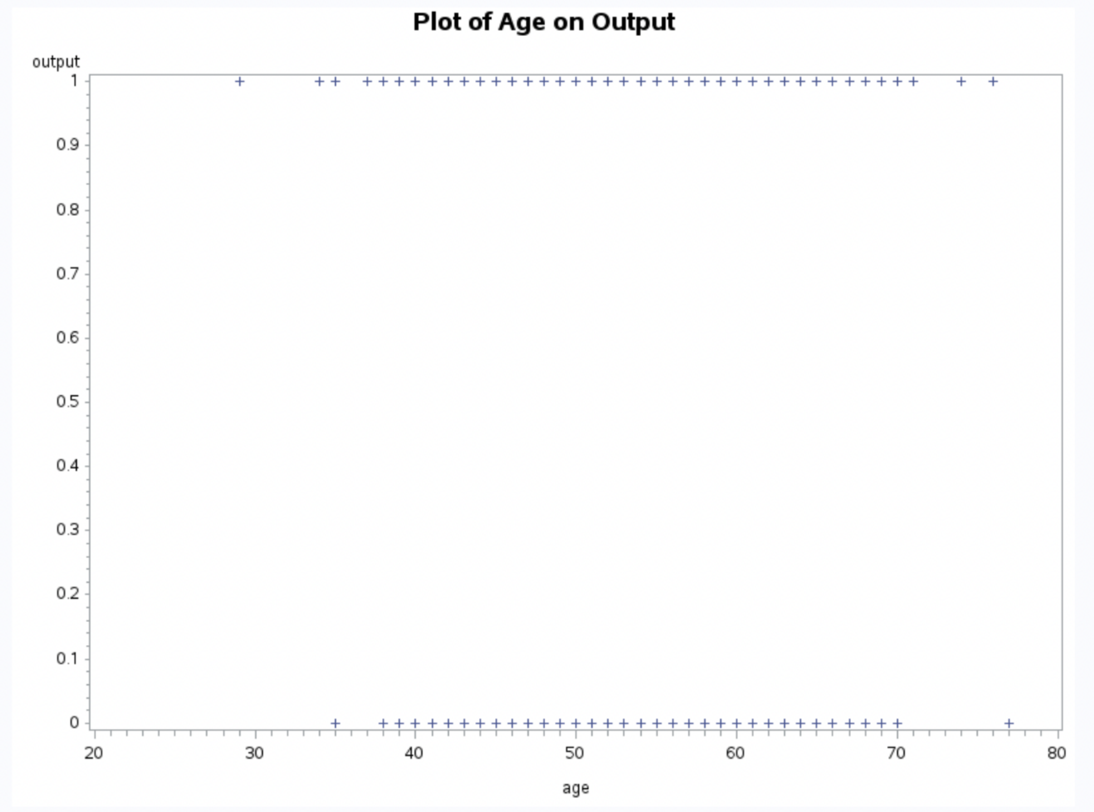
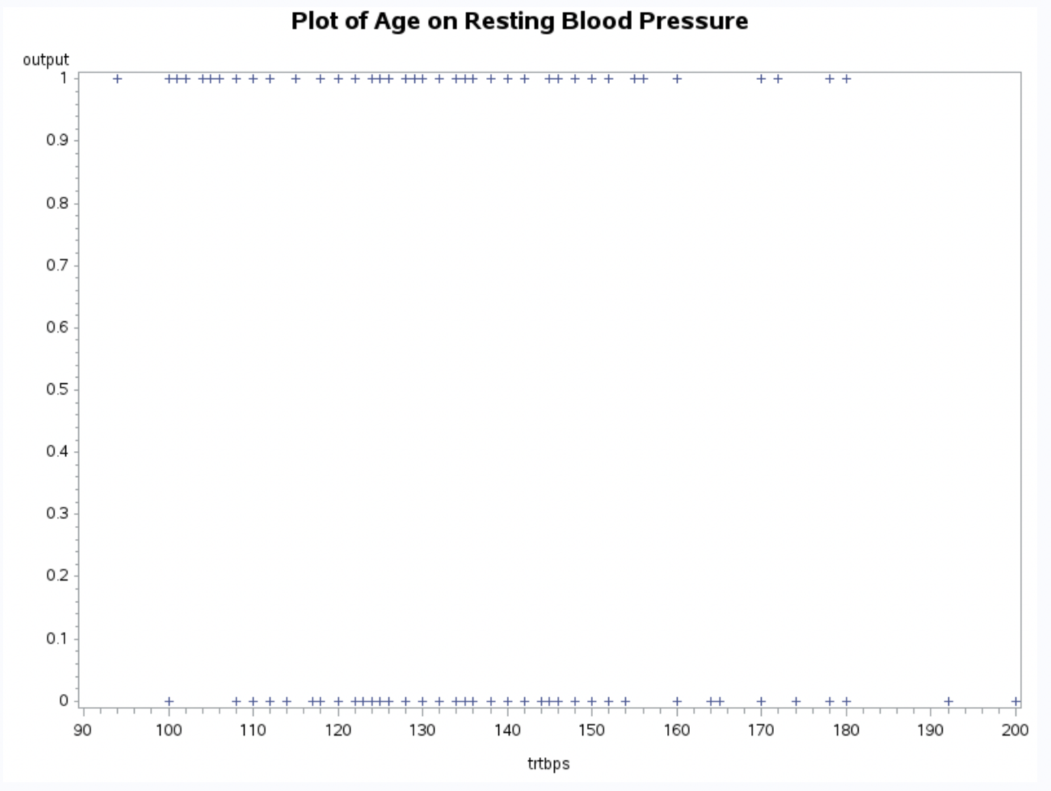
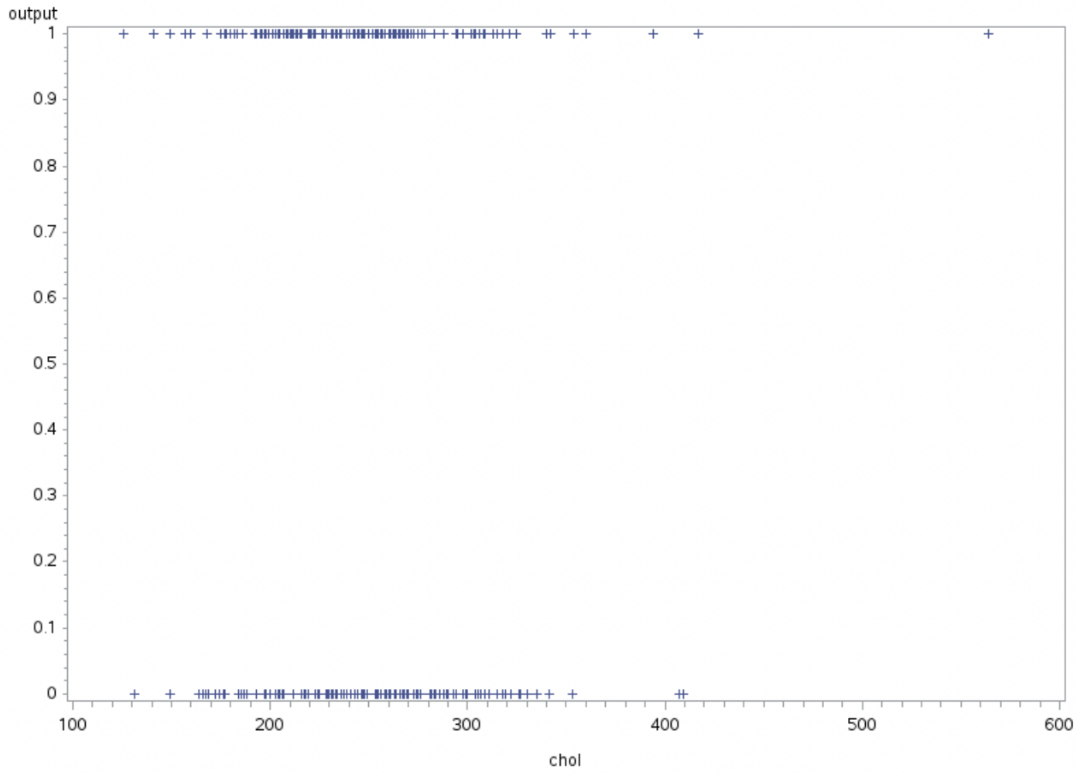

Analysis of High Fasting Blood Sugar
Percent of Frequency
80%
60%
40%
20%
0%
False
True
High Fasting Blood Sugar
output
🟦 0
🟥 1

Figure 4.


Plot of Age on Output
output
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
20
30
40
50
60
70
80
age


Plot of Age on Resting Blood Pressure
output
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
90
100
110
120
130
140
150
160
170
180
190
200
trtbps

Plot of Age on Cholesterol Level



Plot of Age on Maximum Heart Rate Achieved

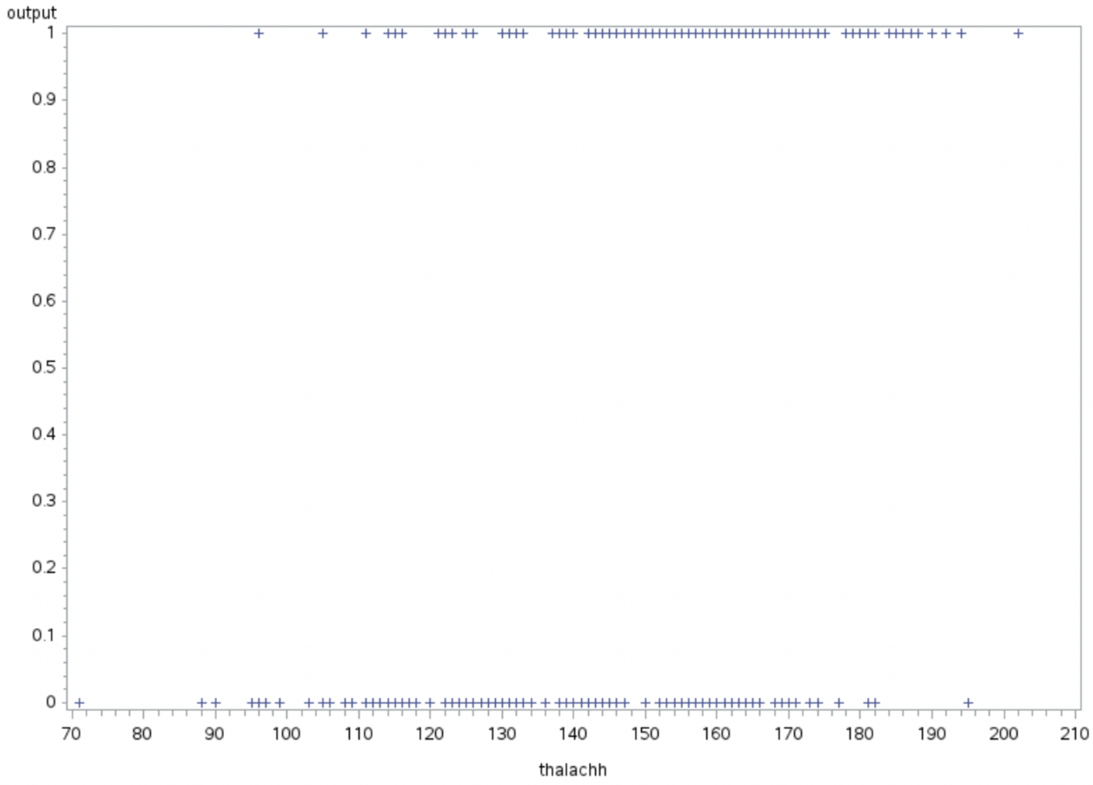
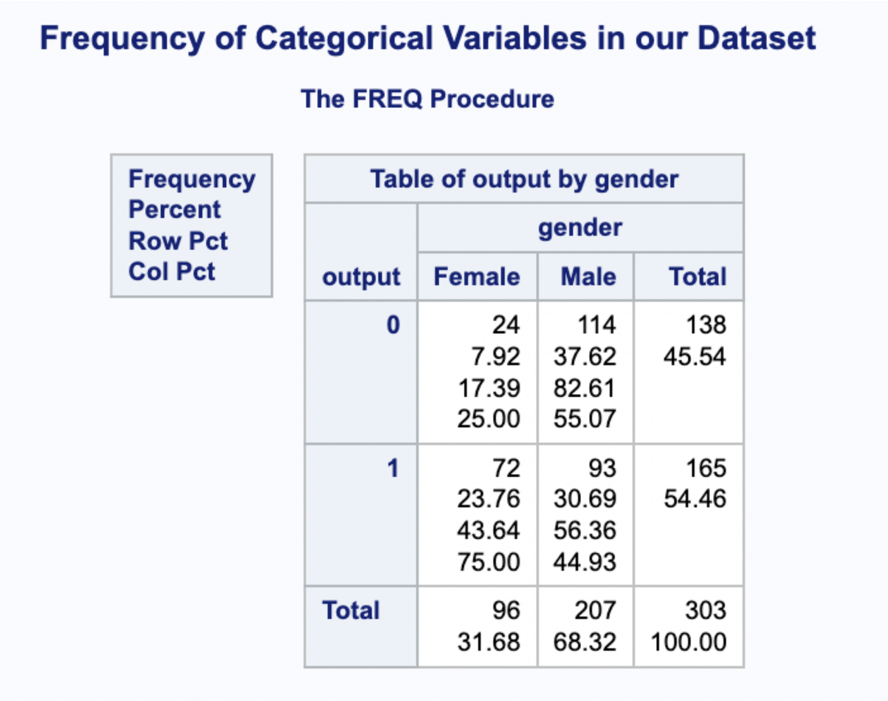


Figure 5.



Frequency of Categorical Variables in our Dataset

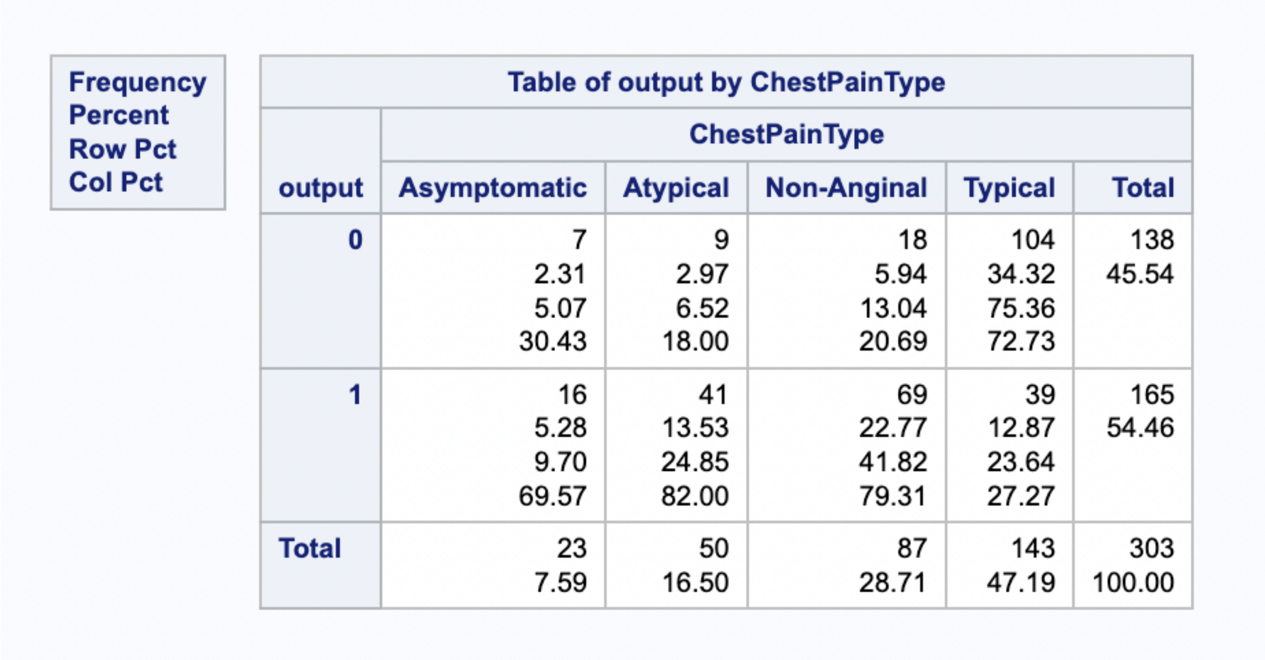
The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of output by gender

	gender		
output	Female	Male	Total
0	24 7.92 17.39 25.00	114 37.62 82.61 55.07	138 45.54
1	72 23.76 43.64 75.00	93 30.69 56.36 44.93	165 54.46
Total	96 31.68	207 68.32	303 100.00

Figure 6.

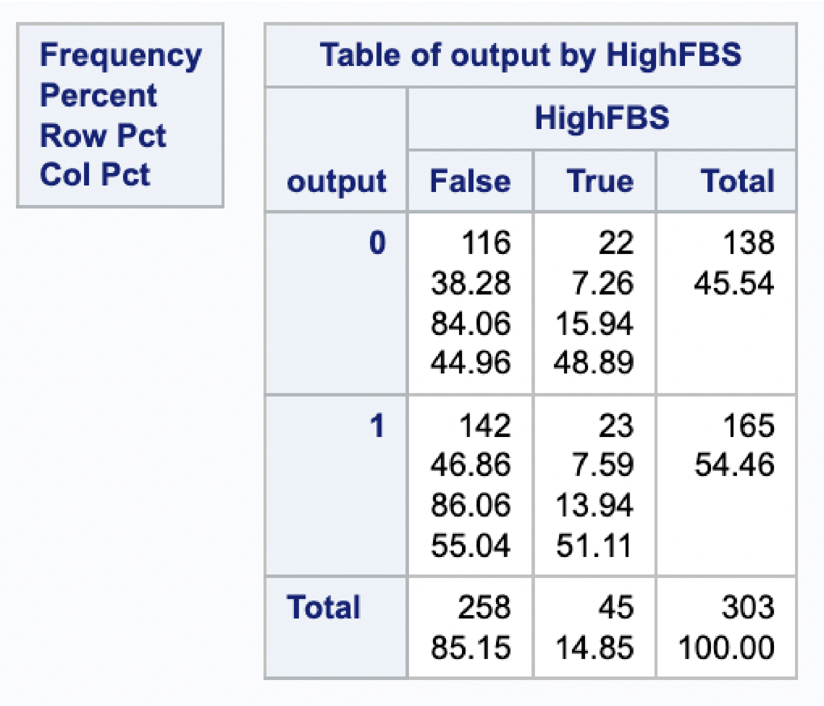


Frequency
Percent
Row Pct
Col Pct

Table of output by ChestPainType

	ChestPainType				
output	Asymptomatic	Atypical	Non-Anginal	Typical	Total
0	7 2.31 5.07 30.43	9 2.97 6.52 18.00	18 5.94 13.04 20.69	104 34.32 75.36 72.73	138 45.54
1	16 5.28 9.70 69.57	41 13.53 24.85 82.00	69 22.77 41.82 79.31	39 12.87 23.64 27.27	165 54.46
Total	23 7.59	50 16.50	87 28.71	143 47.19	303 100.00

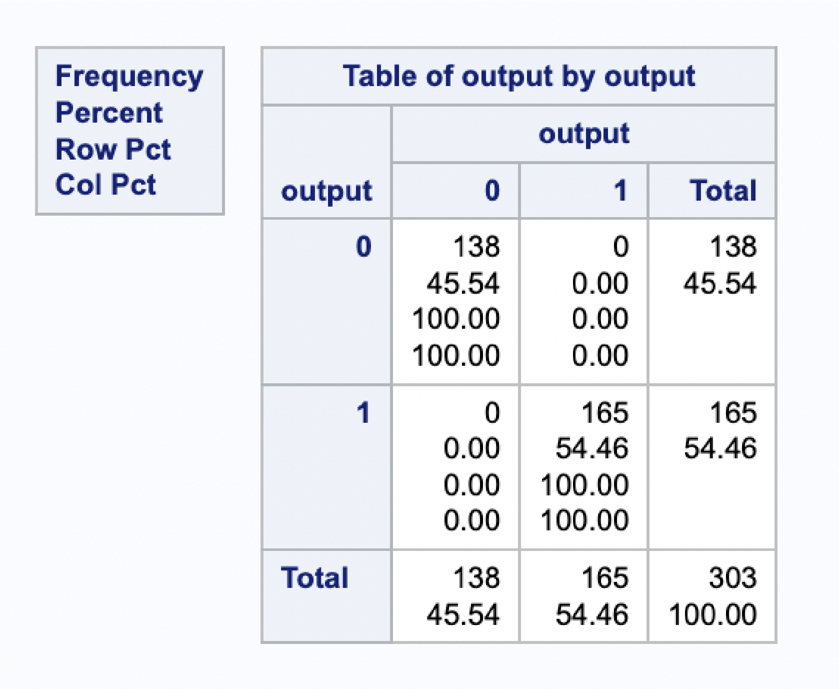
Figure 7.



Frequency
Percent
Row Pct
Col Pct

Table of output by HighFBS			
	HighFBS		
output	False	True	Total
0	116 38.28 84.06 44.96	22 7.26 15.94 48.89	138 45.54
1	142 46.86 86.06 55.04	23 7.59 13.94 51.11	165 54.46
Total	258 85.15	45 14.85	303 100.00

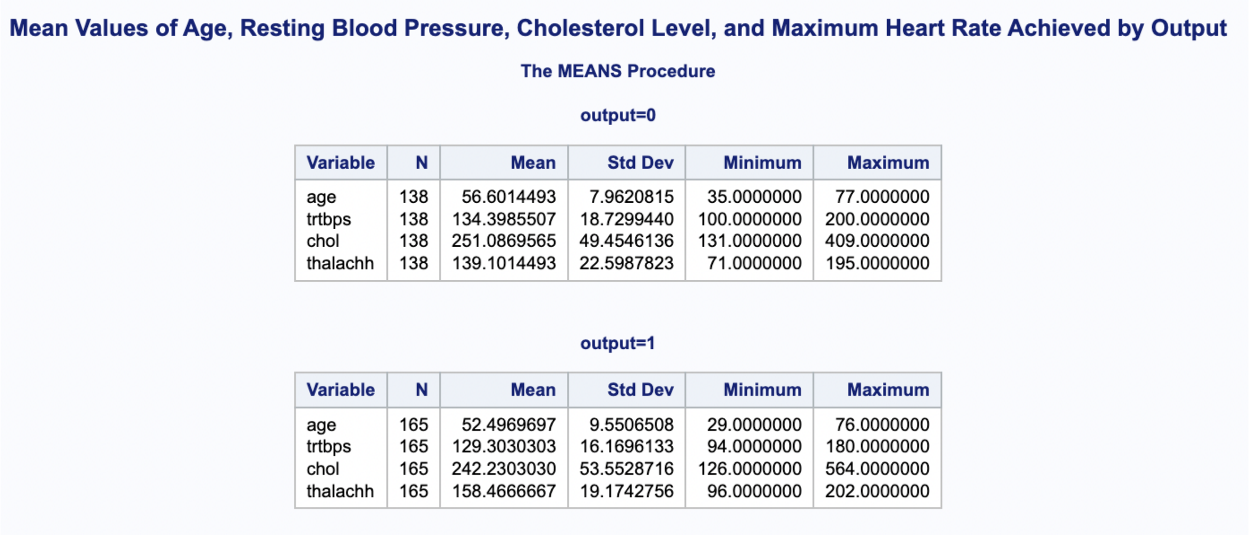
Figure 8.



Frequency
Percent
Row Pct
Col Pct

Table of output by output			
	output		
output	0	1	Total
0	138 45.54 100.00 100.00	0 0.00 0.00 0.00	138 45.54
1	0 0.00 0.00 0.00	165 54.46 100.00 100.00	165 54.46
Total	138 45.54	165 54.46	303 100.00

Figure 9.



Mean Values of Age, Resting Blood Pressure, Cholesterol Level, and Maximum Heart Rate Achieved by Output

The MEANS Procedure

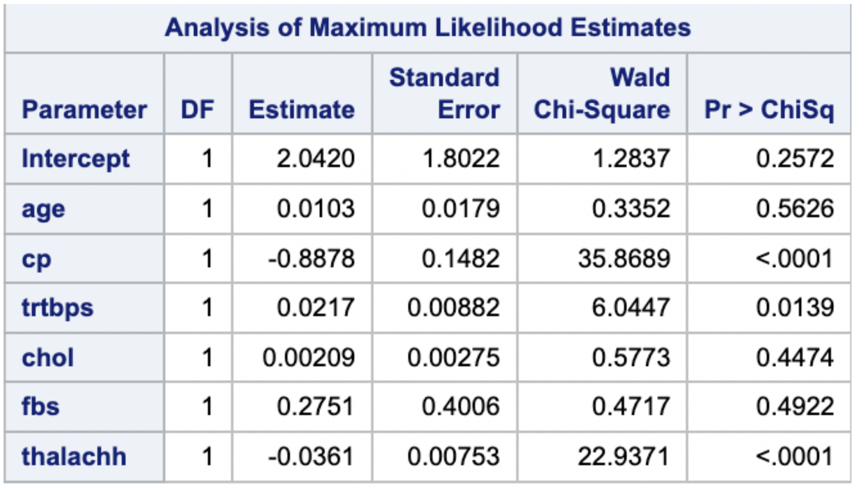
output=0

Variable	N	Mean	Std Dev	Minimum	Maximum
age	138	56.6014493	7.9620815	35.0000000	77.0000000
trtbps	138	134.3985507	18.7299440	100.0000000	200.0000000
chol	138	251.0869565	49.4546136	131.0000000	409.0000000
thalachh	138	139.1014493	22.5987823	71.0000000	195.0000000

output=1

Variable	N	Mean	Std Dev	Minimum	Maximum
age	165	52.4969697	9.5506508	29.0000000	76.0000000
trtbps	165	129.3030303	16.1696133	94.0000000	180.0000000
chol	165	242.2303030	53.5528716	126.0000000	564.0000000
thalachh	165	158.4666667	19.1742756	96.0000000	202.0000000

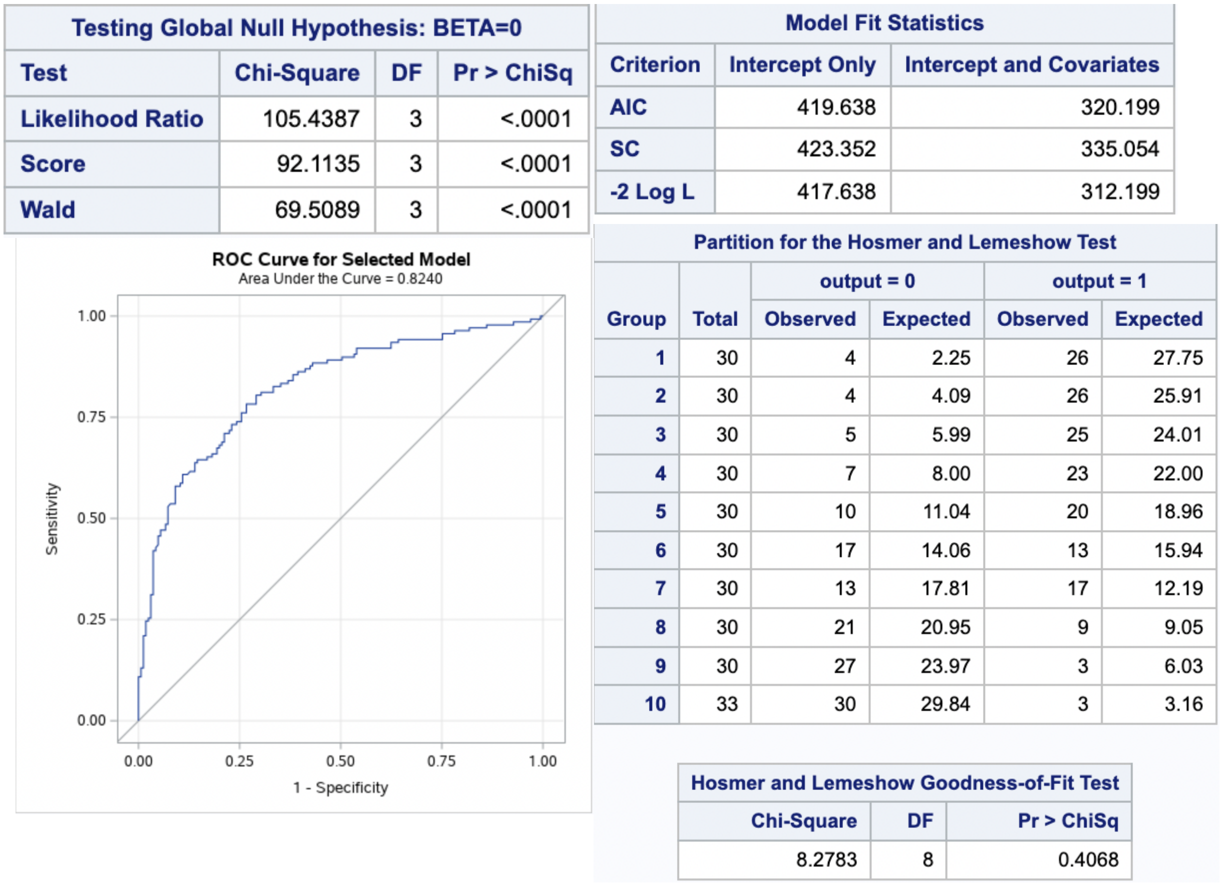
Figure 10.



Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.0420	1.8022	1.2837	0.2572
age	1	0.0103	0.0179	0.3352	0.5626
cp	1	-0.8878	0.1482	35.8689	<.0001
trtbps	1	0.0217	0.00882	6.0447	0.0139
chol	1	0.00209	0.00275	0.5773	0.4474
fbs	1	0.2751	0.4006	0.4717	0.4922
thalachh	1	-0.0361	0.00753	22.9371	<.0001

Figure 11.



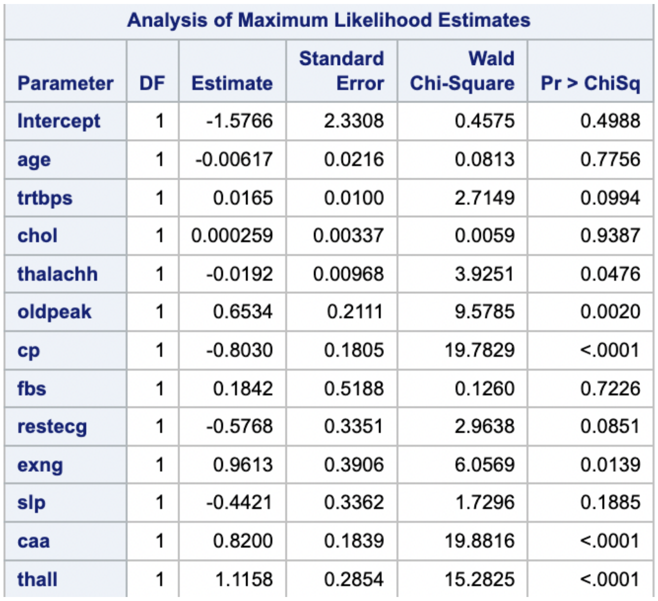
Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	105.4387	3	<.0001
Score	92.1135	3	<.0001
Wald	69.5089	3	<.0001

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	419.638	320.199
SC	423.352	335.054
-2 Log L	417.638	312.199

Partition for the Hosmer and Lemeshow Test					
		output = 0		**output = 1**	
Group	**Total**	**Observed**	**Expected**	**Observed**	**Expected**
1	30	4	2.25	26	27.75
2	30	4	4.09	26	25.91
3	30	5	5.99	25	24.01
4	30	7	8.00	23	22.00
5	30	10	11.04	20	18.96
6	30	17	14.06	13	15.94
7	30	13	17.81	17	12.19
8	30	21	20.95	9	9.05
9	30	27	23.97	3	6.03
10	33	30	29.84	3	3.16

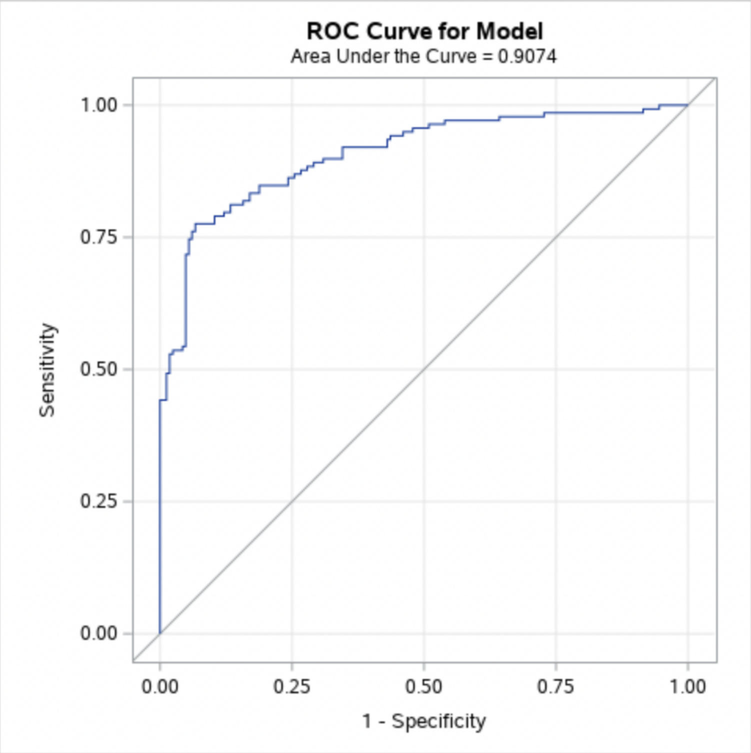
Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
8.2783	8	0.4068

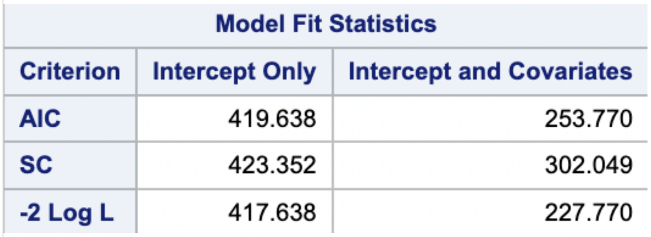
Figure 12.



Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-1.5766	2.3308	0.4575	0.4988
age	1	-0.00617	0.0216	0.0813	0.7756
trtbps	1	0.0165	0.0100	2.7149	0.0994
chol	1	0.000259	0.00337	0.0059	0.9387
thalachh	1	-0.0192	0.00968	3.9251	0.0476
oldpeak	1	0.6534	0.2111	9.5785	0.0020
cp	1	-0.8030	0.1805	19.7829	<.0001
fbs	1	0.1842	0.5188	0.1260	0.7226
restecg	1	-0.5768	0.3351	2.9638	0.0851
exng	1	0.9613	0.3906	6.0569	0.0139
slp	1	-0.4421	0.3362	1.7296	0.1885
caa	1	0.8200	0.1839	19.8816	<.0001
thall	1	1.1158	0.2854	15.2825	<.0001

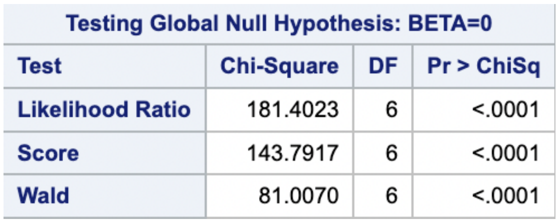
Figure 13.



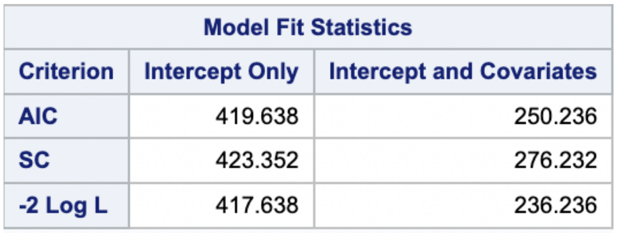


Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	419.638	253.770
SC	423.352	302.049
-2 Log L	417.638	227.770

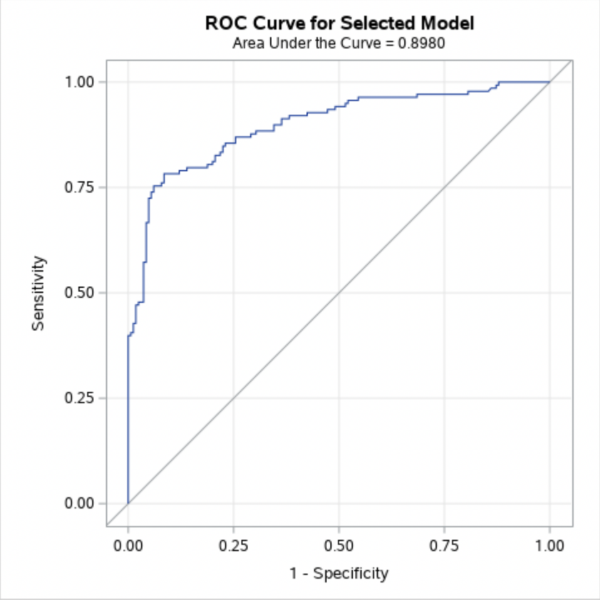
Figure 14.

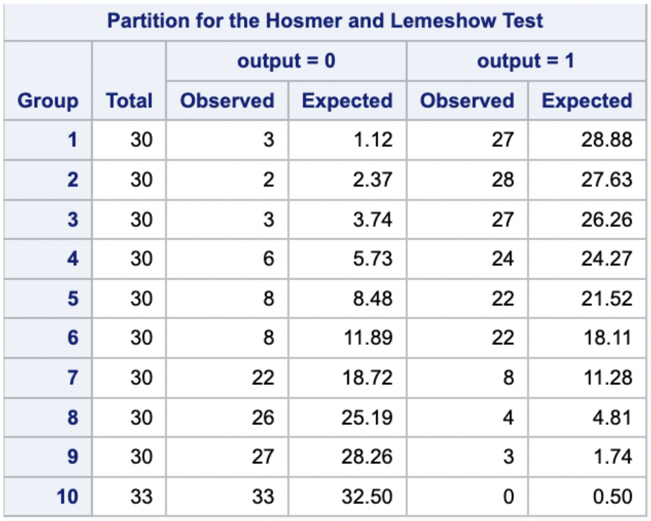


Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	181.4023	6	<.0001
Score	143.7917	6	<.0001
Wald	81.0070	6	<.0001



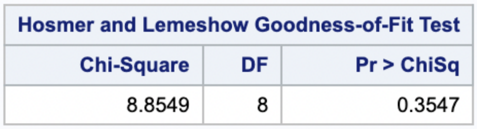
Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	419.638	250.236
SC	423.352	276.232
-2 Log L	417.638	236.236





Partition for the Hosmer and Lemeshow Test

		output = 0		output = 1	
Group	Total	Observed	Expected	Observed	Expected
1	30	3	1.12	27	28.88
2	30	2	2.37	28	27.63
3	30	3	3.74	27	26.26
4	30	6	5.73	24	24.27
5	30	8	8.48	22	21.52
6	30	8	11.89	22	18.11
7	30	22	18.72	8	11.28
8	30	26	25.19	4	4.81
9	30	27	28.26	3	1.74
10	33	33	32.50	0	0.50



Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
8.8549	8	0.3547

5. Resources

“Heart Disease.” *Mayo Clinic*, Mayo Foundation for Medical Education and Research, 9 Feb. 2021, www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/.

“Know Your Risk for Heart Disease.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 9 Dec. 2019, www.cdc.gov/heartdisease/risk_factors.htm.