Heart Disease Diagnosis Analysis

By Melchor Ronquillo, Shrdha Shrestha, and Sophie Srisak

Introduction

Purpose

For our project we decided to analyze factors that could potentially predict the possibility of having a diagnosis of heart disease.

Our dataset consists of 14 variables from 303 individuals. The variables given are both categorical and numerical. They include variables that are easily accessible to the normal everyday person, like one's blood sugar, and also variables that require specialized medical equipment and procedures, like an ECG and Fluoroscopy through X-ray.

Our goal is to provide knowledge and resources to those who cannot easily access medical professionals and guide them to understanding their potential risk of heart disease through the utilization of the accessible variables in this dataset and performing an analysis on them. Through that they can assess their potential risk and seek help from a medical professional, if needed.

Background On Heart Disease

Risk Factors

- Certain risk factors can contribute to fatty plaque buildup inside of narrow arteries and can lead to the risk of a heart attack, angina, or stroke
- Risk factors include: older age, high blood pressure, high blood cholesterol, obesity, and lack of physical activity
- Shortness of breath, chest pain, and racing heartbeat are some symptoms of heart disease

Variables

Categorical Variables

- sex **
  - (1 = male, 0 = female)
- cp - chest pain type **
  - (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 0 = asymptomatic)
- fbs - fasting blood sugar > 120 mg/dl
  - (1 = true, 0 = false)
- restecg - resting electrocardiographic results
  - (1 = normal, 2 = having ST-T wave abnormality, 0 = hypertrophy)
- exng - exercise induced angina
  - (1 = yes, 0 = no)
- slp - the slope of the peak exercise ST segment
  - (2 = upsloping, 1 = flat, 0 = downsloping)
- thall - thall rate
  - (2 = normal, 1 = fixed defect, 3 = reversible defect)
- output - the predicted attribute - diagnosis of heart disease (angiographic disease status) **
  - (Value 0 = < 50% diameter narrowing, Value 1 = > 50% diameter narrowing)

Variables

Numerical Variables

- age - age in years
- trtbps - resting blood pressure **
  - (in mm Hg on admission to the hospital)
- chol - serum cholesterol **
  - (in mg/dl)
- thalachh - maximum heart rate achieved **
- oldpeak - ST depression induced by exercise relative to rest
- caa - number of major vessels (0-3) colored by fluoroscopy

Analysis primarily focuses on seven of the 14 variables: age, sex, type of chest pain, resting blood pressure, cholesterol levels, fasting blood sugar exceeding 120 mg/dl, and maximum heart rate achieved.

We thought these are these variables were more easily accessible for the regular person who does not work in the medical field and based on our outside research, these were the reasons most listed on articles such as the CDC article referenced in our paper.

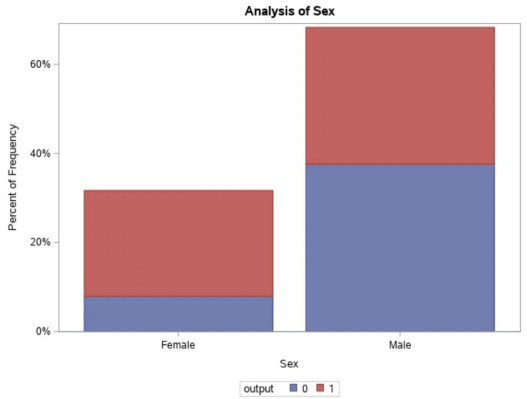
Hypothesis Test Statement

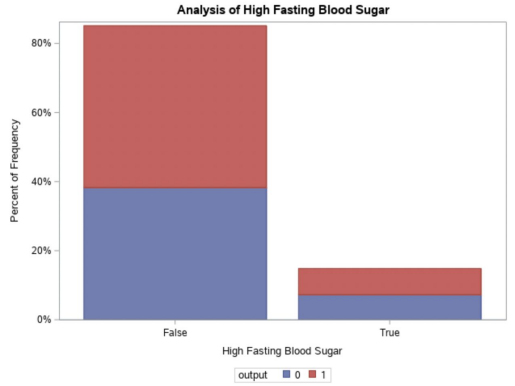
Cholesterol level and chest pain type are the most important and significant variables in predicting the chance of heart disease out of the seven variables that we thought were most easily accessible.

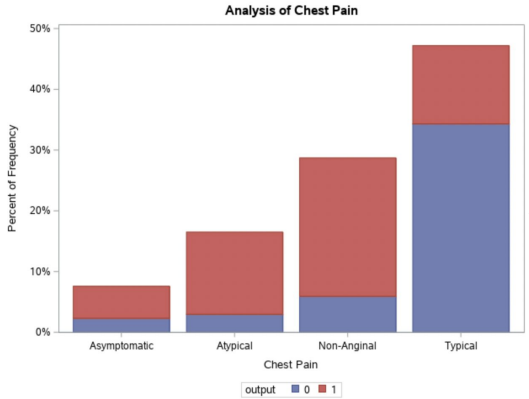
Statistical Test

First Analysis: Vertical Bar Graph

Second Analysis: Frequency Tables







Analyzing Variables

Categorical Variables: Frequency Tables

- **207** males and **96** females in the study
- For high chance of heart disease :
  - **56.36%** male and **43.64%** female
  - **41.82%** had non-anginal chest pain, **24.85%** had atypical, **23.64%** had typical, and **9.7%** had asymptomatic chest pain
  - **86.06%** did **not** have a high fasting blood sugar
- **138** individuals had a low chance of having heart disease (45.54%) and **165** individuals had a high chance of having heart disease(54.46%)

Numerical Variables: Proc Means

Low Chance of Heart Disease:

Mean Age:56.60

Mean Resting BP: 134.3985507 mm Hg

Mean Cholesterol :251.0869565 mg/dl,

Mean Max. Heart Rate: 139.1014493 bpm

High Chance of Heart Disease:

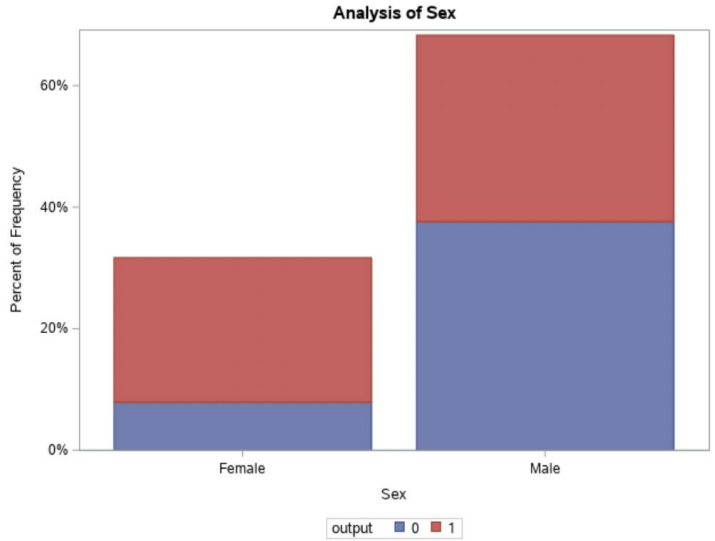
Mean Age: 52.49

Mean Resting BP: 129.3030303mm Hg

Mean Cholesterol : 242.2303030mg/dl,

Mean Max. Heart Rate: 158.4666667 bpm

Conclusion from First Analysis



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

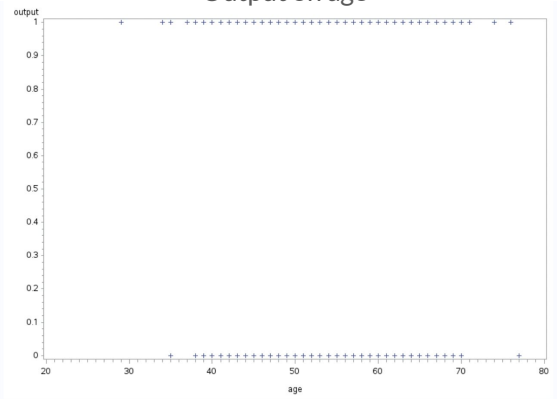
We found that the sex variable was not equal in occurances and thought that it may skew the analysis.

Statistical Test

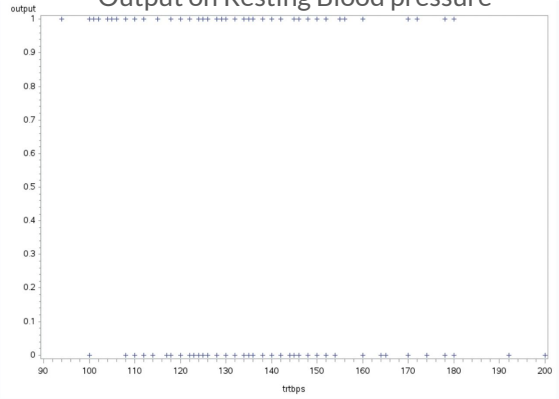
Third Analysis: Scatterplot

- Plot the numerical variables on output
- Looking for complete separation

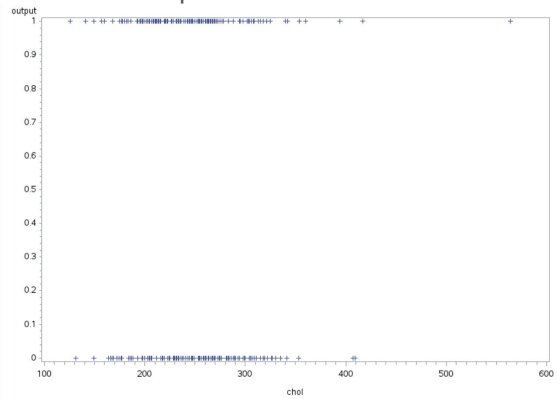
Output on age



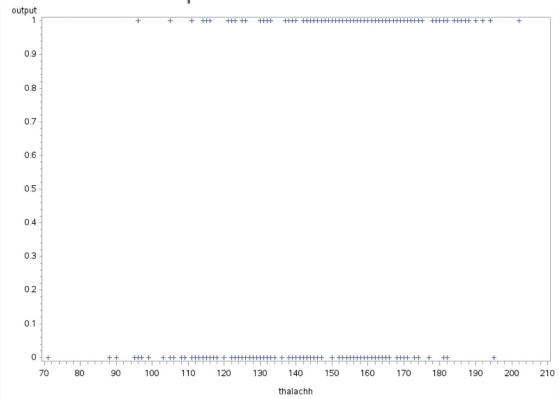
Output on Resting Blood pressure



Output on Cholesterol level



Output on Max Heart Rate

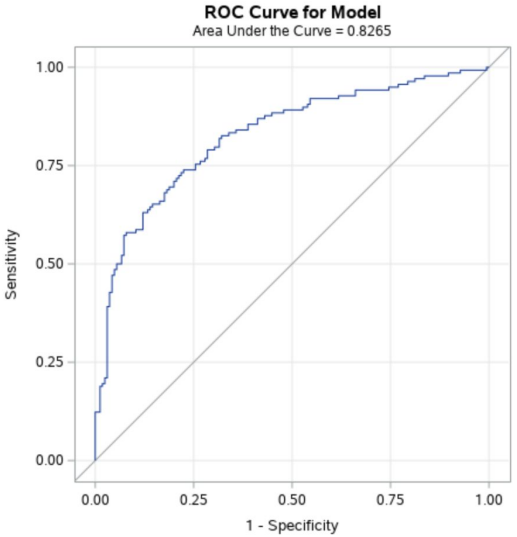


Statistical Test

Fourth Analysis: First Logistic Model

- Ran a logistic model with all of our variables of interest
- AUC = 0.8265
- AIC = 324.534
- Some variables not statistically significant

```
proc logistic data = heart plots(only)=roc;
model output = age  cp trtbps chol fbs thalachh;
run;
```



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	419.638	324.534
SC	423.352	350.530
-2 Log L	417.638	310.534

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

Conclusion from First Logistic Model

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

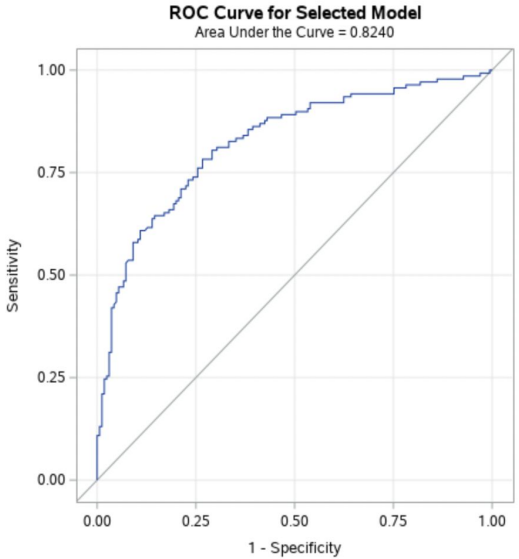
Removed age, chol, and fbs because the p-valued were not statistically significant at alpha = 0.05.

Statistical Test

Fifth Analysis: First Reduced Logistic Model

- Logistic regression of significant variables
- AUC = 0.8240
- AIC = 320.199
- Both improved a little bit

```
proc logistic data = heart plots(only)=roc;
model output = cp trtbps thalachh / selection = stepwise Risklimits lackfit ctable;
run;
```



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.9928	1.4565	4.2220	0.0399
cp	1	-0.8796	0.1462	36.2050	<.0001
trtbps	1	0.0246	0.00839	8.5869	0.0034
thalachh	1	-0.0375	0.00696	28.9875	<.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

Statistical Test

```
proc logistic data = heart plots(only)=roc;
model output = cp trtbps thalachh / selection = stepwise Risklimits lackfit ctable;
run;
```

Fifth Analysis: First Reduced Logistic Model

- Likelihood ratio = <.0001
  - Overall model is significant
- Goodness-of-Fit-Test = 0.4068
  - No statistical difference between observed and expected

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

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

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

Statistical Test

Sixth Analysis: Second Logistic Model

- Variables not significant
  - Age
  - Resting Blood Pressure
  - Cholesterol
  - Fasting Blood Sugar
  - Resting ECG

```
proc logistic data = heart plots(only)=roc;
model output = age trtbps chol thalachh oldpeak cp fbs restecg exng slp caa thall;
run;
```

Analysis of Maximum Likelihood Estimates

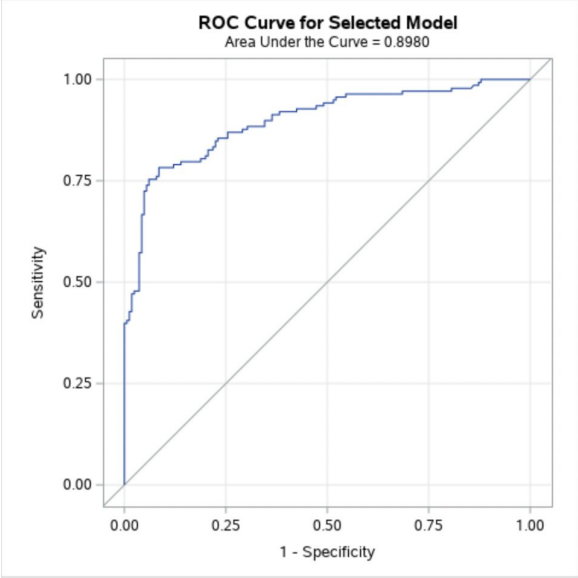
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.5766	2.3308	0.4575	0.4988
age	1	-0.00617	0.0216	0.0813	0.7756
trtbps	1	0.0165	0.0100	2.7149	0.0994
chol	1	0.000259	0.00337	0.0059	0.9387
thalachh	1	-0.0192	0.00968	3.9251	0.0476
oldpeak	1	0.6534	0.2111	9.5785	0.0020
cp	1	-0.8030	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

Statistical Test

Seventh Analysis: Second Reduced Logistic Model

- Logistic regression of significant variables
- AUC = 0.8980
- AIC = 250.236
- Drastic improvements compared to our previous reduced model

```
proc logistic data = heart plots(only)=roc;
model output = thalachh oldpeak cp exng caa thall / selection = stepwise Risklimits lackfit ctable;
run;
```



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.5341	1.3936	0.1469	0.7015
thalachh	1	-0.0197	0.00822	5.7474	0.0165
oldpeak	1	0.7844	0.1844	18.0922	<.0001
cp	1	-0.7402	0.1712	18.6937	<.0001
exng	1	1.0604	0.3767	7.9235	0.0049
caa	1	0.7744	0.1728	20.0849	<.0001
thall	1	1.0427	0.2739	14.4945	0.0001

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

Statistical Test

Seventh Analysis: Second Reduced Model

- Likelihood ratio = <.0001
  - Model is significant
- Goodness-of-Fit-Test = 0.4068
  - No statistical difference between observed and expected

```
proc logistic data = heart plots(only)=roc;
model output = thalachh oldpeak cp exng caa thall / selection = stepwise Risklimits lackfit ctable;
run;
```

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

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
8.8549	8	0.3547

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

Conclusion

From our regression analysis of the 13 variables we found that **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 hypothesis that cholesterol level and chest pain type were the best predicting variables was **incorrect**.

Possible Error: excluding sex variable from regression model and not including all variables in our original logistic regression model

Confounding: medication taken by participants, pre-existing conditions/comorbidities

Importance

- Help medical professionals focus on the most important factors (ex: exercise induced angina, number of major blood vessels, and old peak) since these are the strongest predictors
- We can extend research by repeating study with a new cohort and see if we get the same results