ANALYSIS OF CLEVELAND DATASET

ACHAL SURESH, NIDA KAZI, SERGEY POSTOLSKY, YAOSONG YU

Section T2

Data exploration:

On the outset, we were intrigued by the way the data was structured. For the NUM column, we had values 0, 1, 2, 3, 4 that represent the 2 different levels of diagnosing the heart disease (if num =0 then no heart disease and if num=1, 2, 3, 4 then num = 1, higher chances of having a heart disease) Based on initial data exploration, we wanted to figure out if we should combine, or separate the data based on the NUM level for diagnosis. We were also intrigued by the relationships affecting thalassemia, and the disorder being the basis for the other disorders. The data also seemed to have features, that intuitively we felt aren't actually the best predictors for our models.

Our initial explorations were manual, as we scoured through the data and features. Later we used R for visualizations hoping to find something different in the dataset.

Starting with this thought in mind, we decided to run LASSO for the best subset selection by splicing the data set. The first time we did this, we found the accuracy for ridge regression was below 70%. To counter this, we decided to run LASSO for the different NUM levels. By doing this, our accuracy reached above 90%.

Using Exploratory Data Analysis, we were able to figure out how many patients have the disease depending on the age factor. We were able to notice the frequency of heart disease depending on the chest pain type. We also carried out some basic data preprocessing by removing missing and omitted values. Figures 1-7 in the Appendix that also represent our observations during the data exploration.

Hypothesis:

Looking at the data, here are our four most intuitive hypotheses.

1. Different nums are affected by different parameters. Cholesterol, Age, Sex and Chest pain affect the NUM the most significantly.

   Based on research and study of the dataset we identified that variable "num" can include values from 0 til4, where 0 – no heart disease and 4 – the most serious heart disease. Hypothesis number one is that different factors can differently influence various levels of severity of heart diseases. However, we think that the following factors are the most significant in the prediction of heart disease: cholesterol level, chest pain, age, and sex.

3. Older men, with High Cholesterol and High BP, are more prone to heart disease.

After an initial look at the data, we could gauge men are more prone to heart disease than women, and intuitively, we wanted to test if older men with preconditions like high BP or High Cholesterol are more prone to having a heart disease.

4. Resting blood pressure (trestbps) and maximum heart rate achieved (thalach) can indicate if a patient has exercise induced angina (exang).

For this hypothesis, we are interested in exploring the correlations of a patient's blood pressure, heart rate, and the condition of a patient's pain. We raised this question by our common knowledge that blood pressure and heart rate can be strong indicators of a person's health condition. Therefore, we would like to test this hypothesis.

Methodologies:

We explored the data for proportions, of the total population with and without heart disease, with and without high cholesterol. It also looked into a population that has high BP and diagnosis for heart disease. We also thought the number of men and women having high cholesterol, and heart disease based on the sex of the person.

For the data exploration, we used LASSO, and the best subset selection to figure out the factors that most affect the accurate prediction of NUM. We tried this in two different ways, first by splitting the data based on num value 0 or 1, and the second time lasso based on the 4 different levels of heart disease.

For Hypothesis 1:

Since hypothesis 1 maps a linear relationship between the response variable NUM and the predictors' Age, Sex, Cholesterol, and Chest Pain, we decided to run a linear regression on these factors.

We divided the variable "num" into 4 variables based on the level of severity of the disease: num1, num2, num3 and num4. We divided the dataset into training (70%) and testing (30%). For each of the "nums" we run the Lasso algorithm on training data to identify the value of the best lambda. Using the best value of lambda, we run the logistic regression on training data to identify the most significant factors and their coefficients that influence the presence and severity of heart disease. In order to assess the prediction accuracy of the logistic regression models for each "num" (num1, num2, num3, num4) we compared models output to the test part of the dataset. Please see the results in the table xx in the Appendix:

For Hypothesis 2:

For this hypothesis, we first started off by converting the value of num to indicate the presence of a heart condition (1) and the absence of any heart condition (0). We followed this by performing LASSO on the dataset to determine the best subset of predictors to include in the clustering. We then performed hierarchical clustering, using the complete method. We also performed K means clustering on the dataset, however observed that the accuracy of the clustering was only 20.8% and hence dropped this method.

For Hypothesis 3:

To test our hypothesis, we fit a logistic regression on trestbps and thalach regarding exang to test if the two predictors are of statistical significance. If we find any predictor with no or minor significance, we will drop the predictors until the predictors are all significant.

As we see in Appendix (Exhibit 3a), we figured that trestbps is not significant. We therefore dropped it and ran another logistic regression solely on thalach, shown in Appendix (Exhibit 3b).

Conclusions:

#1: Different factors influence heart disease prediction differently based on disease severity; Surprisingly as it can be seen from the table, neither age nor sex nor cholesterol turned out to be significant factors in prediction of all heart diseases' severity levels. Please see Exhibit 1

Additional observations and insights:

The bigger your maximum heart rate achieved (Thalach) the less likely you have a heart disease;
The following factors can be the best predictors of having heart disease: cp - chest pain (especially type 4), ca - number of major vessels (0-3) colored by fluoroscopy and thal – thalassemia (especially type 7).

#2: Since the number of observations for each cluster was not significantly different it is difficult to make concrete conclusions on the data. However, some of the observations we see are:
The age of person increases from 44(Cluster 2) to 57(Cluster 1) and also probability of heart disease can be more for men.
The Cholesterol levels in the cluster with more heart diseases are significantly more apart from the safe cholesterol levels
The Blood Pressure levels in both the clusters are appearing to be closer than expected. Please see Exhibit 2

Additional observations and insights:

The chances of having high blood sugar level can also increase with presence of heart disease.
People with reversible Thalassemia defect also have a higher chance of heart disease.

#3: After dropping trestbps as the predictor, we achieved 60% accuracy in prediction. We therefore conclude that resting blood pressure (trestbps) is not an indicator of exang as tested in the logistic regression, whereas maximum heart rate achieved (thalach) is a crucial indicator of exang.

The result is somewhat surprising for us as the model shows that resting blood pressures is not significant at all in indicating a person's health condition. On the contrary, maximum heart rate achieved shows meets our expectation in predicting the patient's condition. Even though the accuracy is only 60%, we also take in mind that the data set only contains a few hundreds data vectors. With this volume of data it is hard to tell what the real cause of this low accuracy is.

In the future, we are interested to test more about the significance of maximum heart rate achieved as an indicator of a person's health condition. Please see exhibit 3.

Appendix:

Data Exploration:

- Number of people with the heart disease: 137
- Number of people without the heart disease: 159
- Number of Men in the dataset compared with presence of heart condition: 112/200
- Number of women in the dataset with presence of heart condition: 25/96

Age	57	58	59	60
Proportion of cases	61.53%	83.33%	61.53%	100%

Fig 1: Proportions of disease prevalent in men

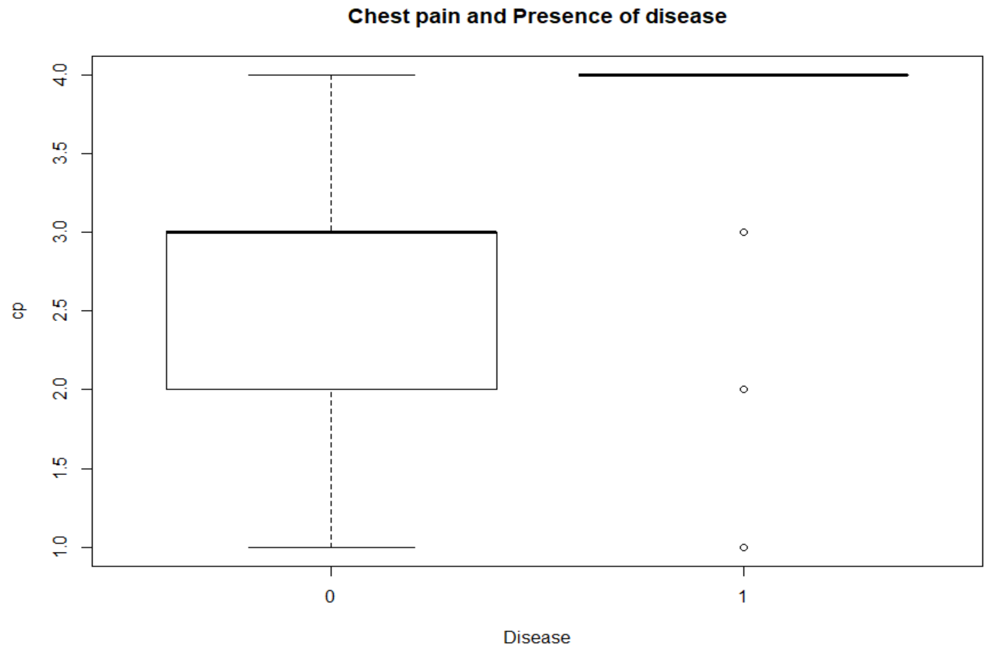


Fig 2: CP wrt prediction of Disease

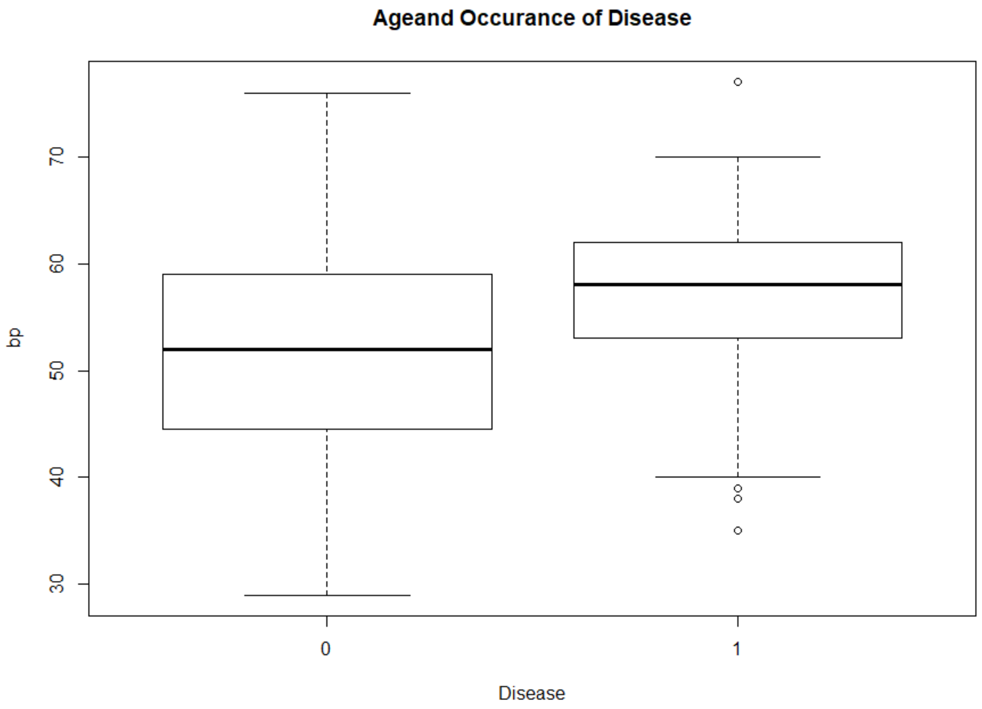


Fig 3: Age wrt to heart disease

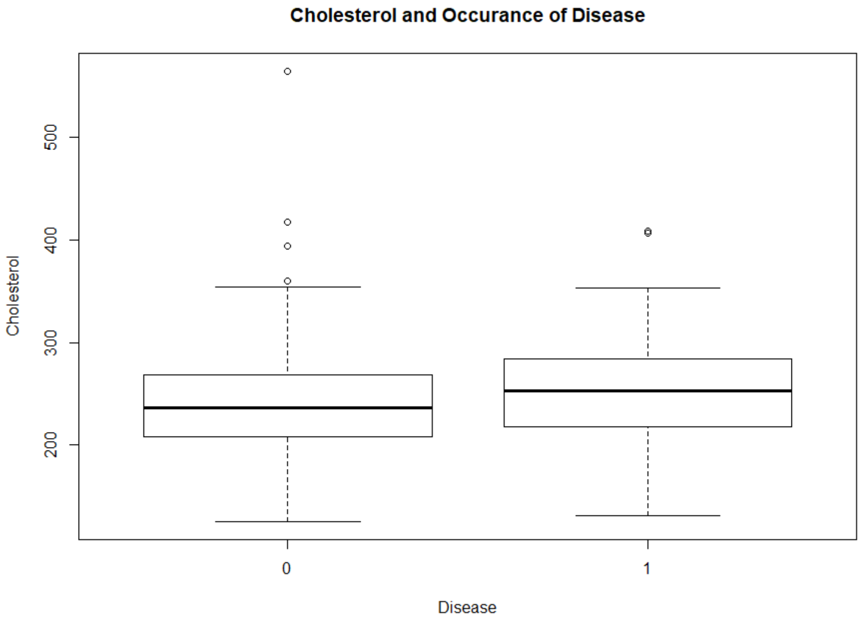


Fig 4: Cholesterol wrt disease

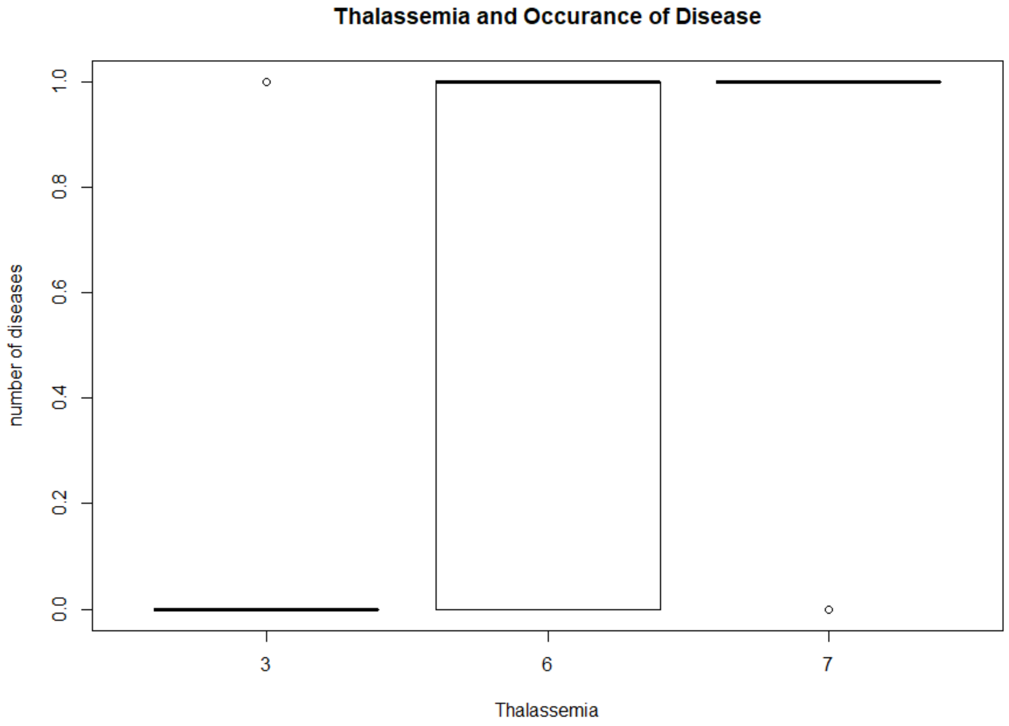


Fig 5: Thalassemia and occurance of disease

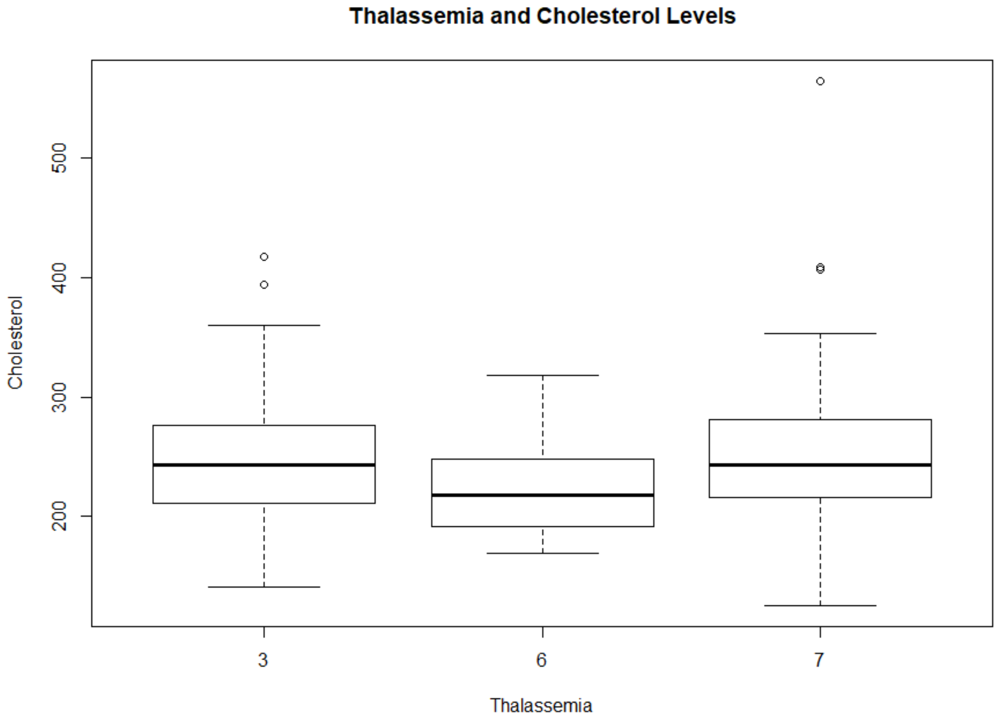


Fig 6: Distribution of thalassemia on cholesterol

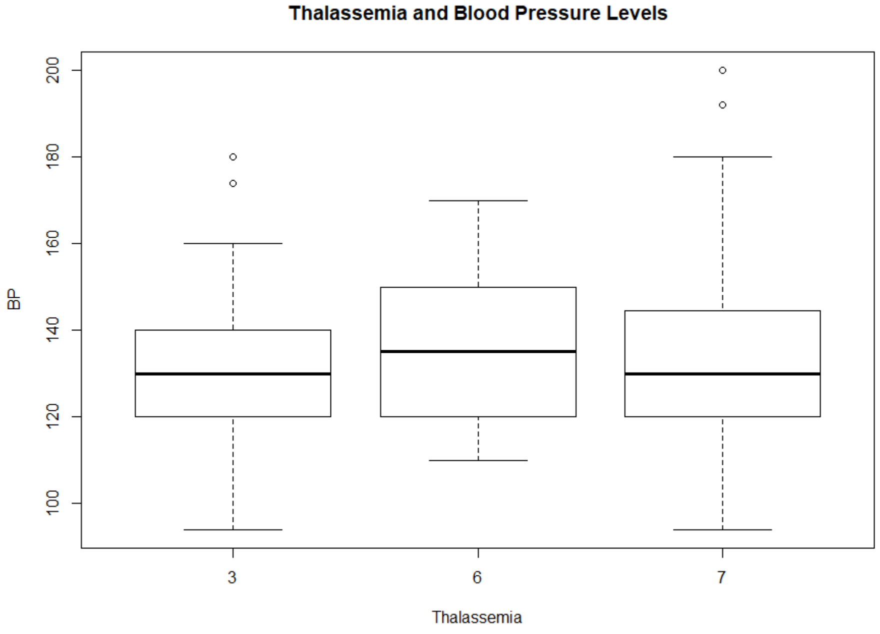


Fig 7: Distribution of thalassemia on BP

Methods:

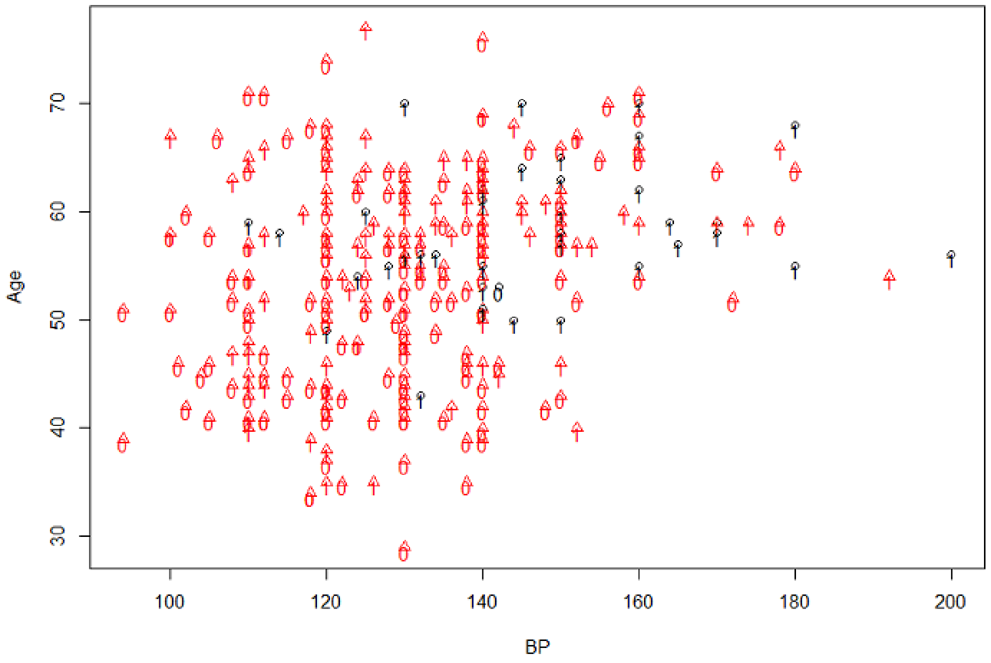


Fig 8: Cluster 1: Clustering on Age and BP

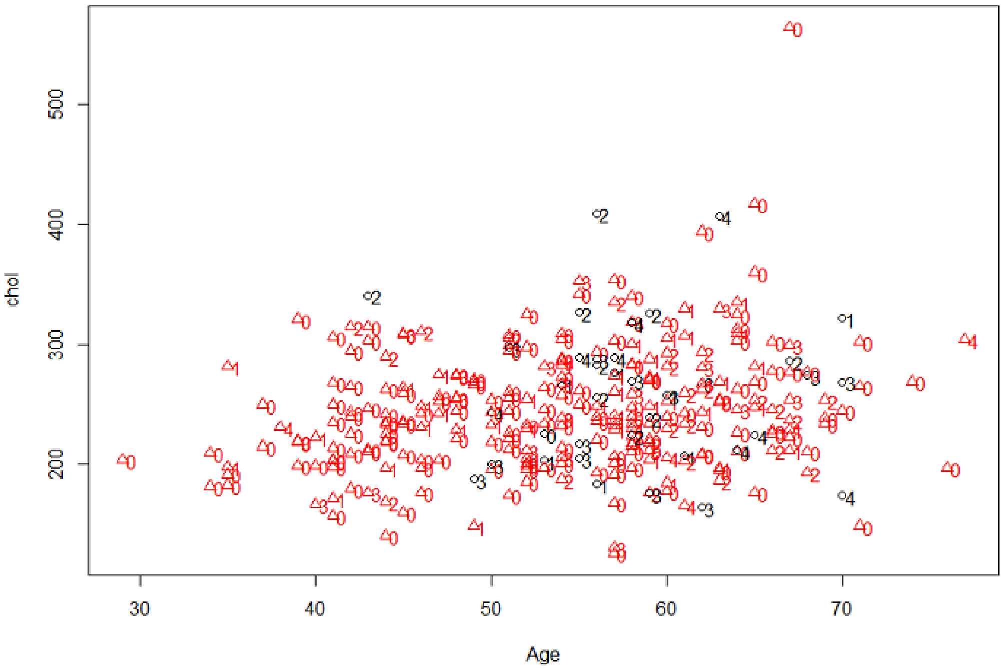
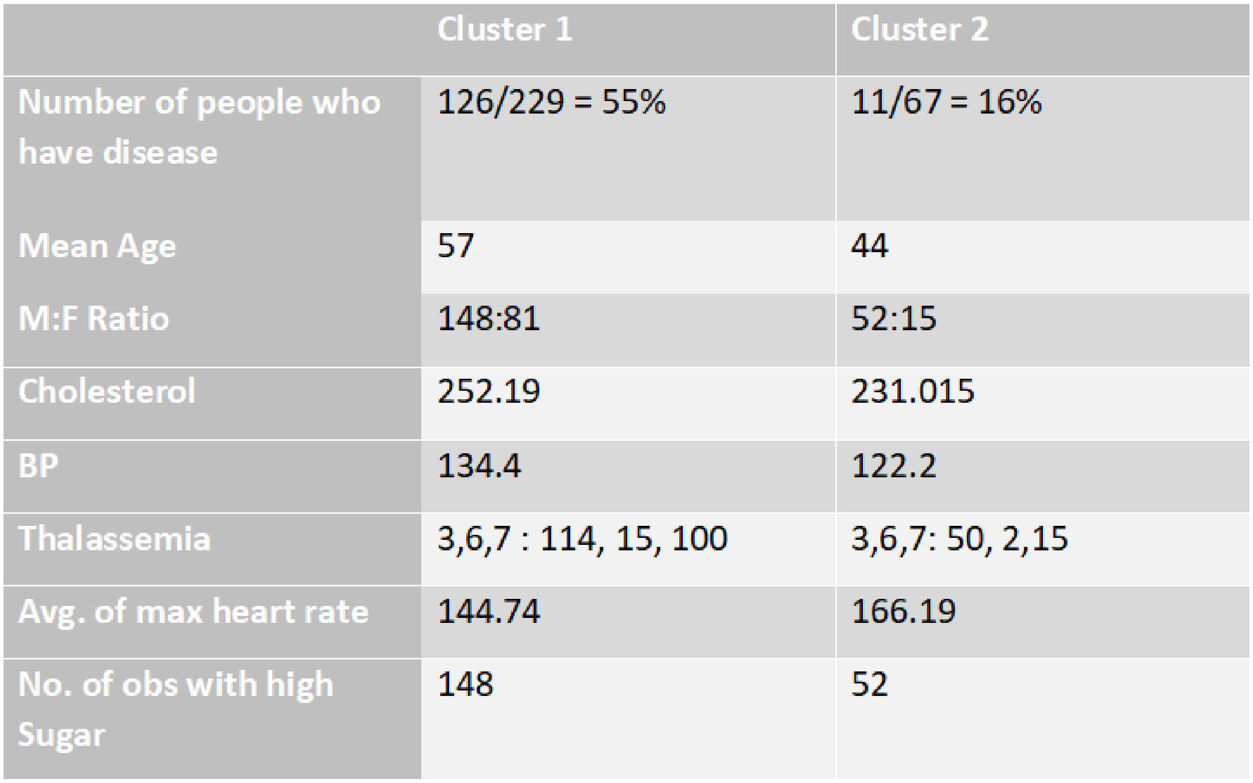


Fig 9: Clustering on Age and cholesterol levels

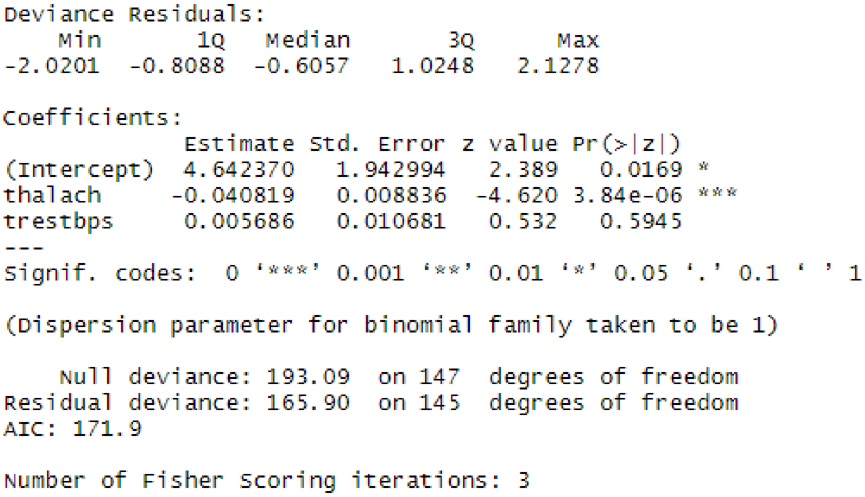
Factors	Num1	Num2	Num3	Num4	NUM
Intercept	-1.71912164	-1.842629990	-1.720616581	-3.2956860	-3.407818623
Age	-	0.003194738	-	-	0.010005656
Sex1	-	-	-	-	0.443771686
Cp2	-	-	-	-	-
Cp3	-0.64376967	-	-	-	-1.359711226
Cp4	0.08783884	0.954952300	0.370854784	-	1.259841472
Trestbps	-	-	-	-	0.010538477
chol	-	0.001543935	-	-	-
Fbs1	-	0.123798457	-	-	-
Restesg1	-	-	-	-	0.116672103
Restesg2	-	-	-	-	0.349273661
Thalach	-	-0.011414000	-0.009358466	-	-0.008388721
Exhang1	-	-	-	-	0.220072892
Oldpeak	-	0.147753916	0.126466508	0.1254514	0.383330112
Slope2	-	0.152809955	-	-	0.534273268
Slope3	-	-	-	-	-
Ca1	0.52313939	0.367766526	-	-	1.600399156
Ca2	-	0.628853632	0.628026126	-	1.509706691
Ca3	-	-	-	1.3890245	0.505074943
Thal6	-	0.018017631	-	0.2455078	0.261996571
Thal7	-	0.291724476	1.081814503	-	1.498038976
ACCURACY	**76.7%**	**93.9%**	**87.8%**	**96.9%**	**78.7%**

Exhibit 1: Output from logistic regression for different levels of num



	Cluster 1	Cluster 2
Number of people who have disease	126/229 = 55%	11/67 = 16%
Mean Age	57	44
M:F Ratio	148:81	52:15
Cholesterol	252.19	231.015
BP	134.4	122.2
Thalassemia	3,6,7 : 114, 15, 100	3,6,7: 50, 2,15
Avg. of max heart rate	144.74	166.19
No. of obs with high Sugar	148	52

Exhibit 2: Output based on clustering.



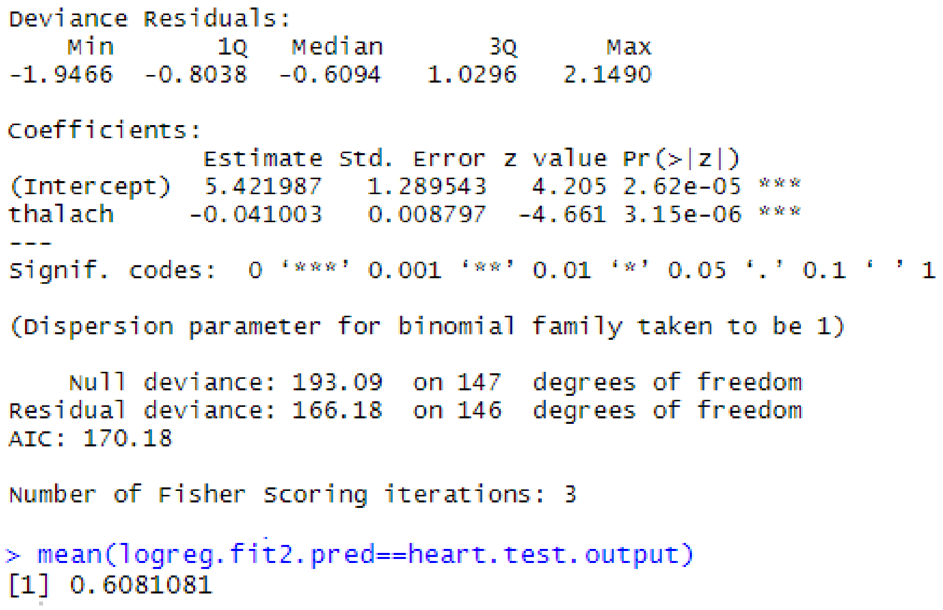
```
Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.0201  -0.8088  -0.6057   1.0248   2.1278  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  4.642370   1.942994   2.389   0.0169 *  
thalach     -0.040819   0.008836  -4.620 3.84e-06 ***
trestbps     0.005686   0.010681   0.532   0.5945    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 193.09  on 147  degrees of freedom
Residual deviance: 165.90  on 145  degrees of freedom
AIC: 171.9

Number of Fisher Scoring iterations: 3
```



```
Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.9466  -0.8038  -0.6094   1.0296   2.1490  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  5.421987   1.289543   4.205 2.62e-05 ***
thalach     -0.041003   0.008797  -4.661 3.15e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 193.09  on 147  degrees of freedom
Residual deviance: 166.18  on 146  degrees of freedom
AIC: 170.18

Number of Fisher Scoring iterations: 3

> mean(logreg.fit2.pred==heart.test.output)
[1] 0.6081081
```

Exhibit 3: Output from Logistic Regression on Exercise Induced Angina (Exang)