Analysis of Heart Disease

Elias Fedai

8/5/2019

Initial Summary of Data

name	type	na	mean	disp	median	mad	min	max	nlevs
Age	numeric	0	56.1805112	8.9253439	57.0	8.89560	29.0	77.0	0
Gender	numeric	0	0.8194888	0.3849202	1.0	0.00000	0.0	1.0	0
CP	numeric	0	3.3753994	0.8879470	4.0	0.00000	1.0	4.0	0
Trestbps	numeric	58	131.8996479	19.7787266	130.0	14.82600	0.0	200.0	0
Chol	numeric	7	176.4878837	118.1847099	216.0	80.06040	0.0	564.0	0
FBS	numeric	82	0.2169118	0.4125214	0.0	0.00000	0.0	1.0	0
RestECG	numeric	1	0.7856000	0.8672000	0.0	0.00000	0.0	2.0	0
Thalach	numeric	54	136.7342657	27.0289981	140.0	29.65200	60.0	202.0	0
Exang	numeric	54	0.4335664	0.4960007	0.0	0.00000	0.0	1.0	0
Oldpeak	numeric	62	1.0313830	1.1466644	0.9	1.33434	-2.6	6.2	0
Slope	numeric	119	1.7455621	0.6596464	2.0	0.00000	1.0	3.0	0
CA	numeric	320	0.6830065	0.9378238	0.0	0.00000	0.0	3.0	0
Thal	numeric	220	5.0492611	1.9341089	6.0	1.48260	3.0	7.0	0
target	integer	0	1.2939297	1.2380093	1.0	1.48260	0.0	4.0	0

*Data needs to be cleaned.

Summary of Data (clean)

name	type	na	mean	disp	median	mad	min	max	nlevs
Age	numeric	0	54.5469799	9.0348823	56.0	8.89560	29	77.0	0
Gender	numeric	0	0.6778523	0.4680852	1.0	0.00000	0	1.0	0
CP	numeric	0	3.1610738	0.9644671	3.0	1.48260	1	4.0	0
Trestbps	numeric	0	131.6543624	17.7458108	130.0	14.82600	94	200.0	0
Chol	numeric	0	246.8557047	52.6070752	242.5	47.44320	100	564.0	0
FBS	numeric	0	0.1442953	0.3519800	0.0	0.00000	0	1.0	0
RestECG	numeric	0	0.9932886	0.9949140	1.0	1.48260	0	2.0	0
Thalach	numeric	0	149.5000000	22.9670017	152.5	22.98030	71	202.0	0
Exang	numeric	0	0.3288591	0.4705889	0.0	0.00000	0	1.0	0
Oldpeak	numeric	0	1.0570470	1.1644426	0.8	1.18608	0	6.2	0
Slope	numeric	0	1.6040268	0.6175742	2.0	1.48260	1	3.0	0
CA	numeric	0	0.6744966	0.9382019	0.0	0.00000	0	3.0	0
Thal	numeric	0	4.7382550	1.9398218	3.0	0.00000	3	7.0	0
target	factor	0	NA	0.4630872	NA	NA	138	160.0	2
Disease	factor	0	NA	0.4630872	NA	NA	138	160.0	2

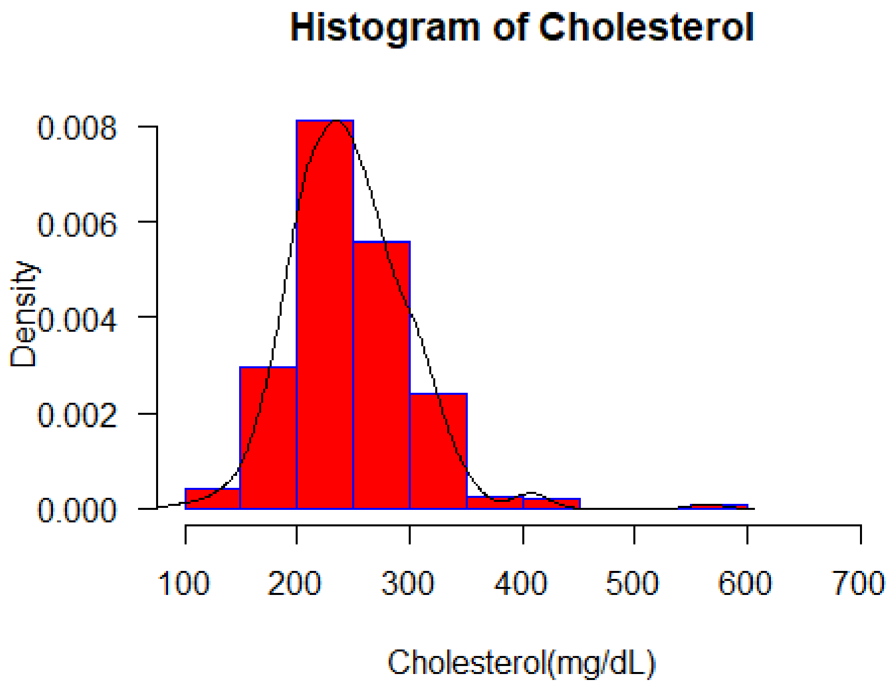
Table of Heart Disease

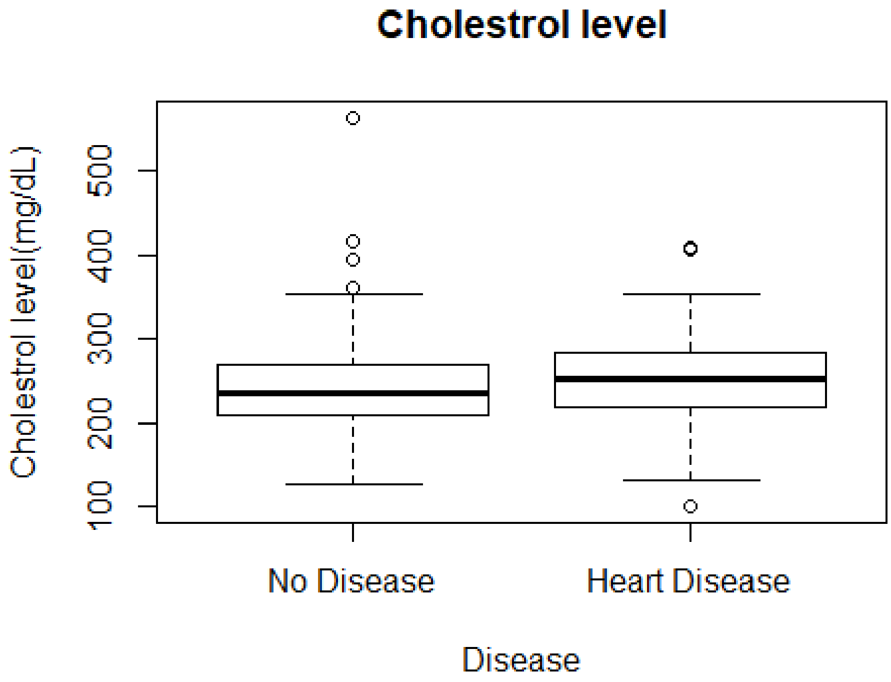
Var1	Freq
0	160
1	138

*There seems to be a lower count on heart disease cases than non heart disease cases.

Cholesterol

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    100.0   211.0   242.5   246.9   275.8   564.0
```

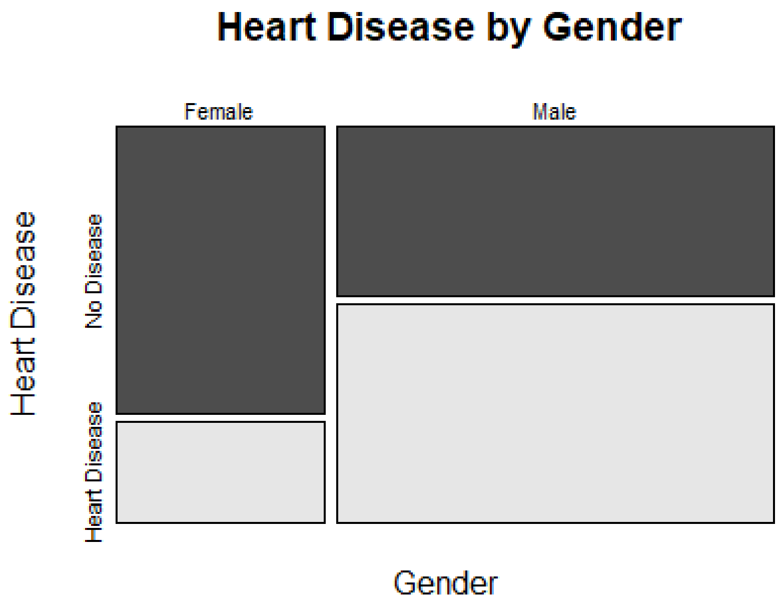


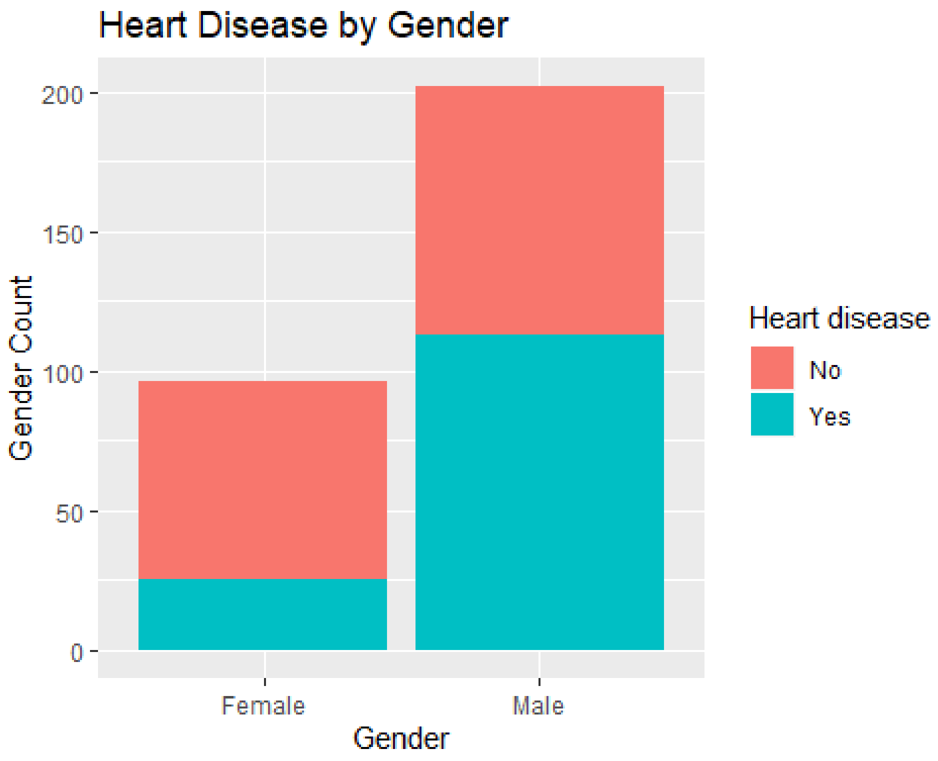


*Both heart and no disease display slightly elevated cholestrol levels.

Gender

```
## Female   Male
##     96    202
```

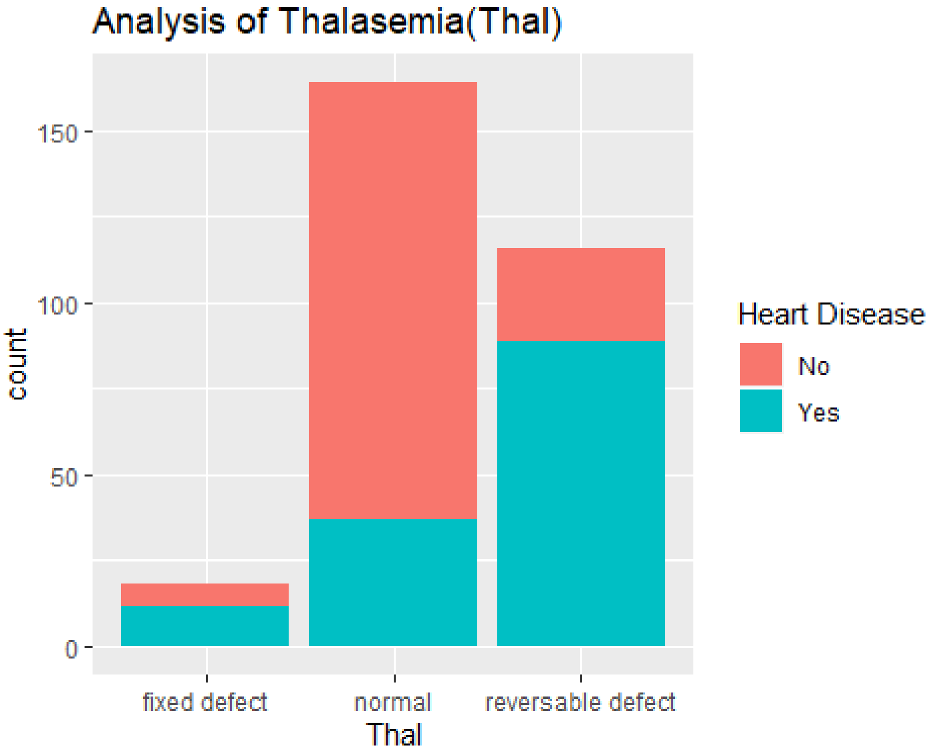




*Graphs/tables display more males where used in this study also there is a higher frequency of males with heart disease.

Thalassemia

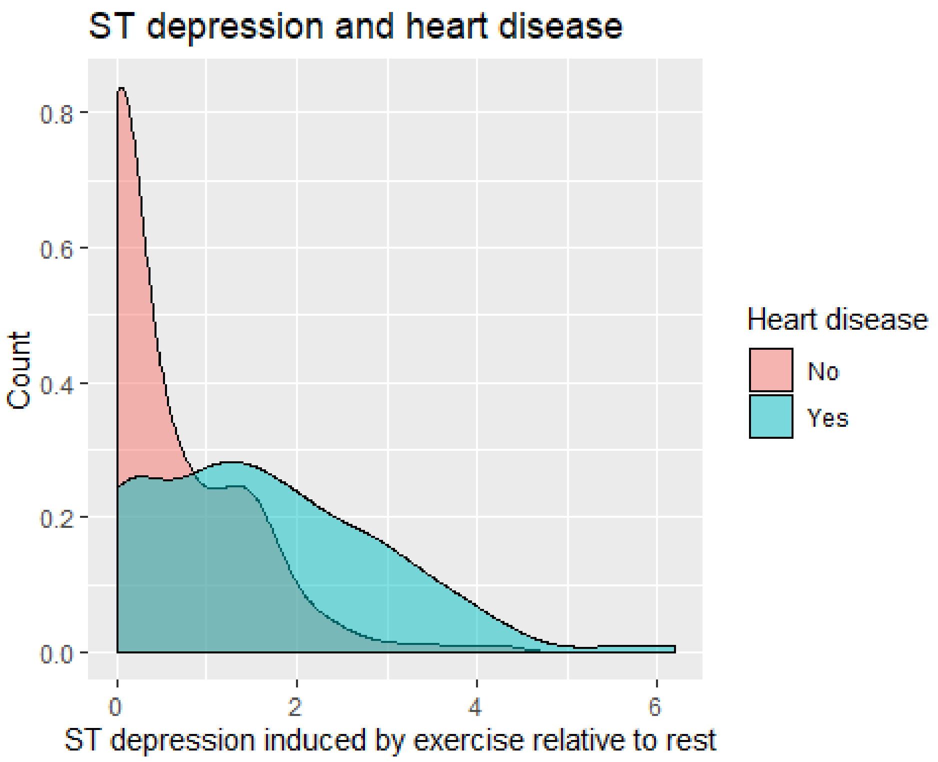
```
##      fixed defect            normal reversable defect
##                18               164               116
```

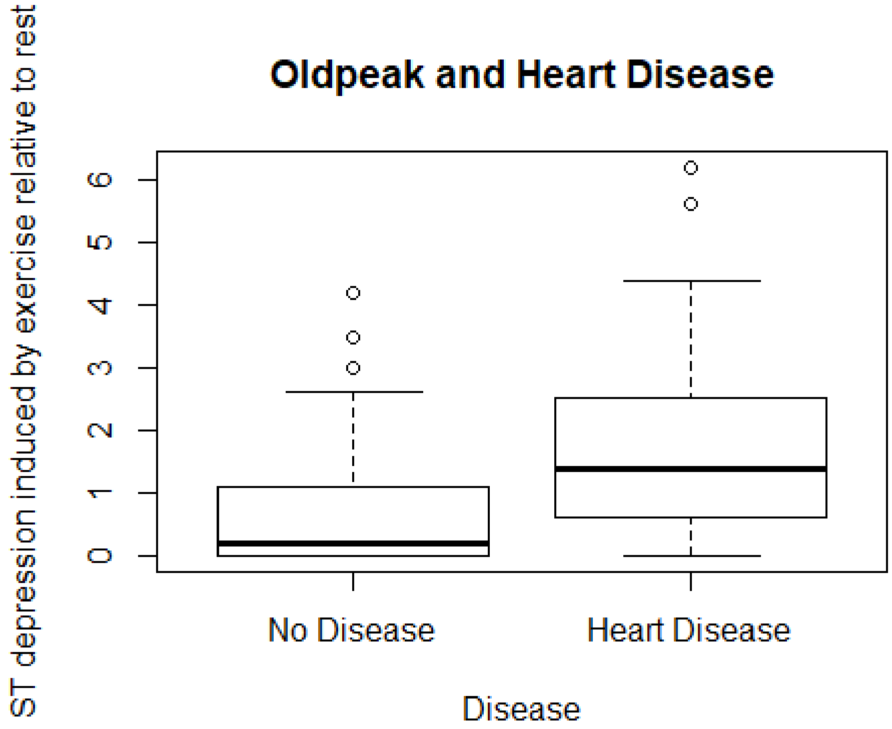


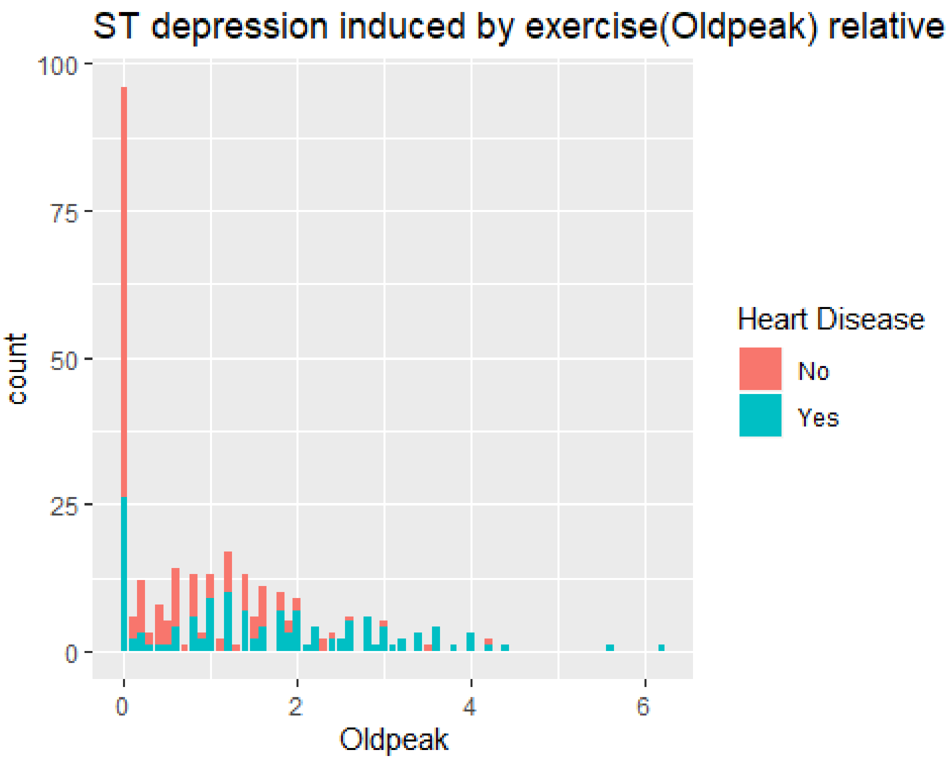
*Reversable defect displays a higher measure for heart disease.

Oldpeak

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   0.000   0.800   1.057   1.600   6.200
```


ST depression and heart disease
Count
0.0
0.2
0.4
0.6
0.8
0
2
4
6
ST depression induced by exercise relative to rest
Heart disease
No
Yes

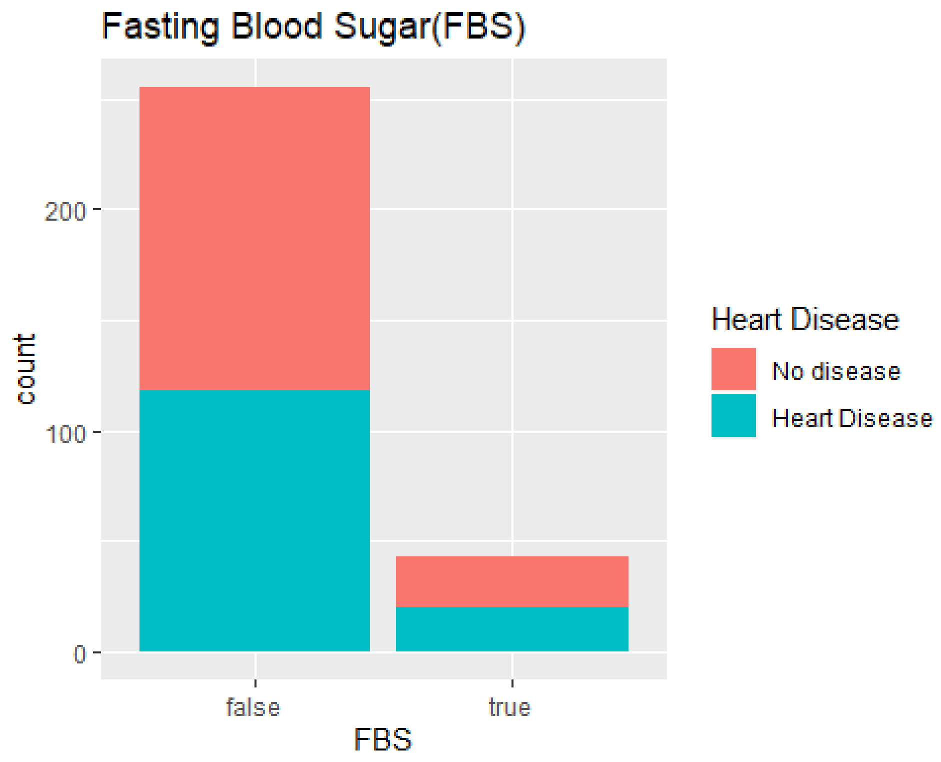

Oldpeak and Heart Disease
ST depression induced by exercise relative to rest
0
1
2
3
4
5
6
No Disease
Heart Disease
Disease

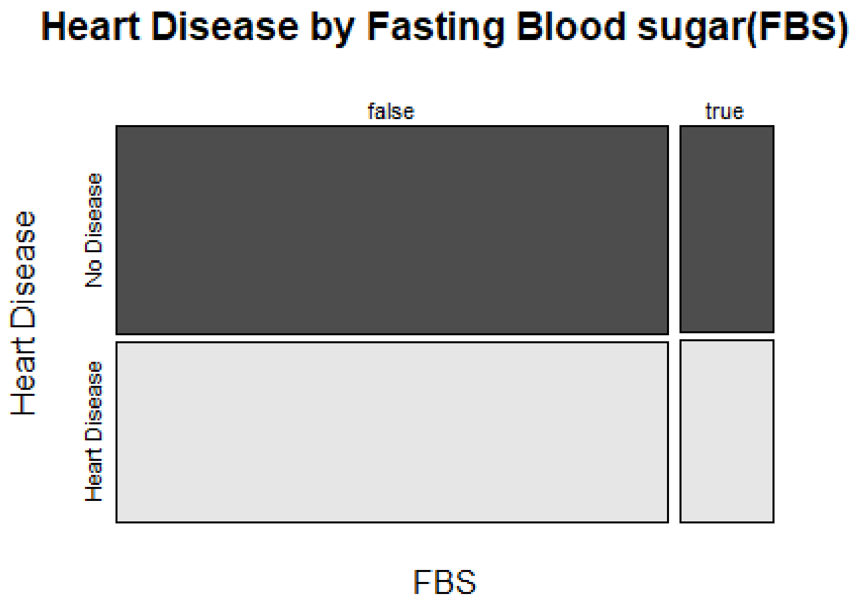


*Visuals for ST levels show a heavy right skewness. The no disease displays slightly more outliers.

Fasting blood sugar

```
## false  true
##   255    43
```

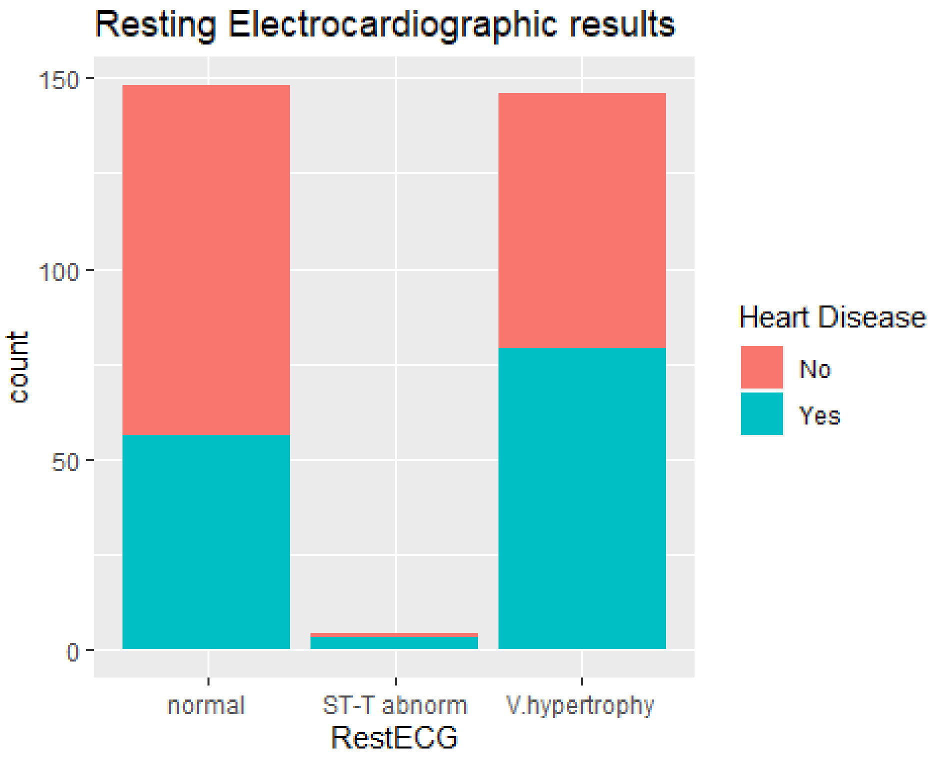




*Visuals appear to show not much possible significance for fasting blood sugar.

Resting Electrocardiographic

```
##        normal   ST-T abnorm V.hypertrophy
##           148             4           146
```



*V hypertrophy displayed a higher count for heart disease.

Logistic regression

```
## 
## Call:
## glm(formula = target ~ Gender + Age + CP + Trestbps + RestECG + 
##     Thalach + Exang + Slope + CA + Thal + Oldpeak + FBS, family = 
binomial, 
##     data = combinedclean)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.7161  -0.5050  -0.1469   0.3476   2.8296  
## 
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)    
## (Intercept)             -2.81103    2.73947  -1.026 0.304835    
## GenderMale               1.39407    0.50438   2.764 0.005712 ** 
## Age                     -0.01034    0.02440  -0.424 0.671677    
## CPatypical angina       -0.82837    0.55808  -1.484 0.137722    
## CPnon-anginal pain      -1.84643    0.50202  -3.678 0.000235 ***
## CPtypical angina        -2.12575    0.66391  -3.202 0.001365 ** 
## Trestbps                 0.02440    0.01125   2.169 0.030095 *  
## RestECGST-T abnorm     0.68976    2.25026   0.307 0.759206    
## RestECGV.hypertrophy   0.54077    0.37750   1.432 0.152004    
## Thalach               -0.01632    0.01097  -1.487 0.136973    
## Exangyes               0.70812    0.43582   1.625 0.104205    
## Slopeflat              0.71177    0.84746   0.840 0.400971    
## Slopeupsloping        -0.45645    0.92176  -0.495 0.620461    
## CA                     1.29252    0.27743   4.659 3.18e-06 ***
## Thalnormal             0.06636    0.78376   0.085 0.932521    
## Thalreversable defect  1.50432    0.76483   1.967 0.049198 *  
## Oldpeak                0.38048    0.22962   1.657 0.097526 .  
## FBStrue               -0.56212    0.60200  -0.934 0.350429    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 411.49  on 297  degrees of freedom
## Residual deviance: 193.03  on 280  degrees of freedom
## AIC: 229.03
## 
## Number of Fisher Scoring iterations: 6
```

*Parameters that were seen to be insignificant will be removed.

```
## 
## Call:
## glm(formula = target ~ Gender + CP + CA + Trestbps + Thal + Oldpeak, 
##     family = binomial, data = training1)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.5514  -0.4789  -0.1509   0.3754   2.5420  
## 
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)    
## (Intercept)           -2.188512   2.129288  -1.028 0.304037    
## GenderMale             1.323882   0.540842   2.448 0.014372 *  
## CPatypical angina     -2.038021   0.694037  -2.936 0.003320 ** 
## CPnon-anginal pain    -2.380433   0.565135  -4.212 2.53e-05 ***
## CPtypical angina      -1.591209   0.741720  -2.145 0.031929 *  
## CA                     0.716567   0.283673   2.526 0.011536 *  
## Trestbps               0.002312   0.013238   0.175 0.861361    
## Thalnormal            -0.195208   1.112786  -0.175 0.860748    
## Thalreversable defect  1.581107   1.117609   1.415 0.157150    
## Oldpeak                0.996666   0.296982   3.356 0.000791 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
```

```
##     Null deviance: 268.61  on 193  degrees of freedom
## Residual deviance: 134.01  on 184  degrees of freedom
## AIC: 154.01
##
## Number of Fisher Scoring iterations: 6
```

```
##          Actual
## Predicted  0  1
##         0 89 23
##         1 12 70
```

```
## [1] 0.1804124
```

*Training model displayed 18% misclassification error. Which will give an Accuracy of 82%

```
## [1] 1.339913e-24
```

*fit test p value indicates this model is significant.

```
##
## Call:
## glm(formula = target ~ Gender + CP + CA + Trestbps + Thal + Oldpeak,
##     family = binomial, data = test1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1894  -0.4441  -0.1346   0.2815   2.6023
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -6.72817    2.91477  -2.308 0.020982 *
## GenderMale              0.86534    0.91177   0.949 0.342580
## CPatypical angina      -0.62171    0.92270  -0.674 0.500439
## CPnon-anginal pain     -2.00075    0.82290  -2.431 0.015042 *
## CPtypical angina       -4.58981    1.68282  -2.727 0.006383 **
## CA                      1.76310    0.48720   3.619 0.000296 ***
## Trestbps                0.03886    0.01725   2.253 0.024255 *
## Thalnormal             -0.92804    1.23453  -0.752 0.452209
## Thalreversable defect   0.90323    1.16946   0.772 0.439908
## Oldpeak                 0.69001    0.31949   2.160 0.030791 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 142.284  on 103  degrees of freedom
## Residual deviance:  64.742  on  94  degrees of freedom
## AIC: 84.742
##
## Number of Fisher Scoring iterations: 6
##           Actual
## Predicted  0  1
##         0 54  9
##         1  5 36

## [1] 0.1346154
```

*Test model displayed a 13.5% misclassification error. Which will give an Accuracy of 86.5%

```
## [1] 4.966e-13
```

*Fit test p value indicates this model is signicant.

```
## 
## Call:
## glm(formula = target ~ RestECG + Age, family = binomial, data = training2)
## 
## Deviance Residuals: 
##    Min      1Q  Median      3Q     Max  
## -2.245  -1.034  -0.596   1.021   1.906  
## 
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)    
## (Intercept)          -4.71102    1.02778  -4.584 4.57e-06 ***
## RestECGST-T abnorm    1.00318    1.20653   0.831   0.4057    
## RestECGV.hypertrophy  0.56444    0.30067   1.877   0.0605 .  
## Age                   0.08084    0.01869   4.325 1.52e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 288.04  on 207  degrees of freedom
## Residual deviance: 260.11  on 204  degrees of freedom
## AIC: 268.11
## 
## Number of Fisher Scoring iterations: 4

##           Actual
## Predicted  0  1
##         0 76 35
##         1 32 65

## [1] 0.3221154
```

*training model displayed 32.2% misclassifcation error. which will give an Accuracy of 67.8%

```
## [1] 3.748741e-06
```

*Fit test p value indicates this model is signifcant.

```
##
## Call:
## glm(formula = target ~ RestECG + Age, family = binomial, data = test2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1837  -1.1750  -0.8519   1.1780   1.5444
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.7868688  1.3239068  -0.594   0.5523
## RestECGV.hypertrophy  0.8230585  0.4594067   1.792   0.0732 .
## Age                  -0.0006296  0.0235104  -0.027   0.9786
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 122.58  on 89  degrees of freedom
## Residual deviance: 119.18  on 87  degrees of freedom
## AIC: 125.18
##
## Number of Fisher Scoring iterations: 4
```

```
##          Actual
## Predicted  0  1
##         0 39 28
##         1 13 10
```

```
## [1] 0.4555556
```

*Test model displayed 45.6% misclassification error. Which will give an Accuracy of 54.4%

```
## [1] 0.1822464
```

*Fit test p value indicates this model is not significant.