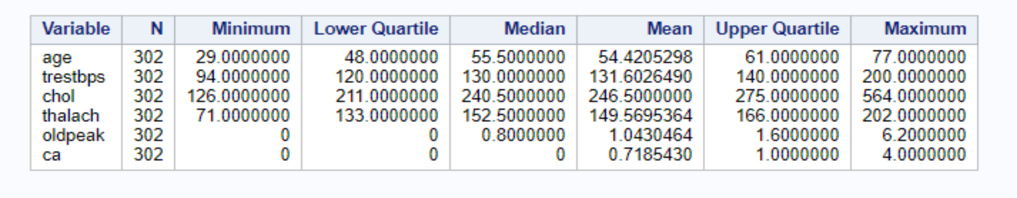
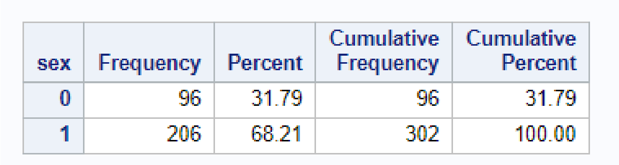
APPENDICES:

Appendix A: Exploratory Data Analysis



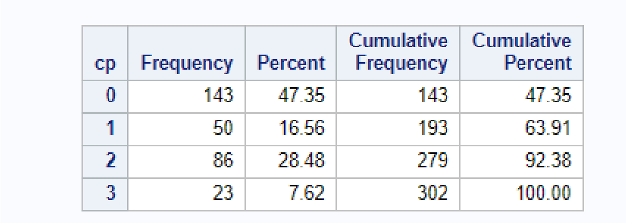
Variable	N	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum
age	302	29.0000000	48.0000000	55.5000000	54.4205298	61.0000000	77.0000000
trestbps	302	94.0000000	120.0000000	130.0000000	131.6026490	140.0000000	200.0000000
chol	302	126.0000000	211.0000000	240.5000000	246.5000000	275.0000000	564.0000000
thalach	302	71.0000000	133.0000000	152.5000000	149.5695364	166.0000000	202.0000000
oldpeak	302	0	0	0.8000000	1.0430464	1.6000000	6.2000000
ca	302	0	0	0	0.7185430	1.0000000	4.0000000

Table A-1: *This table includes the minimum, maximum, 1st Quartile, median, mean, and 3rd Quartile values for 6 numerical variables used in the presence of heart attack analysis.*



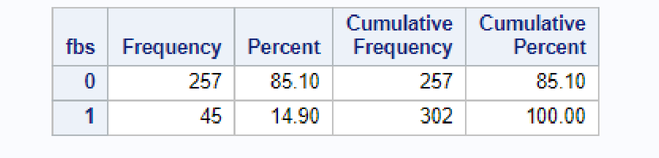
sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	96	31.79	96	31.79
1	206	68.21	302	100.00

Table A-2: *This table displays the counts pf the sex of the subjects (0 = Female; 1 = Male)*



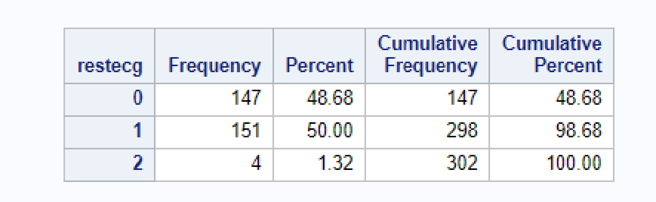
cp	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	143	47.35	143	47.35
1	50	16.56	193	63.91
2	86	28.48	279	92.38
3	23	7.62	302	100.00

Table A-3: *This table displays the counts for each of the chest pain type (0 = typical angina; 1 = atypical angina; 2 = non-anginal pain; 3 = asymptomatic)*



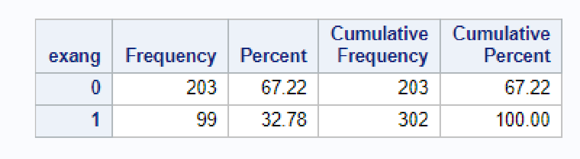
fbs	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	257	85.10	257	85.10
1	45	14.90	302	100.00

Table A-4: *This table displays the counts of the fasting blood sugar > 120 mg/dl (0 = False; 1 = True)*



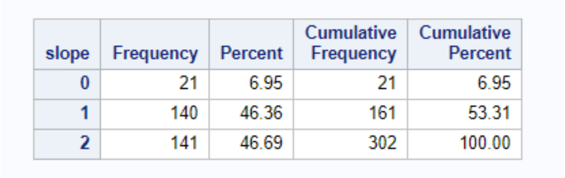
restecg	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	147	48.68	147	48.68
1	151	50.00	298	98.68
2	4	1.32	302	100.00

Table A-5: *This table displays the counts for each of the rest electrocardiographic result (0 = Normal; 1 = Having ST-T wave abnormality T wave inversions and/or ST elevation or depression of > 0.05 mV; 2 = Showing probable or definite left ventricular hypertrophy by Estes' criteria)*



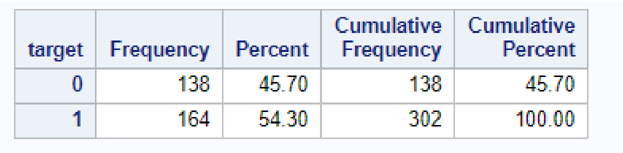
exang	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	203	67.22	203	67.22
1	99	32.78	302	100.00

Table A-6: *This table displays the counts of the exercise induced angina (0 = No; 1 = Yes)*



slope	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	21	6.95	21	6.95
1	140	46.36	161	53.31
2	141	46.69	302	100.00

Table A-7: *This table displays the counts for each of the slope of the peak exercise (0 = Upsloping; 1 = Flat; 2 = Downsloping)*



target	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	138	45.70	138	45.70
1	164	54.30	302	100.00

Table A-8: *This table displays the counts of the response variable, presence of heart attack (0 = No/less chance of getting heart attack; 1 = More chance of getting heart attack).*

Appendix B: Statistical analysis / Modeling

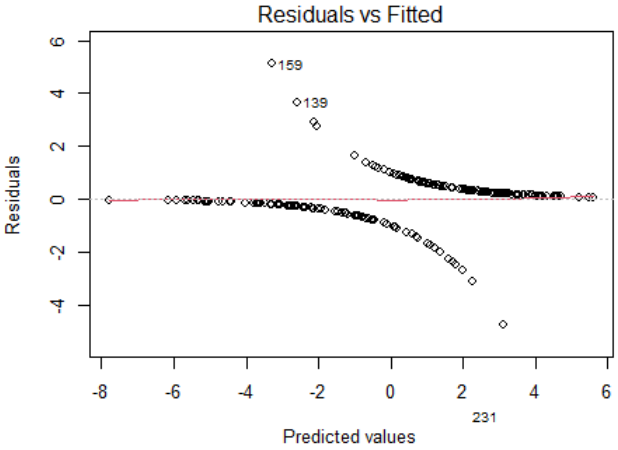


Figure B-1: *This residuals vs fitted plot for the model fitted with conventional logistic regression after stepwise selection into the training dataset shows no obvious pattern and most of the residuals are located near the loess line.*

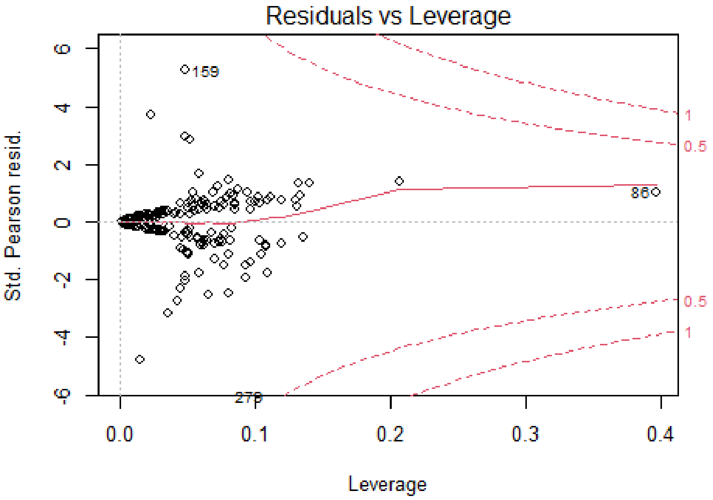
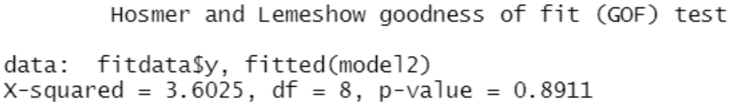


Figure B-2*: This residuals vs leverage for the model fitted with conventional logistic regression after stepwise selection into the training dataset shows there is an obvious outlier observation #86 and there are influential points since all cases are well inside the cook's distance lines. We would not remove the outlier because this is what was actually observed, and therefore it is meaningful to include it into the study.*



```
                Hosmer and Lemeshow goodness of fit (GOF) test

data:  fitdata$y, fitted(model2)
X-squared = 3.6025, df = 8, p-value = 0.8911
```

Figure B-3*: This is the Hosmer and Lemeshow goodness of fit test for the model fitted with conventional logistic regression after stepwise selection into the training dataset . The p-value of this test is not significant when compared to alpha = 0.05 and there is no sign of lack of fit.*

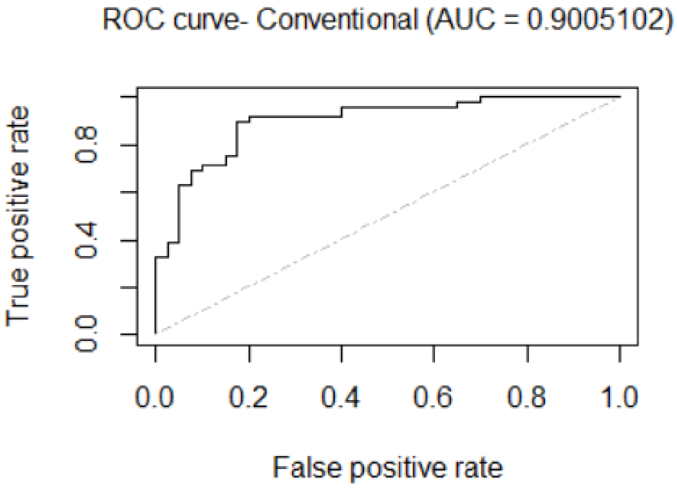


Figure B-4*: This Receiver Operating Characteristics (ROC) curve for the testing set fitted with conventional logistic regression after stepwise selection has an area under the curve of*

0.9005102 in which 0.5 is used as the classification cutoff. The predictive performance of the testing dataset is pretty good since the area under the curve is close to 1.

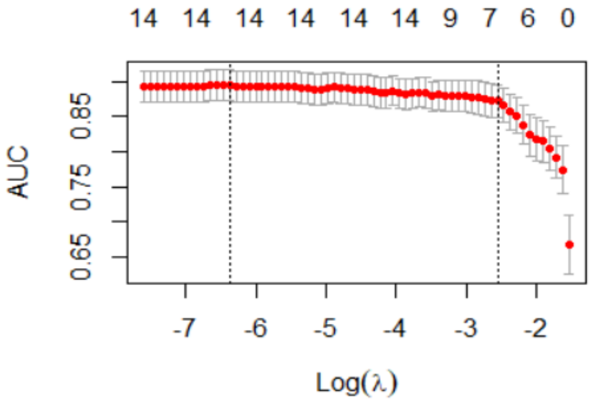


Figure B-5*: This log Lambda vs AUC shows the area under the curve for ROC for each log lambda and the best lambda for the training set fitted with logistic regression with LASSO that gives the largest area under the curve is 0.00171024.*

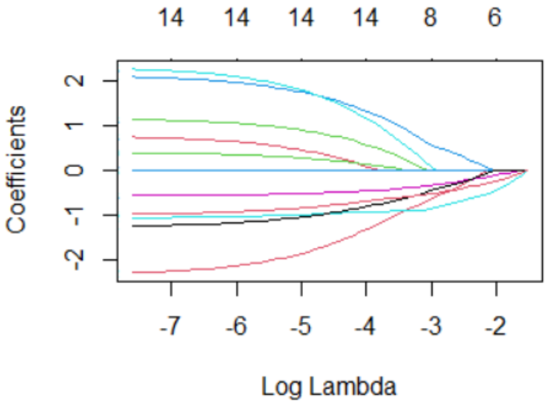


Figure B-6*: This is the solution path of the log lambdas in which 14 covariates are selected and 2 covariates' coefficients are shrunk to 0 after fitted with the LASSO model.*

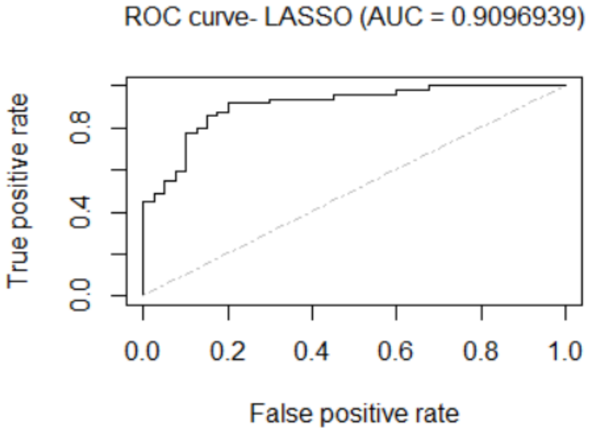


Figure B-7*: This Receiver Operating Characteristics (ROC) curve for the testing set fitted with logistic regression with LASSO penalty has an area under the curve of 0.9096939 in which 0.5 is used as the classification cutoff. The predictive performance of the testing dataset fitted with* *the LASSO model is better than the testing dataset fitted with conventional logistic regression after running stepwise selection.*

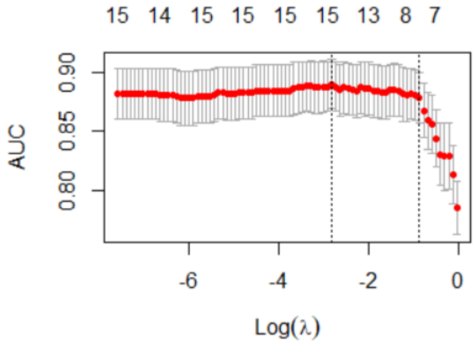


***Figure B-8**: This log Lambda vs AUC shows the area under the curve for ROC for each log lambda and the best lambda for the training set fitted with elastic net regression that gives the largest area under the curve is 0.05958795.*

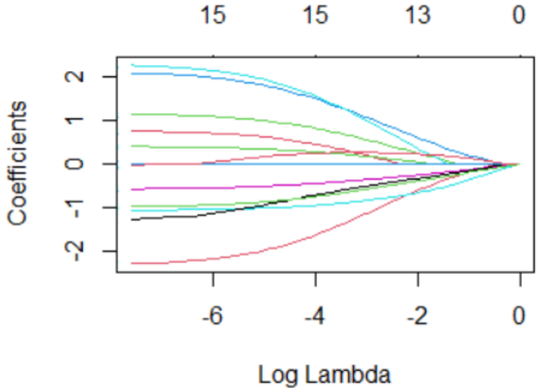


***Figure B-9**: This is the solution path of the log lambdas in which 15 covariates are selected and 1 coefficient of the covariate is shrunk to 0 after fitted with the elastic net regression.*

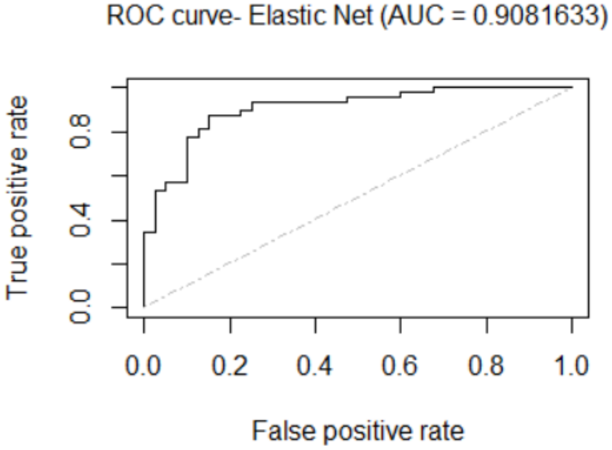


***Figure B-10**: This Receiver Operating Characteristics (ROC) curve for the testing set fitted with elastic net regression has an area under the curve of 0.9081633 in which 0.5 is used as the classification cutoff. The predictive performance of the testing dataset fitted with the elastic net regression is slightly weaker than the LASSO model but is better than the testing dataset fitted with conventional logistic regression after running stepwise selection.*