Introduction:

The goal of this project is to analyze which factors can be used to predict whether or not a patient has a heart condition. What factors affect the existence of a heart condition? This is to help analyze a heat condition using externally available values like heart rate, ECG age etc.

This data set was gathered by Dr. Robert Detrano at the VA Medical Center in Long Beach, CA. The data set consists of 303 observations, 138 observations hold information about patients with heart disease. The dataset contains 10 numeric variables: 4 continuous variables and 6 qualitative variables.

Analysis:

Explaining Variables:

Below are the details of the variables mentioned in the Dataset :

Name	*Description*
Age	Age in years
Sex	Male =1, Female =0;
Chest_Pain_Type	1 – 4 (typical angina, abnormal angina, Non-anginal pain, Asymptomatic)
Blood_Pressure	Resting Blood pressure upon hospital admission (in mmHg)
Cholesterol	Serum cholesterol in mg/dl
Fasting_Blood_Sugar	Is fasting blood sugar < 120 mg/dl? 1=true, 0=false
Resting_ECG	Normal = 0, Abnormal = 1, Left ventricular hypertrophy = 2
Maximum_Heart_Rate	Maximum heart rate achieved after exercising
Angina	Yes = 1, No = 0 – Does the patient experience angina as a result of exercise?
Heart_condition	Healthy =1, Sick =0. Angiographic disease status.

Dependent variable: Heart_condition

Importing the Data in SAS and Creating Dummy Variables for Categorical Data:

The dataset Cardiology has a Dependent variable: Heart Condition as a Binary Variable where when heart_condition = 0 the patient does not have a heart condition and when heart_condition = 1, the patient has a heart condition.

Since there are two variables containing categorical data, dummy variables will have to be created to manage them :

Chest_Pain_Type :
1 – 4 (typical angina, abnormal angina, Non-anginal pain, Asymptomatic)
Thus we will create 3 dummy variables

Where dum_cpt1 = 1 when chest_pain_type = **1**
Where dum_cpt2 = 1 when chest_pain_type = **2**
Where dum_cpt3 = 1 when chest_pain_type = **3**

Resting_ECG
Normal = 0, Abnormal = 1, Left ventricular hypertrophy = 2
(Thus we will create 2 dummy variables=)

Where dum_r_ecg1 = 1 when resting_ecg = **1**
Where dum_r_ecg2 = 1 when resting_ecg = **2**

Exploratory Data Analysis and Visualization:

Since our Dataset has a dependent variable with Qualitative Binary Data, we will pursue a logistic regression model.

From Fig 1, we see that out of the total observed cases, 138 (that is 45.54 %)do not have a heart condition while 165 (54.46%)do have a heart condition. The total number of observations is 303. There are no missing or null values.

The sample size (n) = 303,
The number of Predictors (k) = 9,
The Cases where Y is 1 = 165,
In the Cases where Y is 0 = 138,
The Sample is large enough as 303/9 = 33.67 which is greater than 10 (i.e there are more than 10 observations for every predictor present)

BOXPLOTS:

From the FIG 2, we see a boxplot to form a relationship between age and heart condition. We can say that because the media in heart condition = 0 is about 57, people of that age tend to not get a heart condition whereas those with the age of about 51 tend to have a heart condition.

With FIG 3, we see a boxplot to form a relationship between maximum heart rate. We can say that a heart rate of 162 tends to have a heart condition whereas a heart rate of 140 tends to not.

Separating Training and Test Data:

We will separate the data set into training and test datasets; the training set will be used to create a regression model and the test set will be used to test the accuracy and goodness of the regression model. Since we have a large number of observations (303), I will use an 80% Training, 20% Test ratio to divide the data for future validation.

The random seed picked for random separation = 592587

Fig 4 shows the number of observations retained for the training model being 243 with a seed of 592587 and a sampling rate of 0.8.

The name of the new training heart_condition variable will be new_y.

Building Our Model:

The general model equation:

$$\log(\text{new y= 1/new\_y = 0}) = \beta 0 + \beta 1(\text{age}) + \beta 2(\text{sex}) + \beta 3(\text{dum\_cpt1}) + \beta 4(\text{dum\_cpt2}) + \beta 5(\text{dum\_cpt3}) + \beta 6(\text{blood\_pressure}) + \beta 7(\text{cholestrol}) + \beta 8(\text{Fasting\_blood\_sugar}) + \beta 9(\text{dum\_r\_ecg1}) + \beta 10(\text{dum\_r\_ecg2}) + \beta 11(\text{maximum\_heart\_rate}) + \beta 12(\text{angina})$$

Fitting the full model:

As seen in Fig 5 the order of Standardized Estimate Coefficients, the highest influencers are dum_cpt3 > dum_cpt 2 > maximum_heart_rate > dum_cpt1. All of these variables have a p-value < 0.05 i.e. none of them are insignificant.

From Fig 5 we will generate a primary equation. We are eliminating all the insignificant predictors (that is with the value of p above 0.05).

Equation:

log(new_y= 1/new_y = 0) = -0.3043 -1.8584(sex) +1.8453(dum cpt1) + 2.2244(dum cpt2) + 2.1327(dum_cpt3)+ 0.0322(maximum_heart_rate)

Where sex = 1 (male)

Sex = 0 (female)

Dum_cpt1 = 1 (typical angina)

Dum_cpt2 = 1(abnormal angina)

Dum_cpt3 = 1 (Non-anginal pain)

Dum_cpt1 = Dum_cpt2 =Dum_cpt3 = 0 (Asymtomatic)

The Test of Significance:

The data is associated with two hypotheses.

The null hypothesis is that none of the predictors influence the dependent variable (new_y).

$\beta i = 0$

The Alternate Hypothesis is that there is at least one predictor that influences the dependent variable.

$\beta i \neq 0$

From the equation: log(new_y= 1/new_y = 0) = -0.3043 -1.8584(sex) +1.8453(dum cpt1) + 2.2244(dum cpt2) + 2.1327(dum_cpt3)+ 0.0322(maximum_heart_rate)

Where sex = 1 (male)

Sex = 0 (female)

Dum_cpt1 = 1 (typical angina)

Dum_cpt2 = 1(abnormal angina)

Dum_cpt3 = 1 (Non-anginal pain)

Dum_cpt1 = Dum_cpt2 =Dum_cpt3 = 0 (Asymtomatic)

We can thus, reject the null hypothesis and say that at least one of the variables present in the dataset influences the dependent variables.

Selection Process for Model Selection:

We have already established that we will pursue a logistic regression model and thus the selection options available to us are: Stepwise, Forward, and Backward

Below is a table with Significant Terms to check which model is better.

	Forward	Backward	Stepwise

R-Squared	0.3799	0.3799	0.3799
AIC	231.25	231.25	231.25
SC	252.21	252.21	252.21

As can be observed from the above data, all of the signifiers for selection criteria give us the same result with 5 variables: sex, dum_cpt1, dum_cpt2, dum_cpt3, and maximum_heart_rate. Thus we will build a regression model with these variables.

Likelihood Ratio Test:

Also, as can be seen from FIG9, the likelihood ratio is 116.129, while this is not very high, the fact that we have no other models to compare it to tells us that this is acceptable. If the p-value is less than 0.05, we can say that at least one predictor can account for the outcome of new_y.

Thus again, the null hypothesis is rejected.

The Final Fitted Model:

The Selected Model thus contains the following variables: sex, dum_cpt1, dum_cpt2, dum_cpt3, and maximum_heart_rate. We will fit an equation using FIG 10.

And the equation for the regression model is as follows:

$$\log(\text{new\_y} = 1/\text{new\_y} = 0) = -7.8548 - 2.2898(\text{sex}) + 2.1462(\text{dum\_cpt1}) + 3.7733(\text{dum\_cpt2}) + 2.5989(\text{dum\_cpt3}) + 0.0541(\text{maximum\_heart\_rate})$$

Checking for multicollinearity:

To check whether or not there is multicollinearity, we will produce a correlation table. If the Correlation between any of the predictors is greater than 0.9, it shows that they are influencing each other and thus their individual effect on the dependent variable is compromised.

As seen in FIG 11, the correlation values tell us that there is no multicollinearity present between the predictors as none of the Correlational values are above or equal to 0.9. Had there been a collinearity issue, we would have further checked collinearity signifiers like VIF to delete the variables that have the highest VIF, and check the correlations after every deletion.

Check for Outliers and Influencers:

Sometimes, errors in recording or unusual data can affect the Regression model and skew the results. We will look at the Pearson and Deviance values to check for any outliers. If either of those values is above or below 3, then the observation is an outlier.

As can be seen in the Graphs in FIG12, there are a few outliers. We will go ahead and delete them.

After deleting 4 observations we get the following graph in FIG 13 which still shows a few outliers.

And after deleting 3 more observations, we get the following graph in FIG 14. As can be seen from FIG14, there appear to be no more outliers. So we are done detecting outliers.

Deleting Influential Points :

We will look at the Dfbeta values for all the variables and those with at $|Dfbeta| > 2/\sqrt{n}$ which for our case is 0.11. We are starting with the value of r squared being 0.46 and we will stop when the difference between each r squared value becomes less than 0.03.

The first round of deleting the influential points we get an R sqr value of 0.72 . As can be seen in FIG 16

The second round of influential points and we see an increase to R srq value of 0.73 as can be seen in FIG 17.

Stopping now as there is less than a value if 0.03 increase.

However, this disturbs and greatly affects our odds ratio values as can be seen in FIG 18

So we will stick with the model we had before we removed the influential points.

Is the Model Satisfactory?

There are several ways to check if a model is satisfactory. We can look at the R squared value. Here the R squared value is 0.4611. This tells us that 46.11 percent of the variance in new_y can be explained by the predictors in our regression model. While this is not a very good model, it is barely moderate.

Effect of each Variable on the new_y variable:

The following is the way each variable affects the new_y variable as can be seen in FIG 20.

sex: If all other variables (predictors) stay the same and sex increases by one unit, the odds of new_y decrease by 89.9% [(0.101 - 1) * 100], with a 95% confidence interval of a decrease between 95.9% [(0.041 - 1) * 100] and 74.0% [(0.251 - 1) * 100].

dum_cpt1: If all other variables (predictors) stay the same and dum_cpt1 increases by one unit, the odds of new_y increase by 755.3% [(8.553 - 1) * 100], with a 95% confidence interval of an increase between 1387.0% [(2.387 - 1) * 100] and 2964.8% [(30.648 - 1) * 100].

dum_cpt2: If all other variables (predictors) stay the same and dum_cpt2 increases by one unit, the odds of new_y increase by 4252.5% [(43.525 - 1) * 100], with a 95% confidence interval of an increase between 8847.0% [(9.847 - 1) * 100] and 19139.0% [(192.390 - 1) * 100].

dum_cpt3: If all other variables (predictors) stay the same and dum_cpt3 increases by one unit, the odds of new_y increase by 1244.9% [(13.449 - 1) * 100], with a 95% confidence interval of an increase between 4592.0% [(5.592 - 1) * 100] and 3134.7% [(32.347 - 1) * 100].

maximum_heart_rate: If all other variables (predictors) stay the same and maximum_heart_rate increases by one unit, the odds of new_y increase by 5.6% [(1.056 - 1) * 100], with a 95% confidence interval of an increase between 3.4% [(1.034 - 1) * 100] and 7.8% [(1.078 - 1) * 100].

Thus the strongest predictor of the response variable is dum_cpt2 i.e. when the chest pain type is abnormal angina.

Predictions:

We will use the following values to create two predictions for the training dataset:

	age	sex	chest_pain_type	blood_pressure	cholesterol	Fasting_blood_sugar	resting_ecg	maximum_heart_rate	angina
1	62	1	4	128	210	0	2	130	1
2	54	1	2	135	260	0	0	168	0

The Predictions can be seen in FIG 21.

Thus the predicted probability that the patient who is male, 62 years old, has a chest pain type of 4, a blood pressure of 128 mmHg, a cholesterol level of 210 mg/dL, fasting blood sugar is <120 mg/dL, resting electrocardiographic results indicate abnormality (resting_ecg=2), the maximum heart rate is 130 beats per minute, experiences angina (angina=1) will have a heart condition is phat = 0.04 and the odds of that patient having a heart condition will increase between 1.75% to 10.53%.

And, the predicted probability that the patient who is male, 54 years old, has a chest pain type of 2, a blood pressure of 135 mmHg, a cholesterol level of 260 mg/dL, fasting blood sugar <120 mg/dL, resting electrocardiographic results is no abnormality, the maximum heart rate is 168 beats per minute, does not have angina (angina=0) will have a heart condition is phat = 0.93 and the odds of that patient having a heart condition will increase between 120.45% to 167.41%.

Validation:

To validate our trained regression model with the test dataset, we will first calculate our threshold value p where if predicted probability > p then heart condition = 1 else if predicted probability < p then heart condition = 0.
We will do this by creating the Classification table and Prob Level value with the greatest sum of Sensitivity and Speciality, our threshold value.

Thus from FIG 22 we see that the P value = 0.4

Now with the cutoff value, we will create a confusion matrix to assess our performance metrics.

With FIG 23,

TP = 25
FP = 7
TN = 19
FN = 9

Thus the 5 metrics are:
Sensitivity/ recall which is The probability that the screening is positive given that the person has a heart condition is 0.74
Specificity that is The probability that the screening is negative given that the person does not have a heart condition is 0.79
The accuracy which is how close is the predicted value to the actual value is 0.73
Precision The believability of the model, when an instance is positive, is 0.78
And the harmonic mean of precision and recall (F metric) is 0.759.

APPENDIX:

Frequency of Heart Condition

The FREQ Procedure

heart_condition	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	138	45.54	138	45.54
1	165	54.46	303	100.00

Fig 1: Frequency Table

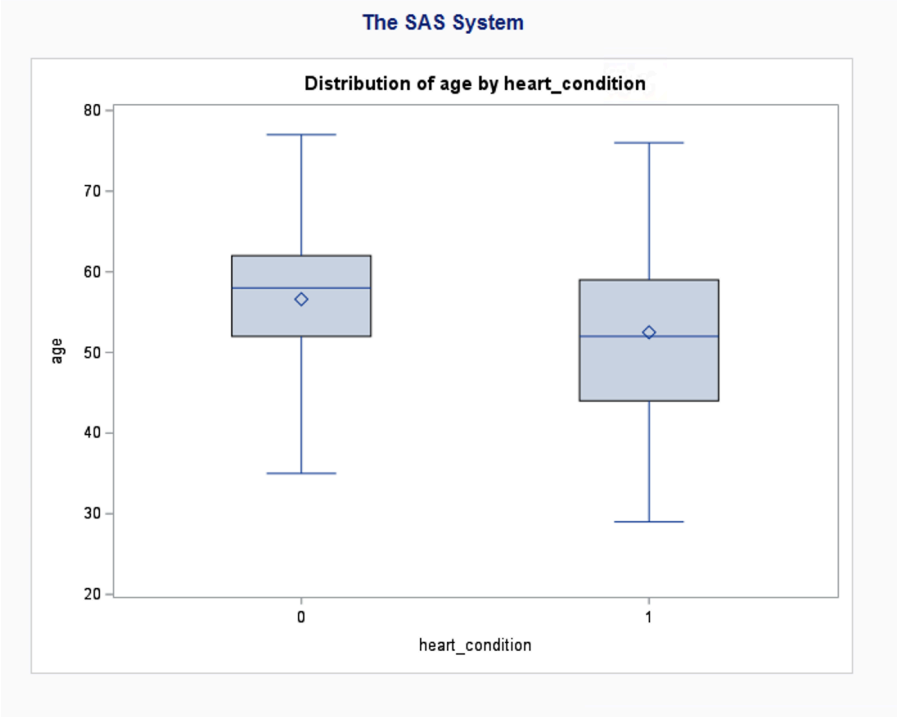

Fig 2

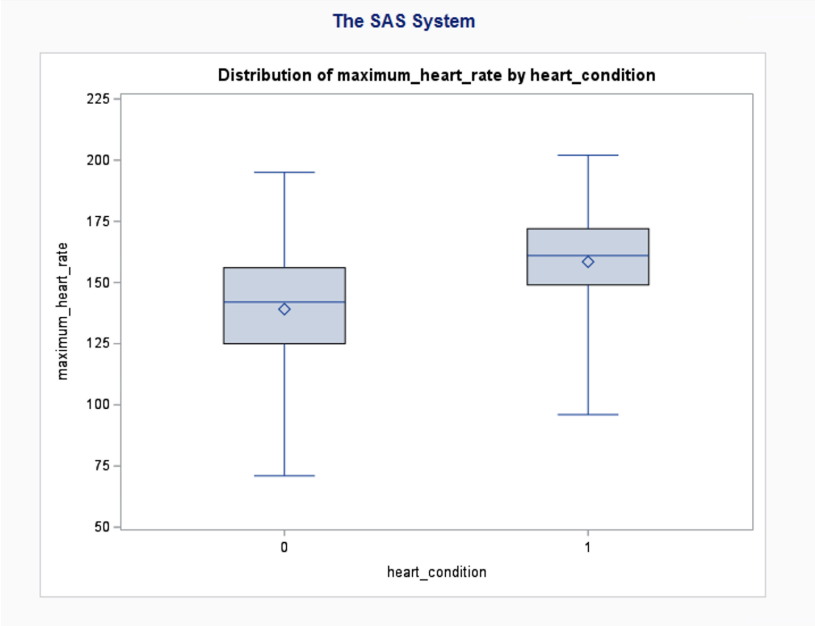

Fig 3

Test and train set for Heart Condition

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	CARDIOLOGY
Random Number Seed	592587
Sampling Rate	0.8
Sample Size	243
Selection Probability	0.80198
Sampling Weight	0
Output Data Set	XV_ALL

Fig 4

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-0.3043	2.3832	0.0163	0.8984	
age	1	-0.0250	0.0230	1.1852	0.2763	-0.1240
sex	1	-1.8584	0.4344	18.2999	<.0001	-0.4840
dum_cpt1	1	1.8453	0.6347	8.4520	0.0036	0.2600
dum_cpt2	1	2.2244	0.5732	15.0609	0.0001	0.4464
dum_cpt3	1	2.1327	0.4578	21.7000	<.0001	0.5358
blood_pressure	1	-0.0117	0.0107	1.2041	0.2725	-0.1115
cholesterol	1	-0.00381	0.00353	1.1625	0.2809	-0.1126
Fasting_blood_sugar	1	-0.1802	0.4835	0.1389	0.7094	-0.0365
dum_r_ecg1	1	-0.8173	1.4084	0.3367	0.5617	-0.0574
dum_r_ecg2	1	-0.2181	0.3582	0.3710	0.5425	-0.0602
maximum_heart_rate	1	0.0322	0.00993	10.4991	0.0012	0.3929
angina	1	-0.5286	0.4165	1.6112	0.2043	-0.1377

Fig 5

Step 5. Effect dum_cpt1 entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	337.382	231.253
SC	340.875	252.212
-2 Log L	335.382	219.253

R-Square	0.3799	Max-rescaled R-Square	0.5076

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	116.1290	5	<.0001
Score	98.6331	5	<.0001
Wald	66.0015	5	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
9.1321	7	0.2433

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-5.5251	1.2967	18.1557	<.0001	
sex	1	-1.5975	0.3837	17.3346	<.0001	-0.4161
dum_cpt1	1	1.7411	0.5993	8.4410	0.0037	0.2454
dum_cpt2	1	2.4661	0.5435	20.5916	<.0001	0.4949
dum_cpt3	1	2.2728	0.4049	31.5162	<.0001	0.5710
maximum_heart_rate	1	0.0374	0.00873	18.3382	<.0001	0.4569

Fig 6

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	337.382	231.253
SC	340.875	252.212
-2 Log L	335.382	219.253

R-Square	0.3799	Max-rescaled R-Square	0.5076

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	116.1290	5	<.0001
Score	98.6331	5	<.0001
Wald	66.0015	5	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
9.1321	7	0.2433

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-5.5251	1.2967	18.1557	<.0001	
sex	1	-1.5975	0.3837	17.3346	<.0001	-0.4161
dum_cpt1	1	1.7411	0.5993	8.4410	0.0037	0.2454
dum_cpt2	1	2.4661	0.5435	20.5916	<.0001	0.4949
dum_cpt3	1	2.2728	0.4049	31.5162	<.0001	0.5710
maximum_heart_rate	1	0.0374	0.00873	18.3382	<.0001	0.4569

Fig 7

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	337.382	231.253
SC	340.875	252.212
-2 Log L	335.382	219.253

R-Square	0.3799	Max-rescaled R-Square	0.5076

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	116.1290	5	<.0001
Score	98.6331	5	<.0001
Wald	66.0015	5	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
9.1321	7	0.2433

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-5.5251	1.2967	18.1557	<.0001	
sex	1	-1.5975	0.3837	17.3346	<.0001	-0.4161
dum_cpt1	1	1.7411	0.5993	8.4410	0.0037	0.2454
dum_cpt2	1	2.4661	0.5435	20.5916	<.0001	0.4949
dum_cpt3	1	2.2728	0.4049	31.5162	<.0001	0.5710
maximum_heart_rate	1	0.0374	0.00873	18.3382	<.0001	0.4569

Fig 8

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	116.1290	5	<.0001
Score	98.6331	5	<.0001
Wald	66.0015	5	<.0001

Fig 9

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-7.8548	1.5506	25.6618	<.0001	
sex	1	-2.2898	0.4623	24.5278	<.0001	-0.5964
dum_cpt1	1	2.1462	0.6512	10.8625	0.0010	0.3060
dum_cpt2	1	3.7733	0.7583	24.7623	<.0001	0.7396
dum_cpt3	1	2.5989	0.4478	33.6912	<.0001	0.6577
maximum_heart_rate	1	0.0541	0.0105	26.5147	<.0001	0.6543

Fig 10

Estimated Correlation Matrix						
Parameter	Intercept	sex	dum_cpt1	dum_cpt2	dum_cpt3	maximum_heart_rate
Intercept	1.0000	0.2513	-0.1556	-0.1863	-0.2069	-0.9752
sex	0.2513	1.0000	-0.2729	-0.2738	-0.2980	-0.3905
dum_cpt1	-0.1556	-0.2729	1.0000	0.2181	0.3228	0.1234
dum_cpt2	-0.1863	-0.2738	0.2181	1.0000	0.2904	0.1628
dum_cpt3	-0.2069	-0.2980	0.3228	0.2904	1.0000	0.1462
maximum_heart_rate	-0.9752	-0.3905	0.1234	0.1628	0.1462	1.0000

Fig 11

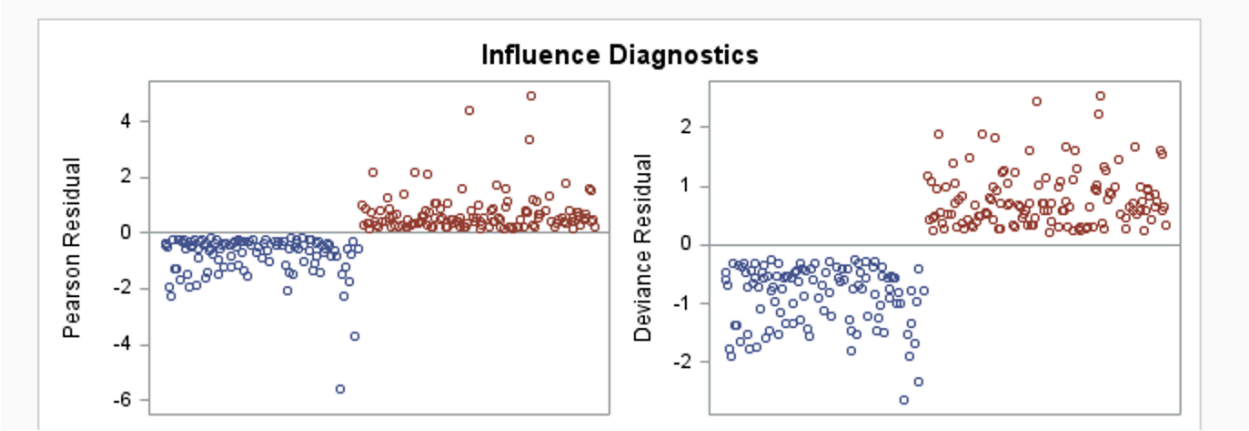

Fig 12

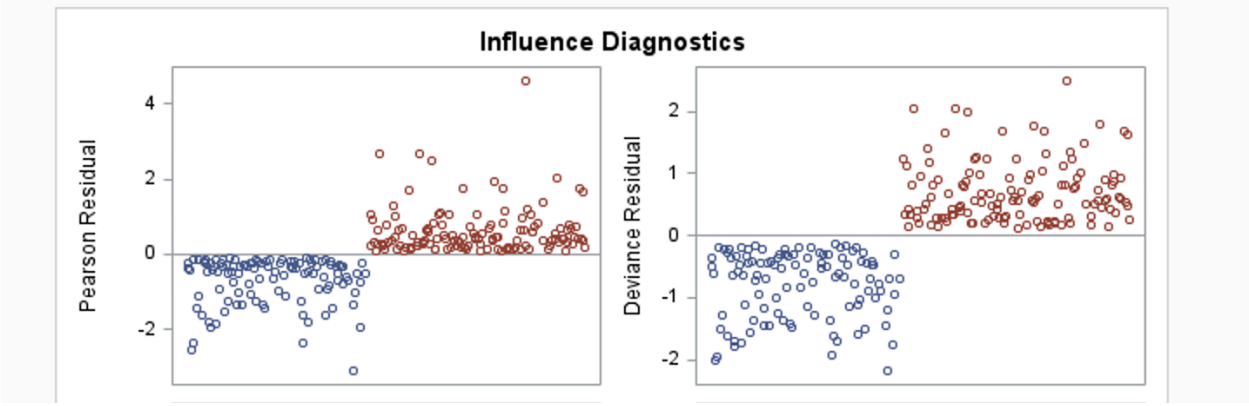

Fig 13

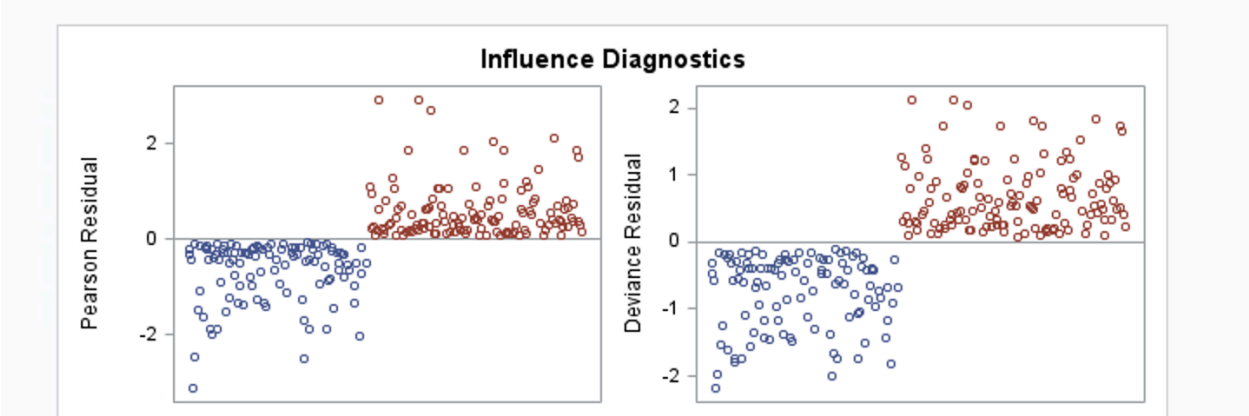

Fig 14

Fig 15

R-Square	0.7229	Max-rescaled R-Square	0.9681

Fig 16

R-Square	0.7394	Max-rescaled R-Square	0.9914

Fig 17

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
sex	<0.001	<0.001	>999.999
dum_cpt1	>999.999	<0.001	>999.999
dum_cpt2	>999.999	<0.001	>999.999
dum_cpt3	>999.999	<0.001	>999.999
maximum_heart_rate	1.295	1.080	1.552

Fig 18

R-Square	0.4611	Max-rescaled R-Square	0.6164

Fig19

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
sex	0.101	0.041	0.251
dum_cpt1	8.553	2.387	30.648
dum_cpt2	43.525	9.847	192.390
dum_cpt3	13.449	5.592	32.347
maximum_heart_rate	1.056	1.034	1.078

Fig 20

Obs	age	sex	chest_pain_type	blood_pressure	cholesterol	Fasting_blood_sugar	resting_ecg	maximum_heart_rate	angina	Selected	heart_condition	dum_cpt1
1	62	1	4	128	210	0	2	130	1	.	.	0
2	54	1	2	135	260	0	0	168	0	.	.	0

dum_cpt2	dum_cpt3	dum_r_ecg1	dum_r_ecg2	new_y	_FROM_	_INTO_	IP_0	IP_1	_LEVEL_	phat	lcl	ucl
0	0	0	1	.	.	0	0.95754	0.04246	1	0.04246	0.01736	0.10018
1	0	0	0	.	.	1	0.06226	0.93774	1	0.93774	0.79054	0.98364

Fig 21

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	Pos Pred	Neg Pred
0.200	119	63	46	9	76.8	93.0	57.8	72.1	87.5
0.250	118	68	41	10	78.5	92.2	62.4	74.2	87.2
0.300	117	70	39	11	78.9	91.4	64.2	75.0	86.4
0.350	117	72	37	11	79.7	91.4	66.1	76.0	86.7
0.400	114	77	32	14	80.6	89.1	70.6	78.1	84.6
0.450	108	79	30	20	78.9	84.4	72.5	78.3	79.8
0.500	105	81	28	23	78.5	82.0	74.3	78.9	77.9
0.550	105	84	25	23	79.7	82.0	77.1	80.8	78.5
0.600	100	87	22	28	78.9	78.1	79.8	82.0	75.7
0.650	95	90	19	33	78.1	74.2	82.6	83.3	73.2
0.700	90	95	14	38	78.1	70.3	87.2	86.5	71.4
0.750	82	99	10	46	76.4	64.1	90.8	89.1	68.3
0.800	76	101	8	52	74.7	59.4	92.7	90.5	66.0

Fig 22

Validation Test Set

The FREQ Procedure

Frequency	Table of heart_condition by pred_dis		
heart_condition	pred_dis		
	0	1	Total
0	19	7	26
1	9	25	34
Total	28	32	60

Fig 23

Code:

```
data Cardiology;
infile 'Cardiology.txt' delimiter = '09'x firstobs = 2 missover;
input age  sex   chest_pain_type blood_pressure  cholesterol
Fasting_blood_sugar  resting_ecg     maximum_heart_rate   angina
heart_condition;
dum_cpt1 = (chest_pain_type = 1);
dum_cpt2 = (chest_pain_type = 2);
dum_cpt3 = (chest_pain_type = 3);
dum_r_ecg1 = (resting_ecg = 1);
dum_r_ecg2 = (resting_ecg = 2);
run;
proc print;
run;
Title "Frequency of Heart Condition";
Proc Freq data = cardiology;
table heart_condition;
run;
PROC SORT;
BY heart_condition;
RUN;
Proc Boxplot;
Plot age * heart_condition;
RUN;
Proc Boxplot;
Plot maximum_heart_rate * heart_condition;
RUN;
Title 'Test and train set for Heart Condition';
Proc surveyselect data = cardiology out = train seed = 592587
samprate = 0.8 outall;
run;
proc print data = train;
run;
proc freq data = train;
tables selected;
run;
data train;
set train;
if selected then new_y = Heart_condition;
run;
proc logistic data = train;
model new_y  (event = '1')= age sex   dum_cpt1 dum_cpt2 dum_cpt3
blood_pressure  cholesterol     Fasting_blood_sugar   dum_r_ecg1
dum_r_ecg2 maximum_heart_rate    angina /  stb;
```

```
run;
proc logistic data = train;
model new_y  (event = '1')= age sex  dum_cpt1 dum_cpt2 dum_cpt3
blood_pressure  cholesterol     Fasting_blood_sugar  dum_r_ecg1
dum_r_ecg2 maximum_heart_rate   angina / selection = stepwise rsquare
stb;
run;
proc logistic data = train;
model new_y  (event = '1')= age sex  dum_cpt1 dum_cpt2 dum_cpt3
blood_pressure  cholesterol     Fasting_blood_sugar  dum_r_ecg1
dum_r_ecg2 maximum_heart_rate   angina / selection = forward rsquare
stb;
run;
proc logistic data = train;
model new_y  (event = '1')= age sex  dum_cpt1 dum_cpt2 dum_cpt3
blood_pressure  cholesterol     Fasting_blood_sugar  dum_r_ecg1
dum_r_ecg2 maximum_heart_rate   angina / selection = backward rsquare
stb;
run;
Title "Check of Collinearity, Outliers and Influencers";
proc logistic data = train;
model new_y  (event = '1')= sex dum_cpt1 dum_cpt2 dum_cpt3
maximum_heart_rate / influence iplots corrb stb;
run;
data train_02;
SET train;
IF _N_ IN (124,134,215,258) THEN DELETE;
RUN;
proc logistic data = train_02;
model new_y  (event = '1')=     sex  dum_cpt1 dum_cpt2 dum_cpt3
maximum_heart_rate / rsquare influence iplots corrb stb;
run;
data train_03;
SET train_02;
IF _N_ IN (126,254) THEN DELETE;
RUN;
proc logistic data = train_03;
model new_y  (event = '1')= sex dum_cpt1 dum_cpt2 dum_cpt3
maximum_heart_rate /  rsquare influence iplots corrb stb;
run;
*deleting influential points;
data train_04;
SET train_03;
IF _N_ IN
(4,5,8,9,12,17,18,23,29,31,38,42,48,52,55,57,59,74,86,87,88,91,99,105,
```

```
110,121,125,129,137,138,139,144,149,154,155,166,174,183,188,189,191,19
6,207,217,230,237,238,251,251,254,258,262,276,293,295) THEN DELETE;
RUN;
proc logistic data = train_04;
model new_y  (event = '1')= sex dum_cpt1 dum_cpt2 dum_cpt3
maximum_heart_rate /  rsquare influence iplots corrb stb;
run;
data train_05;
SET train_04;
IF _N_ IN (28,83,99,214) THEN DELETE;
RUN;
proc logistic data = train_05;
model new_y  (event = '1')= sex dum_cpt1 dum_cpt2 dum_cpt3
maximum_heart_rate /  rsquare influence iplots corrb stb;
run;
Title "Final Model" ;
proc logistic data = train_03;
model new_y  (event = '1')= sex dum_cpt1 dum_cpt2 dum_cpt3
maximum_heart_rate / rsquare stb;
run;
Title "Predictions";
data new ;
input age  sex   chest_pain_type      blood_pressure  cholesterol
Fasting_blood_sugar    resting_ecg    maximum_heart_rate    angina;
datalines;
62    1     4     128   210   0     2     130   1
54    1     2     135   260   0     0     168   0
;
data pred;
set new train_03;
dum_cpt1 = (chest_pain_type = 1);
dum_cpt2 = (chest_pain_type = 2);
dum_cpt3 = (chest_pain_type = 3);
dum_r_ecg1 = (resting_ecg = 1);
dum_r_ecg2 = (resting_ecg = 2);
run;
proc logistic;
model new_y  (event = '1')=     sex   dum_cpt1 dum_cpt2 dum_cpt3
maximum_heart_rate;
output out = pred p = phat lower = lcl  upper = ucl predprobs =
(individual);
run;
proc print data = pred;
Title " Predicted Probabilities";
run;
```

```
proc freq data = train_03;
tables new_y;
run;
Title 'Validation Test Set ';
proc logistic data = train_03;
model new_y  (event = '1')=     sex   dum_cpt1 dum_cpt2 dum_cpt3
maximum_heart_rate /ctable pprob = (0.2 to 0.8 by 0.05)     ;
output out = pred_01 (where = (new_y = .)) p = phat lower = lcl
upper = ucl predprobs = (individual);
run;
data probs;
set pred_01;
pred_dis = 0;
threshold = 0.4;
if phat>threshold then pred_dis = 1;
run;
proc freq data= probs;
tables heart_condition*pred_dis / norow nocol nopercent;
run;
```