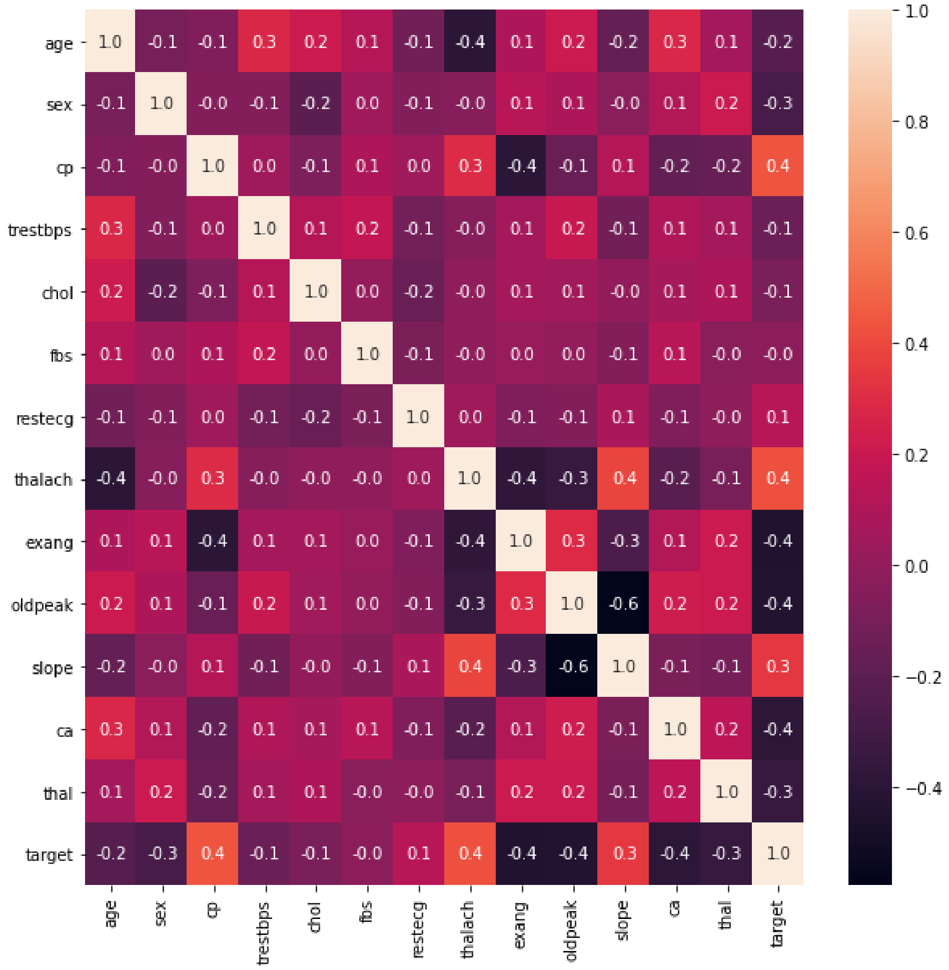
Curiosity project - Group 06 - Heart Disease

Submitters:
Tal Carmi, ID: 039161203
Anna Mosenzon, ID: 200320836
Rami Skolozub, ID: 316736396

Part I:

1. The Learning problem:
   a. a classification problem to distinguish between presence or absence of heart disease based on dataset contains 13 features and 1 Label column (there is a heart disease or there is no heart disease.
   The target variable: Heart Disease – Yes\No distributes as follows:
   54% of the observations have heart disease and 46% have no heart disease.

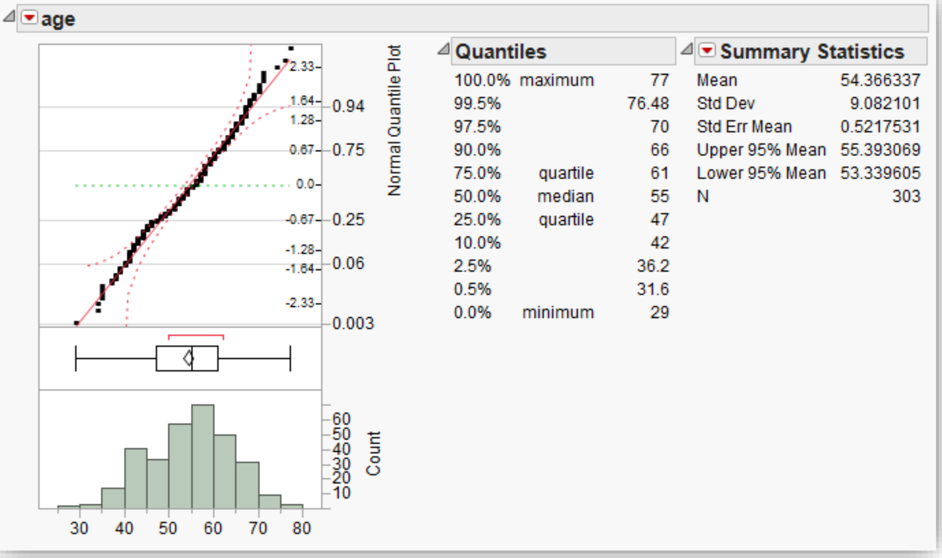
   Due to low correlation between the following features and the target we removed them from the data set (colored grey in the table below):
   - trestbps
   - chol
   - fbs
   - restecg



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.0	-0.1	-0.1	0.3	0.2	0.1	-0.1	-0.4	0.1	0.2	-0.2	0.3	0.1	-0.2
sex	-0.1	1.0	-0.0	-0.1	-0.2	0.0	-0.1	-0.0	0.1	0.1	-0.0	0.1	0.2	-0.3
cp	-0.1	-0.0	1.0	0.0	-0.1	0.1	0.0	0.3	-0.4	-0.1	0.1	-0.2	-0.2	0.4
trestbps	0.3	-0.1	0.0	1.0	0.1	0.2	-0.1	-0.0	0.1	0.2	-0.1	0.1	0.1	-0.1
chol	0.2	-0.2	-0.1	0.1	1.0	0.0	-0.2	-0.0	0.1	0.1	-0.0	0.1	0.1	-0.1
fbs	0.1	0.0	0.1	0.2	0.0	1.0	-0.1	-0.0	0.0	0.0	-0.1	0.1	-0.0	-0.0
restecg	-0.1	-0.1	0.0	-0.1	-0.2	-0.1	1.0	0.0	-0.1	-0.1	0.1	-0.1	-0.0	0.1
thalach	-0.4	-0.0	0.3	-0.0	-0.0	-0.0	0.0	1.0	-0.4	-0.3	0.4	-0.2	-0.1	0.4
exang	0.1	0.1	-0.4	0.1	0.1	0.0	-0.1	-0.4	1.0	0.3	-0.3	0.1	0.2	-0.4
oldpeak	0.2	0.1	-0.1	0.2	0.1	0.0	-0.1	-0.3	0.3	1.0	-0.6	0.2	0.2	-0.4
slope	-0.2	-0.0	0.1	-0.1	-0.0	-0.1	0.1	0.4	-0.3	-0.6	1.0	-0.1	-0.1	0.3
ca	0.3	0.1	-0.2	0.1	0.1	0.1	-0.1	-0.2	0.1	0.2	-0.1	1.0	0.2	-0.4
thal	0.1	0.2	-0.2	0.1	0.1	-0.0	-0.0	-0.1	0.2	0.2	-0.1	0.2	1.0	-0.3
target	-0.2	-0.3	0.4	-0.1	-0.1	-0.0	0.1	0.4	-0.4	-0.4	0.3	-0.4	-0.3	1.0

Column Name	Description	Comment	Data Type
age	Age in years		numeric
sex	The person's sex	1 = male, 0 = female	categorical
cp	The chest pain experienced (4 values)	Value 0: typical angina, Value 1: atypical angina, Value 2: non-anginal pain, Value 3: asymptomatic	categorical
trestbps	The person's resting blood pressure	mm Hg on admission to the hospital	numeric
chol	The person's cholesterol measurement in mg/dl		numeric
fbs	The person's fasting blood sugar	> 120 mg/dl, 1 = true; 0 = false	categorical
restecg	Resting electrocardiographic measurement	0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria	categorical
thalach	The person's maximum heart rate achieved		numeric
exang	Exercise induced angina	1 = yes; 0 = no	categorical
oldpeak	ST depression induced by exercise relative to rest	ST' relates to positions on the ECG plot	Float
slope	the slope of the peak exercise ST segment	Value 0: upsloping, Value 1: flat, Value 2: down sloping	categorical
ca	number of major vessels colored by fluoroscopy		numeric
thal	A blood disorder called thalassemia	0 = null; 1 = fixed defect; 2 = normal; 3 = reversable defect	categorical
target	**Heart disease**	**0 = no, 1 = yes**	**categorical**

b. In this dataset we have 303 rows, 13 features and one label column called “target”. Below are the distributions of each feature that was included in the calculations:



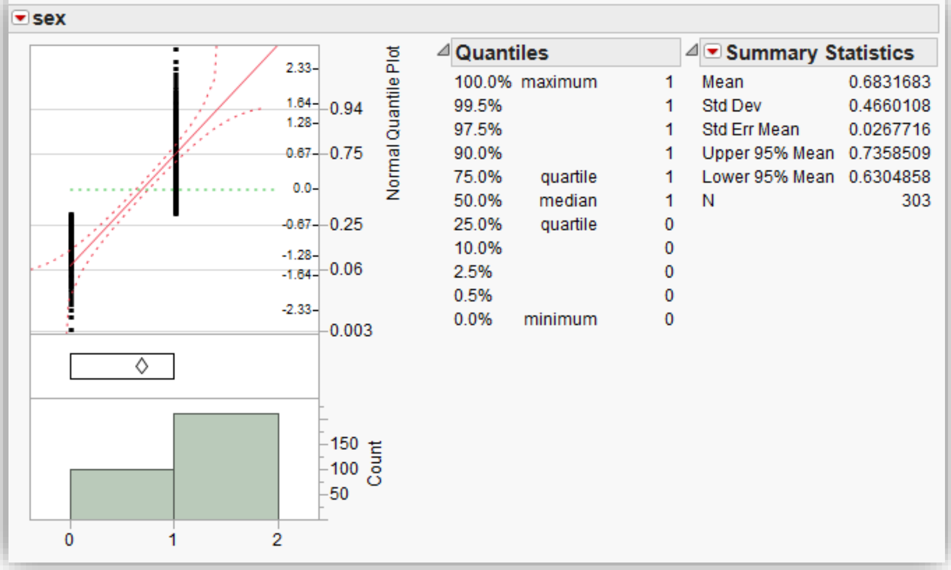
age

Quantiles

100.0%	maximum	77
99.5%		76.48
97.5%		70
90.0%		66
75.0%	quartile	61
50.0%	median	55
25.0%	quartile	47
10.0%		42
2.5%		36.2
0.5%		31.6
0.0%	minimum	29

Summary Statistics

Mean	54.366337
Std Dev	9.082101
Std Err Mean	0.5217531
Upper 95% Mean	55.393069
Lower 95% Mean	53.339605
N	303



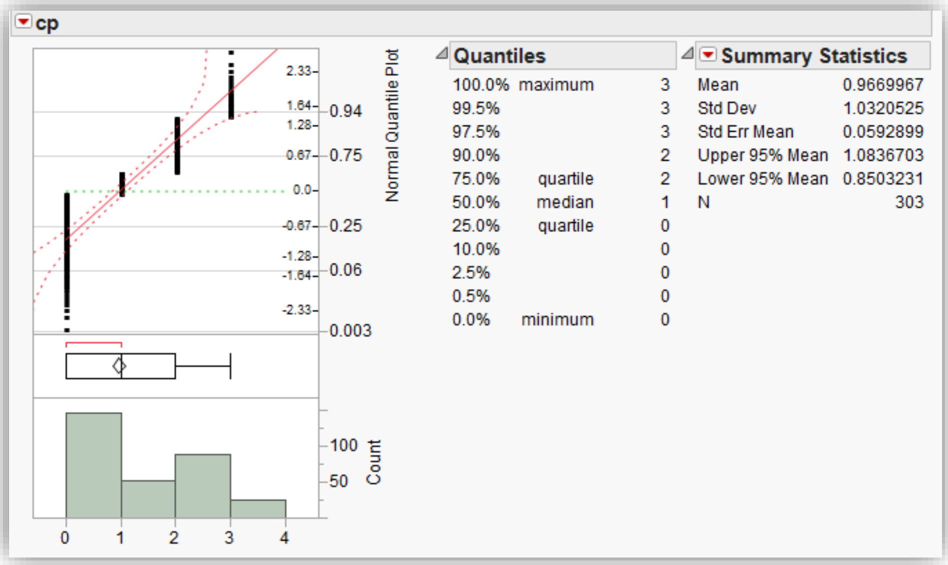
sex

Quantiles

100.0%	maximum	1
99.5%		1
97.5%		1
90.0%		1
75.0%	quartile	1
50.0%	median	1
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	0.6831683
Std Dev	0.4660108
Std Err Mean	0.0267716
Upper 95% Mean	0.7358509
Lower 95% Mean	0.6304858
N	303



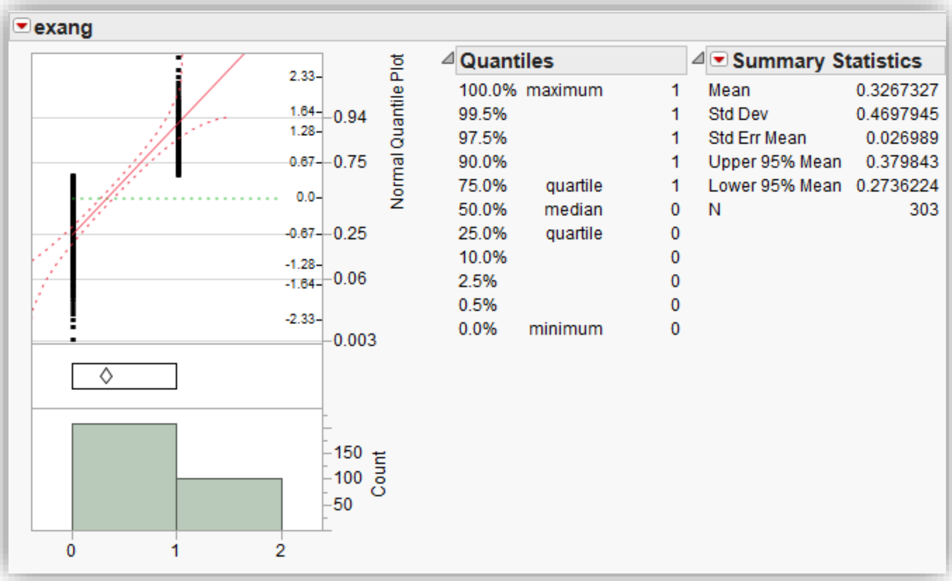
cp

Quantiles

100.0%	maximum	3
99.5%		3
97.5%		3
90.0%		2
75.0%	quartile	2
50.0%	median	1
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	0.9669967
Std Dev	1.0320525
Std Err Mean	0.0592899
Upper 95% Mean	1.0836703
Lower 95% Mean	0.8503231
N	303



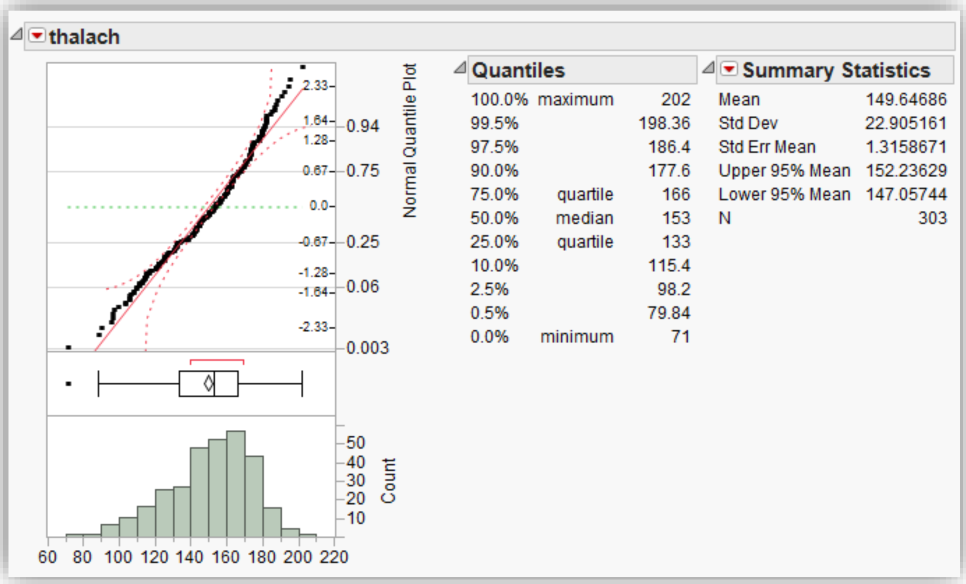
exang

Quantiles

100.0%	maximum	1
99.5%		1
97.5%		1
90.0%		1
75.0%	quartile	1
50.0%	median	0
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	0.3267327
Std Dev	0.4697945
Std Err Mean	0.026989
Upper 95% Mean	0.379843
Lower 95% Mean	0.2736224
N	303



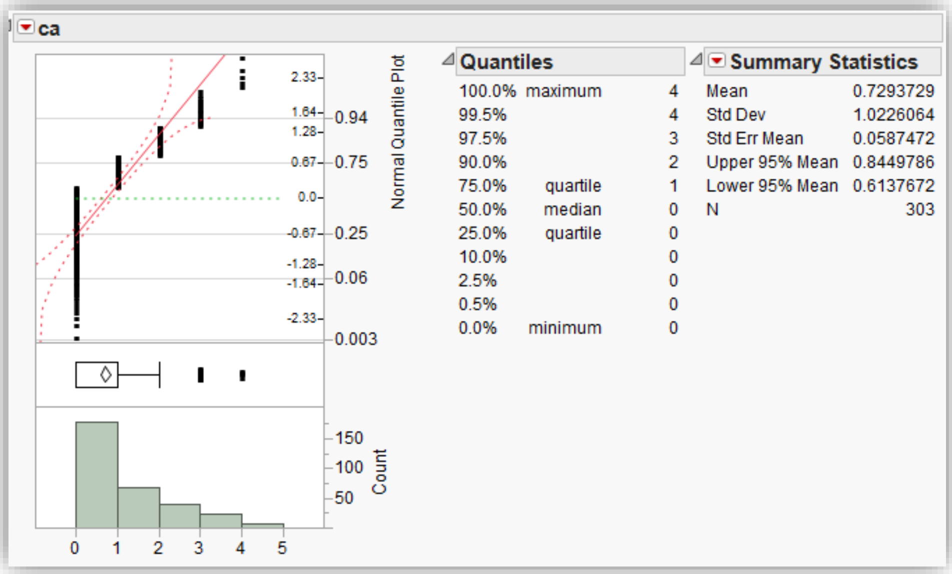
thalach

Quantiles

100.0%	maximum	202
99.5%		198.36
97.5%		186.4
90.0%		177.6
75.0%	quartile	166
50.0%	median	153
25.0%	quartile	133
10.0%		115.4
2.5%		98.2
0.5%		79.84
0.0%	minimum	71

Summary Statistics

Mean	149.64686
Std Dev	22.905161
Std Err Mean	1.3158671
Upper 95% Mean	152.23629
Lower 95% Mean	147.05744
N	303



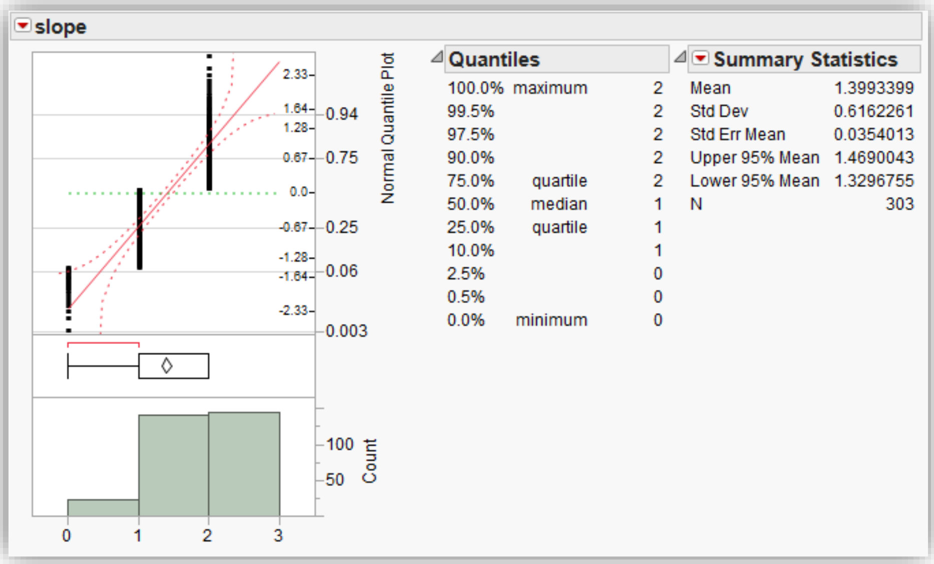
ca

Quantiles

100.0%	maximum	4
99.5%		4
97.5%		3
90.0%		2
75.0%	quartile	1
50.0%	median	0
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	0.7293729
Std Dev	1.0226064
Std Err Mean	0.0587472
Upper 95% Mean	0.8449786
Lower 95% Mean	0.6137672
N	303

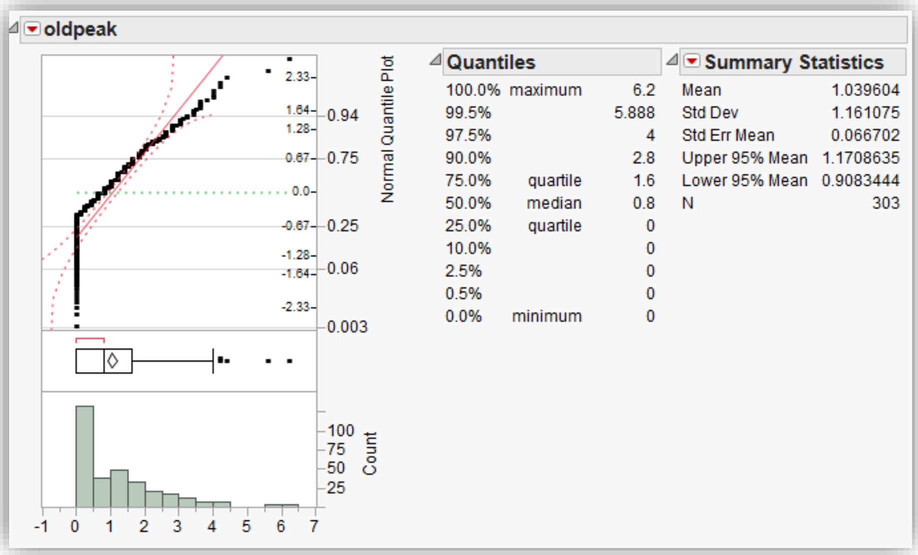


Quantiles

100.0%	maximum	2
99.5%		2
97.5%		2
90.0%		2
75.0%	quartile	2
50.0%	median	1
25.0%	quartile	1
10.0%		1
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	1.3993399
Std Dev	0.6162261
Std Err Mean	0.0354013
Upper 95% Mean	1.4690043
Lower 95% Mean	1.3296755
N	303

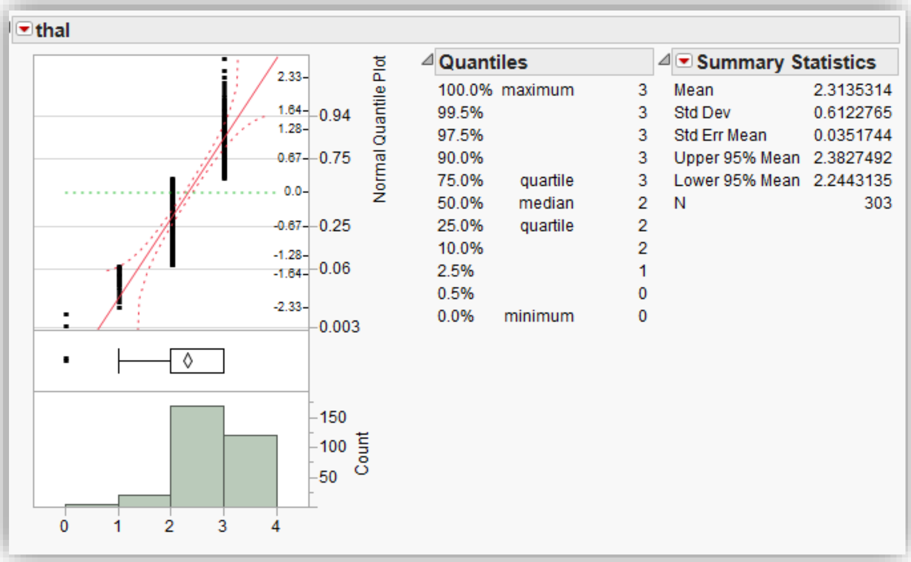


Quantiles

100.0%	maximum	6.2
99.5%		5.888
97.5%		4
90.0%		2.8
75.0%	quartile	1.6
50.0%	median	0.8
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	1.039604
Std Dev	1.161075
Std Err Mean	0.066702
Upper 95% Mean	1.1708635
Lower 95% Mean	0.9083444
N	303



Quantiles

100.0%	maximum	3
99.5%		3
97.5%		3
90.0%		3
75.0%	quartile	3
50.0%	median	2
25.0%	quartile	2
10.0%		2
2.5%		1
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	2.3135314
Std Dev	0.6122765
Std Err Mean	0.0351744
Upper 95% Mean	2.3827492
Lower 95% Mean	2.2443135
N	303

2. A. problem formulation with Bayesian inference:

    i. The parameters of the Posterior are the mean and STD for Normal distribution data for each target value – 0 and 1 as follows:

        i. Target = 0

	Mean	STD
age	56.60145	9.067102
sex	0.826087	0.465241
cp	0.478261	1.030348
thalach	139.1014	22.86733
exang	0.550725	0.469019
oldpeak	1.585507	1.159157
slope	1.166667	0.615208
ca	1.166667	1.020918
thal	2.543478	0.611265

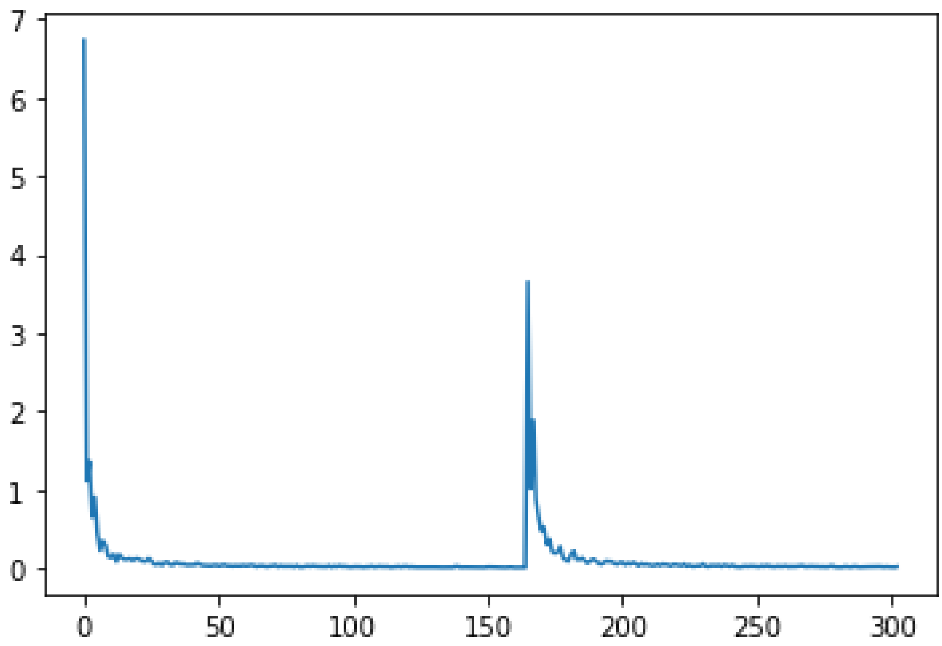
        ii. Target = 1

	Mean	STD
age	52.49697	9.067102
sex	0.563636	0.465241
cp	1.375758	1.030348
thalach	158.4667	22.86733
exang	0.139394	0.469019
oldpeak	0.58303	1.159157
slope	1.593939	0.615208
ca	0.363636	1.020918
thal	2.121212	0.611265

    i. The Prior for each variable are the μ and the σ according to the data for both target value – 0 and 1 as follows:

	μ	σ
age	54.36634	9.067102
sex	0.683168	0.465241
cp	0.966997	1.030348
thalach	149.6469	22.86733
exang	0.326733	0.469019
oldpeak	1.039604	1.159157
slope	1.39934	0.615208
ca	0.729373	1.020918
thal	2.313531	0.611265

3. a. ii. The information gain graph is:



b.i. You can see that there is a jump in the information gain after 165 samples, the reason is because our data is organize so that the first 165 samples are the “1” target value (have heart disease) while the rest of the 138 samples are the “0” target value (have no heart disease). So, we would expect that if we order the data so that there will “1” target value and “0” target value alternately we may gain the most of the information quicker.