HEART DISEASE

SAS PROJECT

Li Wu

Feb 8, 2025

OBJECTIVES

I. Identify Key Risk Factors

Analyze features contributing to heart disease to understand their influence.

II. Develop a Predictive Model

Build a reliable model for early detection, enabling timely medical interventions.

AGENDA

- Data Description
- Univariate Analysis
- Bivariate Analysis
- Feature Engineering
- Predictive Modeling
- Conclusions
- Appendix (Statistic Tests)

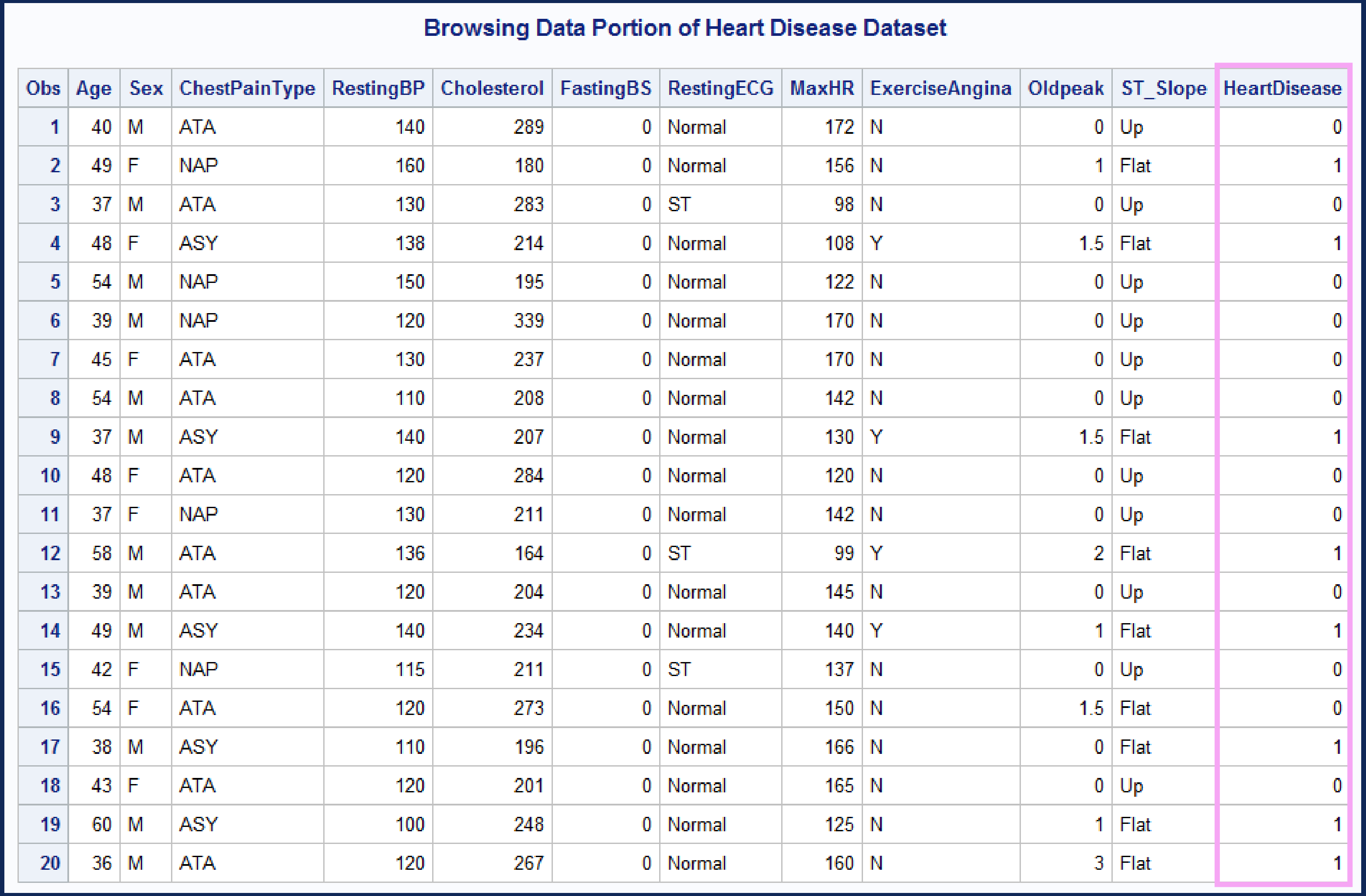
DATA DESCRIPTION

Total: 918 observations | 11 features

Heart failure is a common event caused by Cardiovascular diseases (CVDs) and this dataset contains 11 features that can be used to predict a possible heart disease.

DATA DESCRIPTION

- Categorical Variables: 6
  - Sex
  - ChestPainType
  - FastingBS
  - RestingECG
  - ExerciseAngina
  - ST_Slope
- Numerical Variables: 5
  - Age
  - RestingBP
  - Cholesterol
  - MaxHR
  - Oldpeak
- Target: HeartDisease
- No Duplicates
- No Missing Values



Browsing Data Portion of Heart Disease Dataset

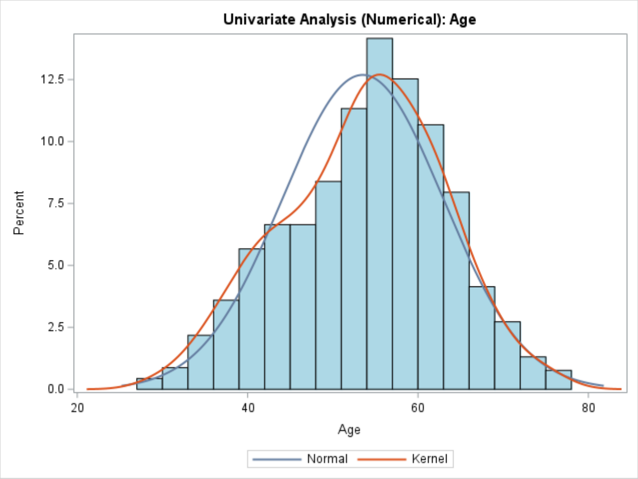
Obs	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
1	40	M	ATA	140	289	0	Normal	172	N	0	Up	0
2	49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
3	37	M	ATA	130	283	0	ST	98	N	0	Up	0
4	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
5	54	M	NAP	150	195	0	Normal	122	N	0	Up	0
6	39	M	NAP	120	339	0	Normal	170	N	0	Up	0
7	45	F	ATA	130	237	0	Normal	170	N	0	Up	0
8	54	M	ATA	110	208	0	Normal	142	N	0	Up	0
9	37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
10	48	F	ATA	120	284	0	Normal	120	N	0	Up	0
11	37	F	NAP	130	211	0	Normal	142	N	0	Up	0
12	58	M	ATA	136	164	0	ST	99	Y	2	Flat	1
13	39	M	ATA	120	204	0	Normal	145	N	0	Up	0
14	49	M	ASY	140	234	0	Normal	140	Y	1	Flat	1
15	42	F	NAP	115	211	0	ST	137	N	0	Up	0
16	54	F	ATA	120	273	0	Normal	150	N	1.5	Flat	0
17	38	M	ASY	110	196	0	Normal	166	N	0	Flat	1
18	43	F	ATA	120	201	0	Normal	165	N	0	Up	0
19	60	M	ASY	100	248	0	Normal	125	N	1	Flat	1
20	36	M	ATA	120	267	0	Normal	160	N	3	Flat	1

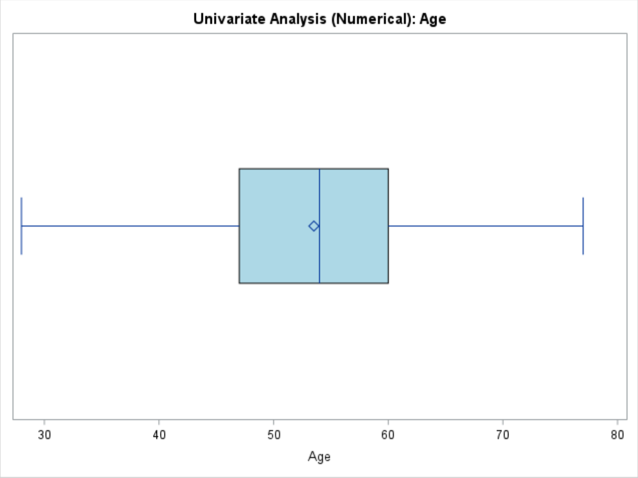
Data Source: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction?resource=download

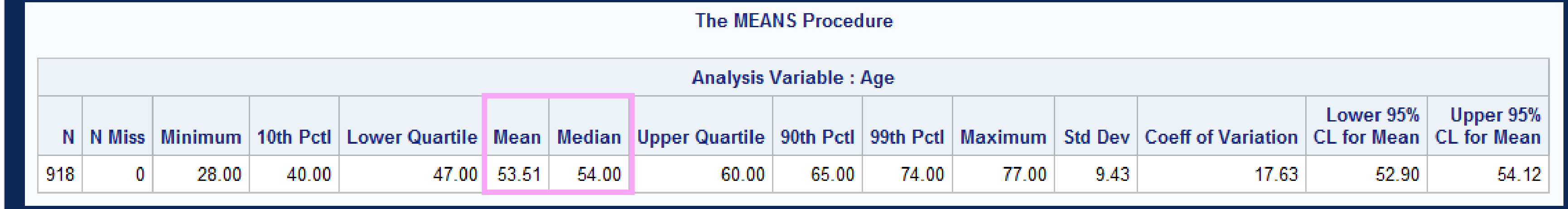
UNIVARIATE ANALYSIS

NUMERICAL VARIABLES

AGE – Not Normally Distributed



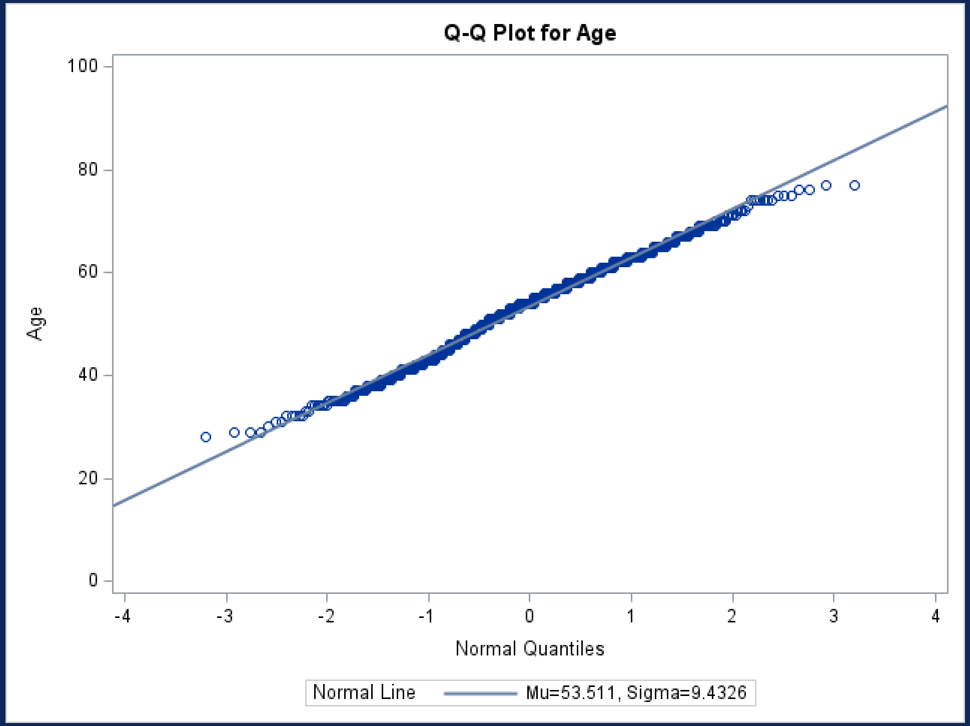


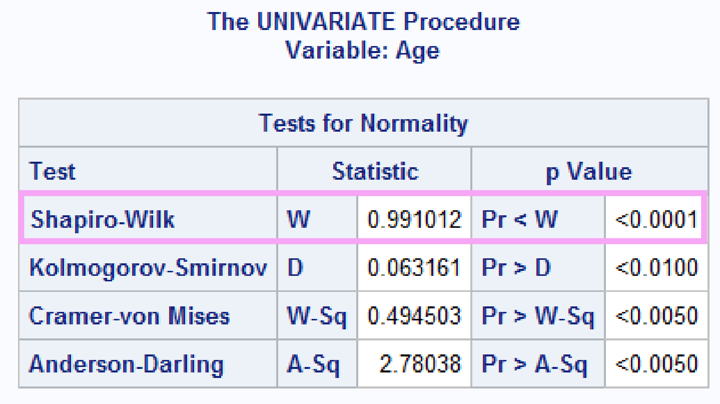


The MEANS Procedure

Analysis Variable : Age

N	N Miss	Minimum	10th Pctl	Lower Quartile	Mean	Median	Upper Quartile	90th Pctl	99th Pctl	Maximum	Std Dev	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean
918	0	28.00	40.00	47.00	53.51	54.00	60.00	65.00	74.00	77.00	9.43	17.63	52.90	54.12





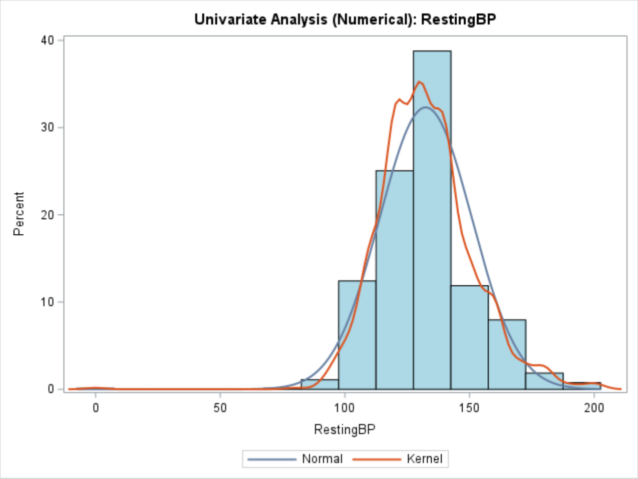
The UNIVARIATE Procedure
Variable: Age

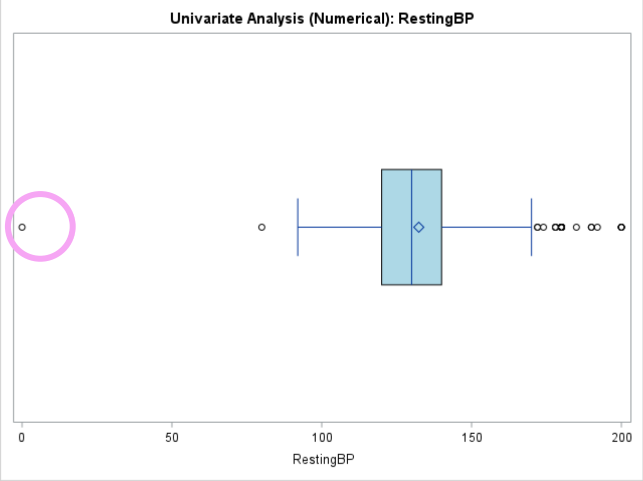
Tests for Normality

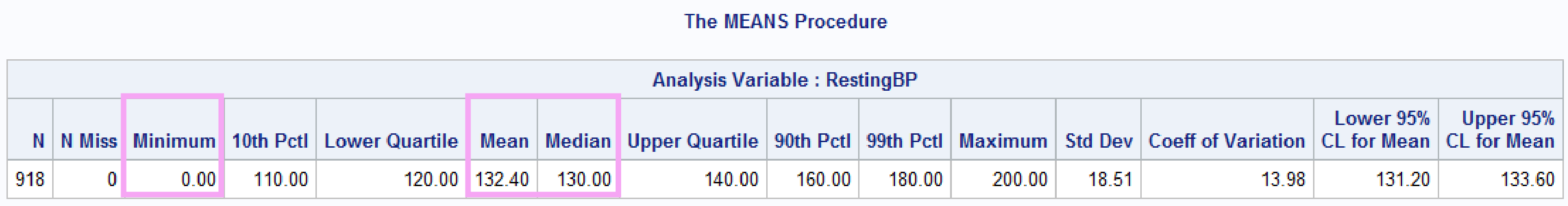
Test	Statistic		p Value	
Shapiro-Wilk	W	0.991012	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.063161	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.494503	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.78038	Pr > A-Sq	<0.0050

Turkey Method (1.5*IQR | 3*IQR): 0 obs | 0 obs
Z-Score (1.96 Std | 3 Std): 51 obs | 0 obs

RESTINGBP (Resting Blood Pressure) – Not Normally Distributed

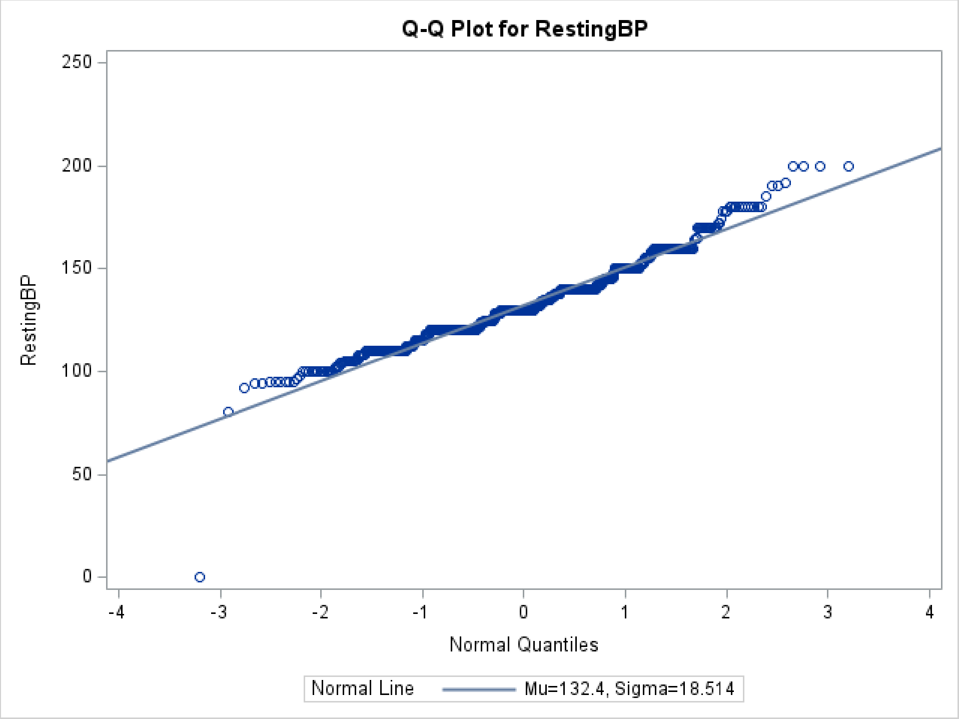


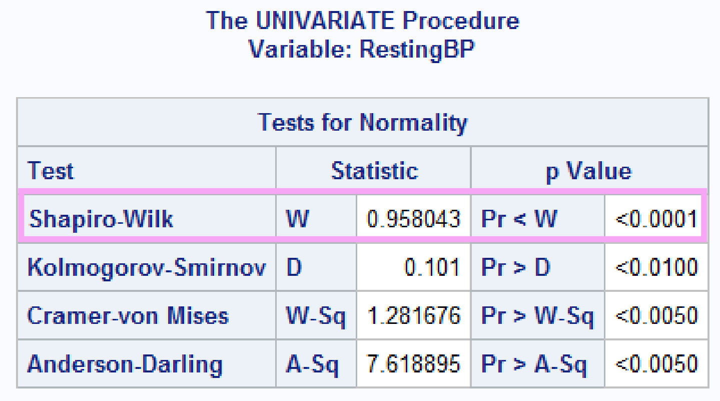



The MEANS Procedure

Analysis Variable : RestingBP

N	N Miss	Minimum	10th Pctl	Lower Quartile	Mean	Median	Upper Quartile	90th Pctl	99th Pctl	Maximum	Std Dev	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean
918	0	0.00	110.00	120.00	132.40	130.00	140.00	160.00	180.00	200.00	18.51	13.98	131.20	133.60



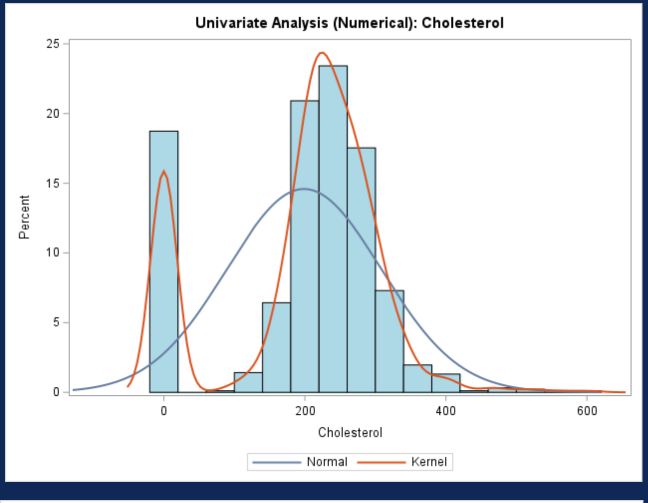

The UNIVARIATE Procedure
Variable: RestingBP

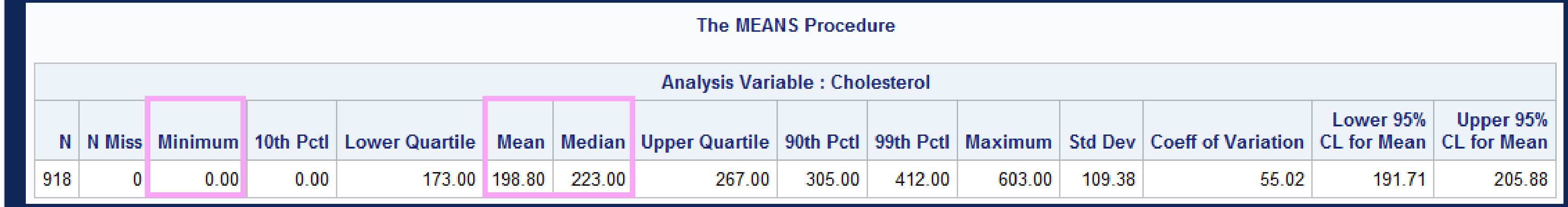
Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.958043	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.101	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.281676	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	7.618895	Pr > A-Sq	<0.0050

Turkey Method (1.5*IQR | 3*IQR): 28 obs | 1 obs
Z-Score (1.96 Std | 3 Std): 52 obs | 8 obs

CHOLESTEROL – Not Normally Distributed

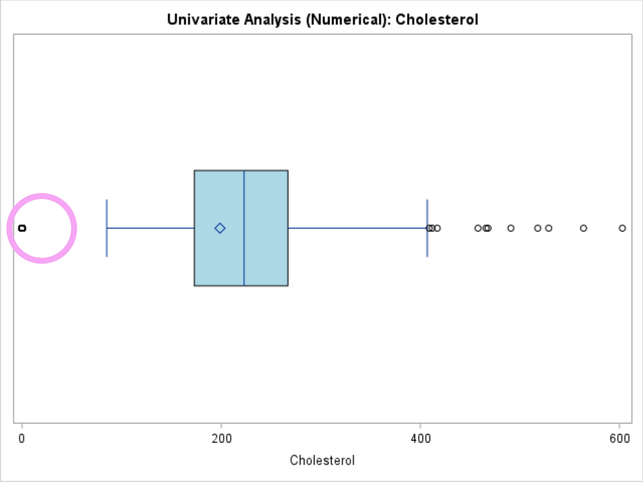


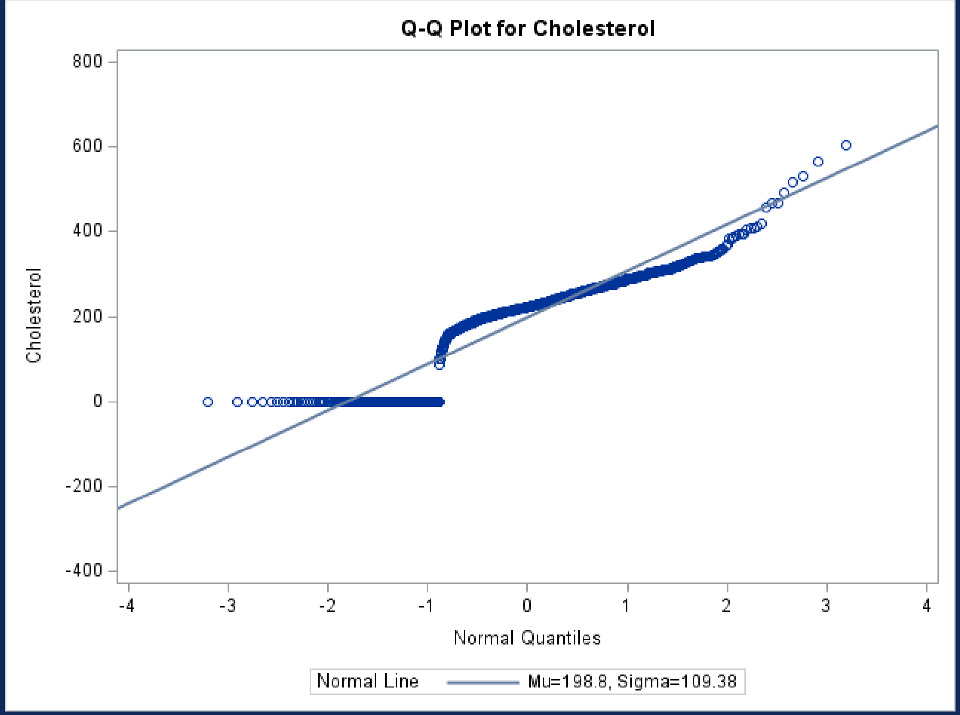


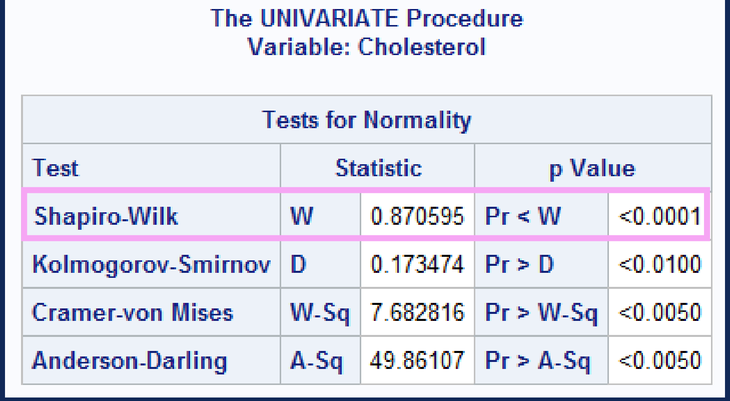
The MEANS Procedure

Analysis Variable : Cholesterol

N	N Miss	Minimum	10th Pctl	Lower Quartile	Mean	Median	Upper Quartile	90th Pctl	99th Pctl	Maximum	Std Dev	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean
918	0	0.00	0.00	173.00	198.80	223.00	267.00	305.00	412.00	603.00	109.38	55.02	191.71	205.88







The UNIVARIATE Procedure
Variable: Cholesterol

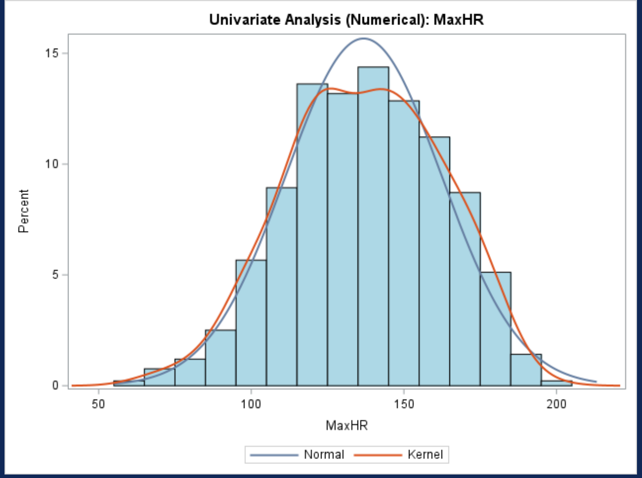
Tests for Normality

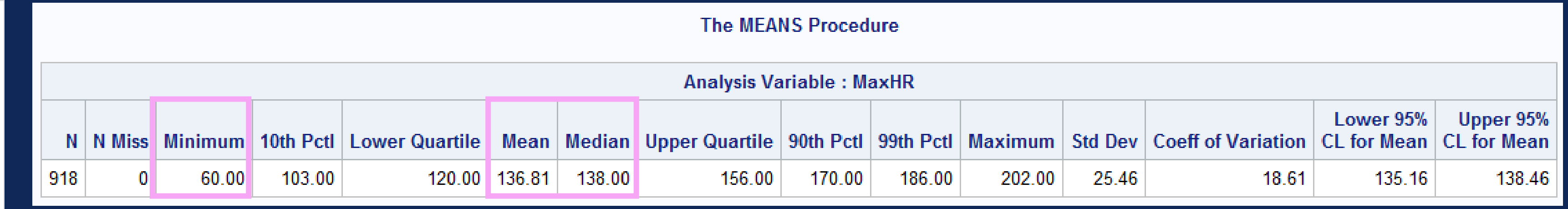
Test	Statistic		p Value	
Shapiro-Wilk	W	0.870595	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.173474	Pr > D	<0.0100
Cramer-von Mises	W-Sq	7.682816	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	49.86107	Pr > A-Sq	<0.0050

Turkey Method (1.5*IQR | 3*IQR): 183 obs | 2 obs

Z-Score (1.96 Std | 3 Std): 9 obs | 3 obs

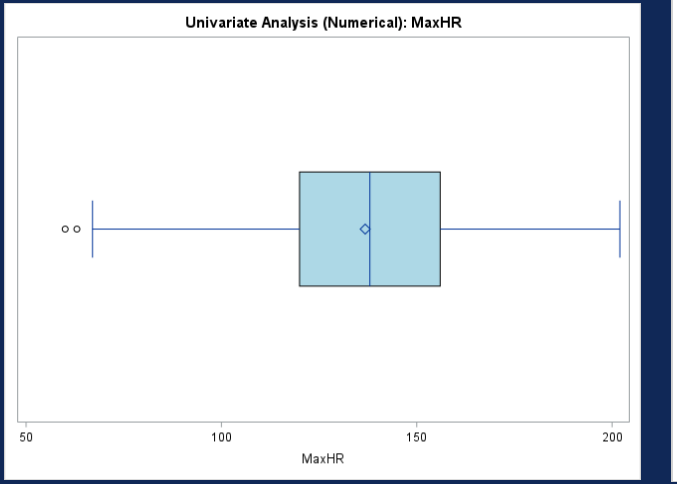
MAXHR – Not Normally Distributed

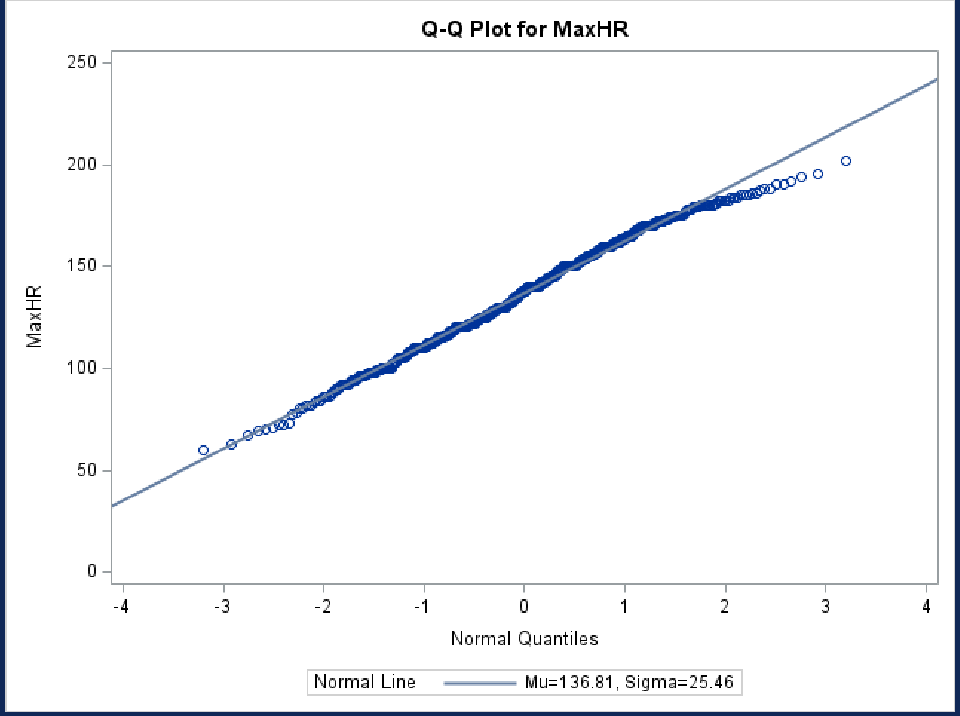

Univariate Analysis (Numerical): MaxHR

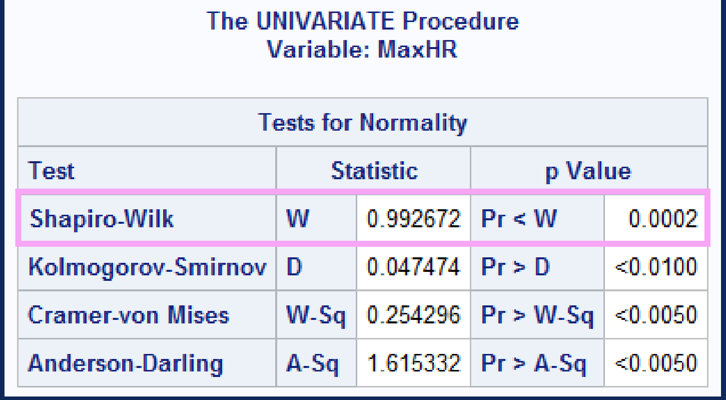

The MEANS Procedure

Analysis Variable : MaxHR

N	N Miss	Minimum	10th Pctl	Lower Quartile	Mean	Median	Upper Quartile	90th Pctl	99th Pctl	Maximum	Std Dev	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean
918	0	60.00	103.00	120.00	136.81	138.00	156.00	170.00	186.00	202.00	25.46	18.61	135.16	138.46


Univariate Analysis (Numerical): MaxHR


Q-Q Plot for MaxHR

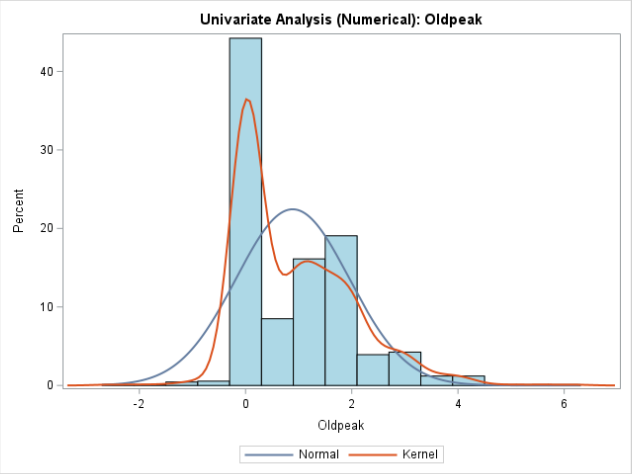

The UNIVARIATE Procedure
Variable: MaxHR

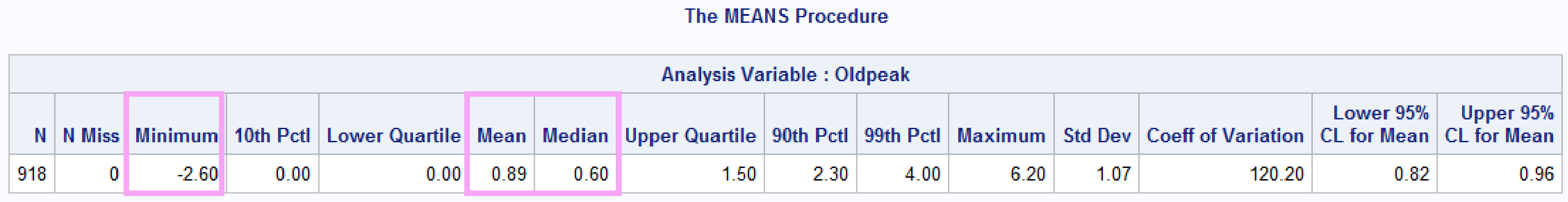
Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.992672	Pr < W	0.0002
Kolmogorov-Smirnov	D	0.047474	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.254296	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.615332	Pr > A-Sq	<0.0050

Turkey Method (1.5*IQR | 3*IQR):	2 obs	|	0 obs
Z-Score (1.96 Std | 3 Std):	33 obs	|	1 obs

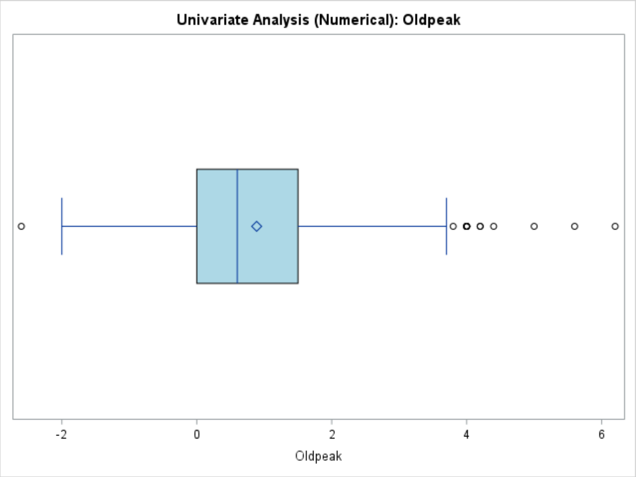
OLDPEAK – Not Normally Distributed

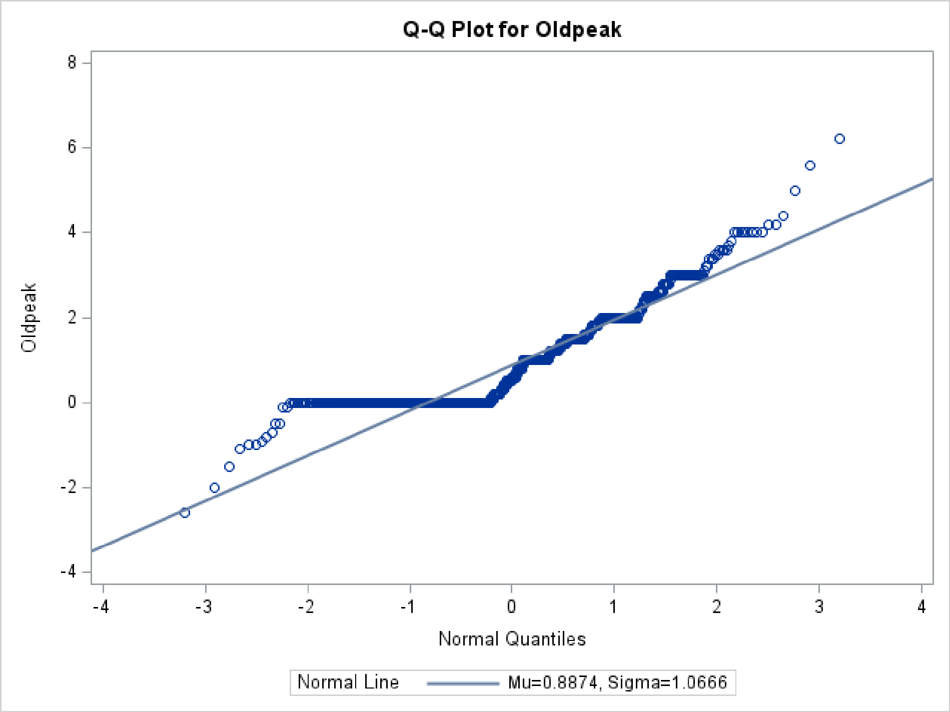

Univariate Analysis (Numerical): Oldpeak

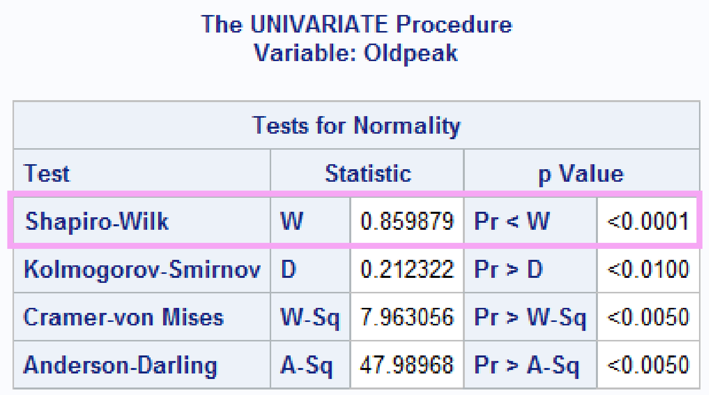

The MEANS Procedure

Analysis Variable : Oldpeak

N	N Miss	Minimum	10th Pctl	Lower Quartile	Mean	Median	Upper Quartile	90th Pctl	99th Pctl	Maximum	Std Dev	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean
918	0	-2.60	0.00	0.00	0.89	0.60	1.50	2.30	4.00	6.20	1.07	120.20	0.82	0.96


Univariate Analysis (Numerical): Oldpeak


Q-Q Plot for Oldpeak


The UNIVARIATE Procedure
Variable: Oldpeak

Tests for Normality

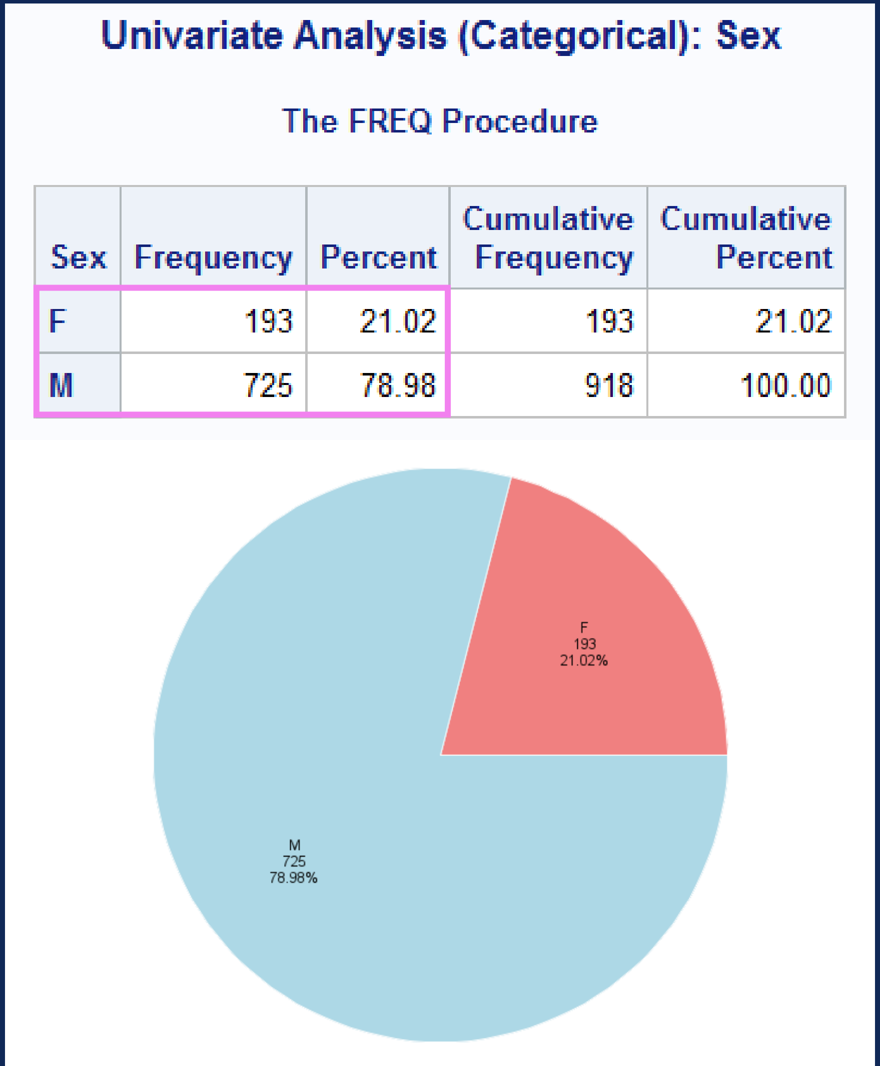
Test	Statistic		p Value	
Shapiro-Wilk	W	0.859879	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.212322	Pr > D	<0.0100
Cramer-von Mises	W-Sq	7.963056	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	47.98968	Pr > A-Sq	<0.0050

Turkey Method (1.5*IQR | 3*IQR): 16 obs | 1 obs
Z-Score (1.96 Std | 3 Std): 59 obs | 7 obs

UNIVARIATE ANALYSIS

CATEGORICAL VARIABLES

SEX

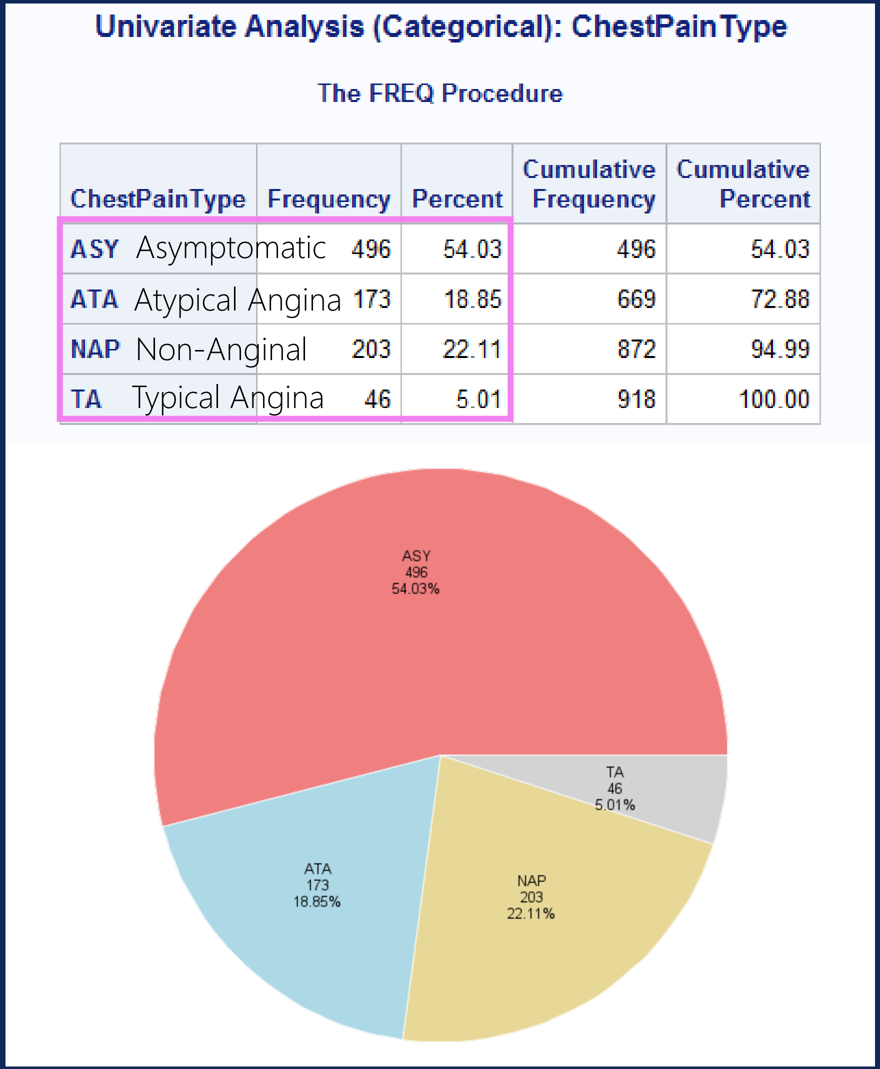


Univariate Analysis (Categorical): Sex

The FREQ Procedure

Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	193	21.02	193	21.02
M	725	78.98	918	100.00

CHESTPAINTYPE

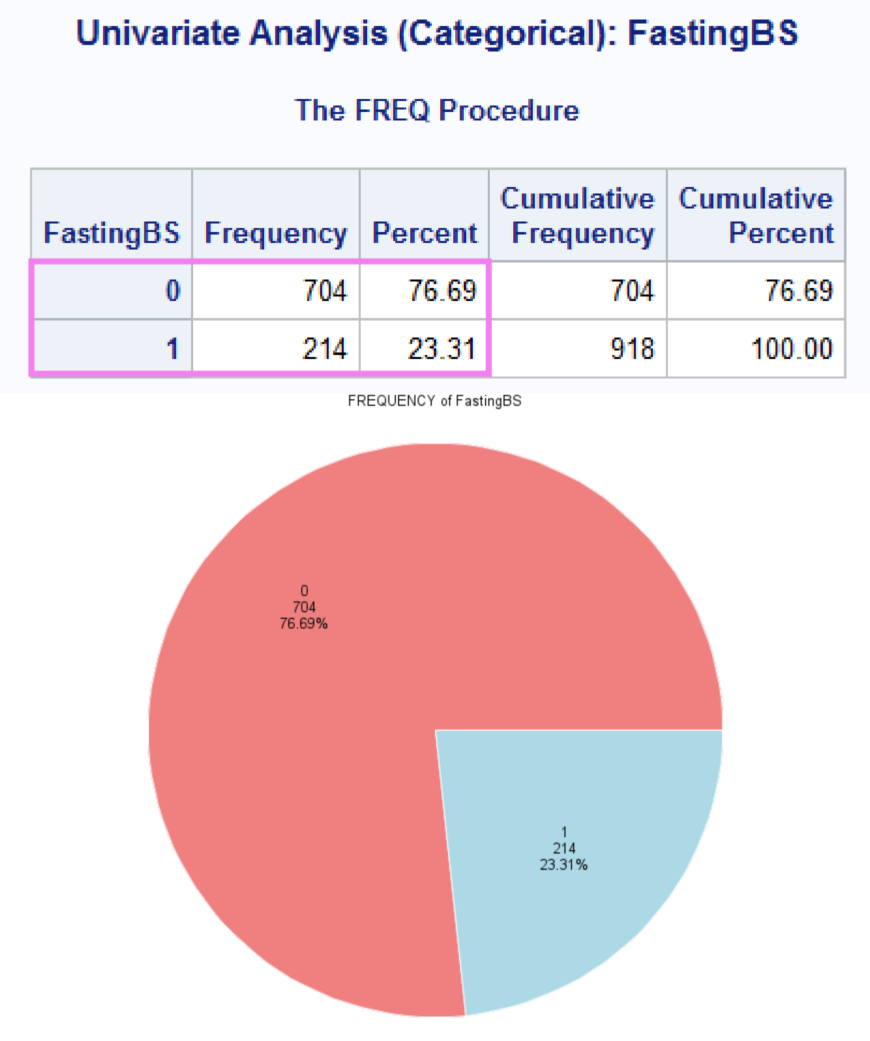


Univariate Analysis (Categorical): ChestPainType

The FREQ Procedure

ChestPainType	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ASY Asymptomatic	496	54.03	496	54.03
ATA Atypical Angina	173	18.85	669	72.88
NAP Non-Anginal	203	22.11	872	94.99
TA Typical Angina	46	5.01	918	100.00

FASTINGBS



Univariate Analysis (Categorical): FastingBS

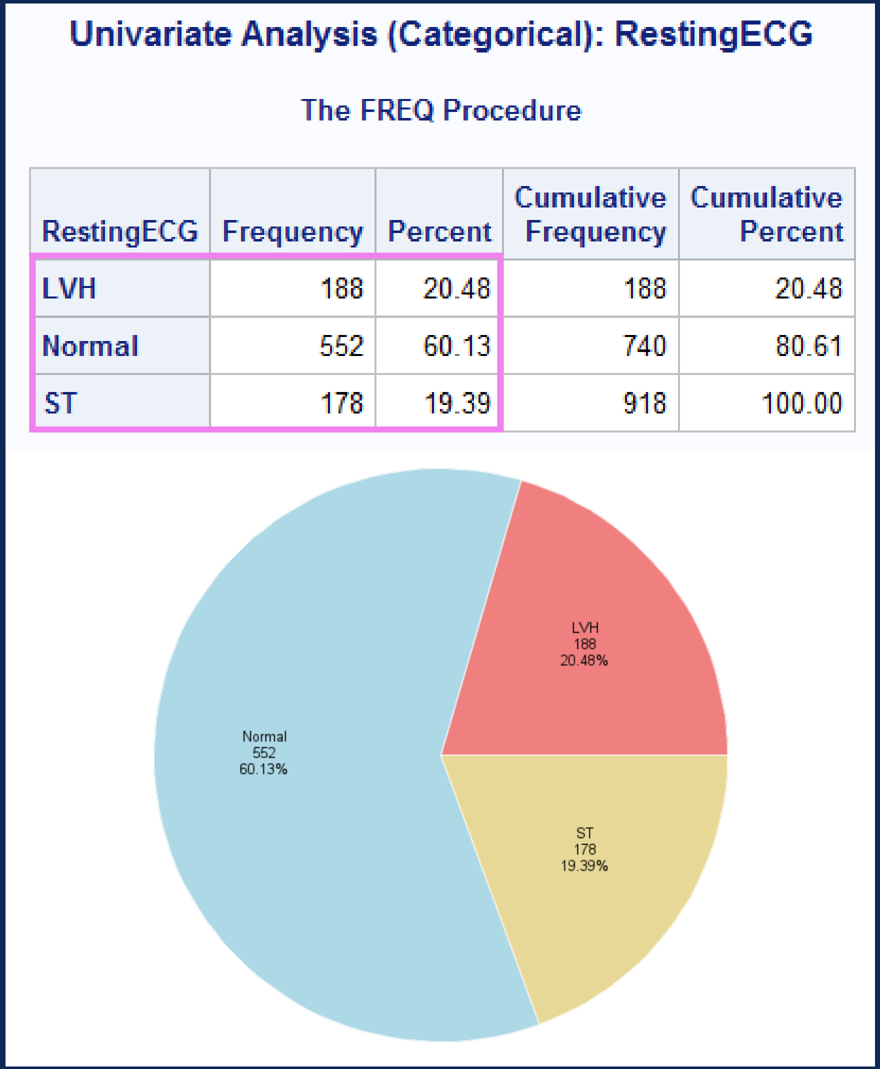
The FREQ Procedure

FastingBS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	704	76.69	704	76.69
1	214	23.31	918	100.00

FREQUENCY of FastingBS

1: FastingBS > 120 mg/dl 0: otherwise

RESTINGECG



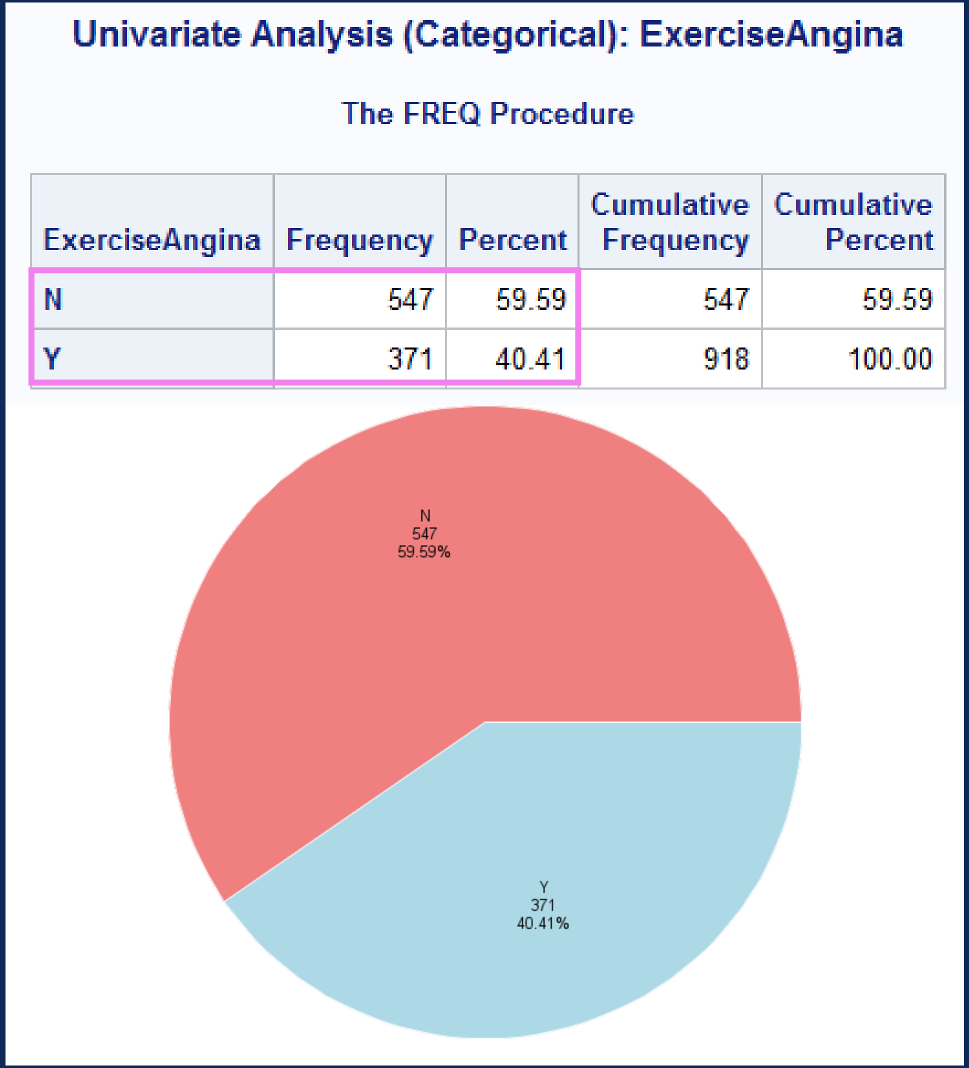
Univariate Analysis (Categorical): RestingECG

The FREQ Procedure

RestingECG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
LVH	188	20.48	188	20.48
Normal	552	60.13	740	80.61
ST	178	19.39	918	100.00

Resting Electrocardiogram Results

EXERCISEANGINA

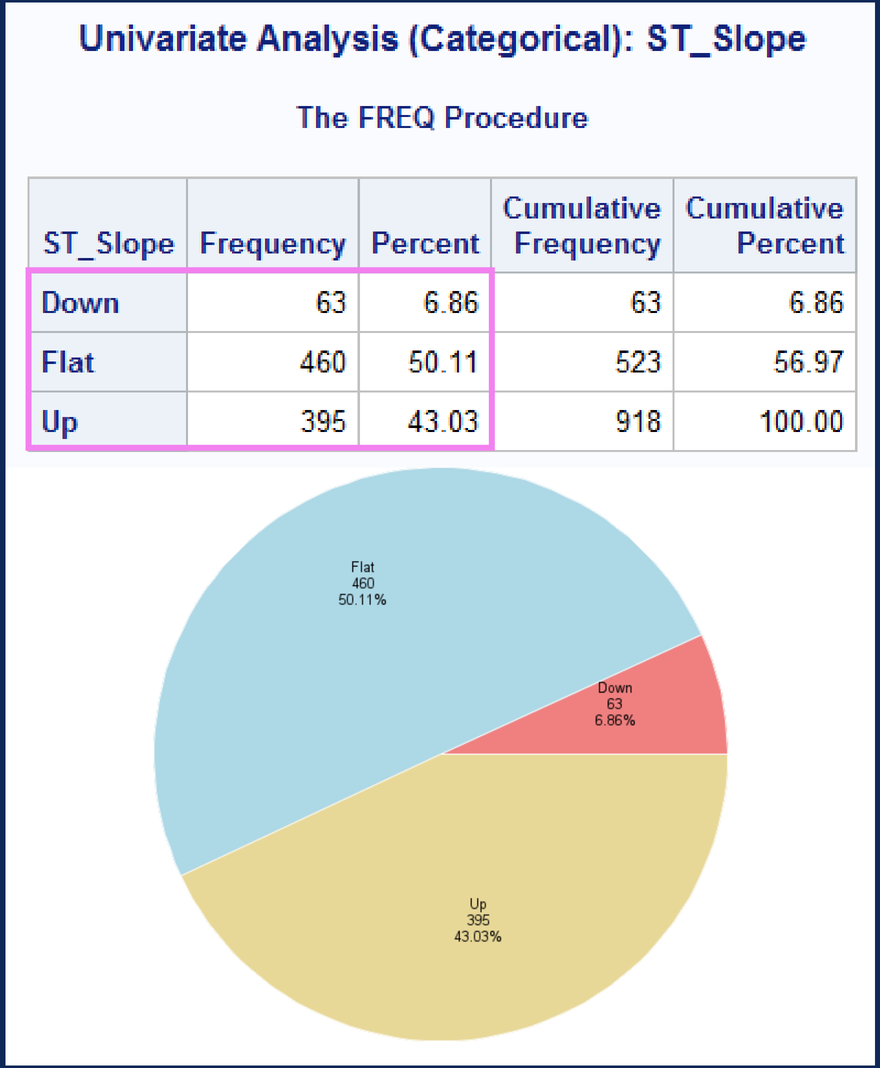


Univariate Analysis (Categorical): ExerciseAngina

The FREQ Procedure

ExerciseAngina	Frequency	Percent	Cumulative Frequency	Cumulative Percent
N	547	59.59	547	59.59
Y	371	40.41	918	100.00

ST_SLOPE

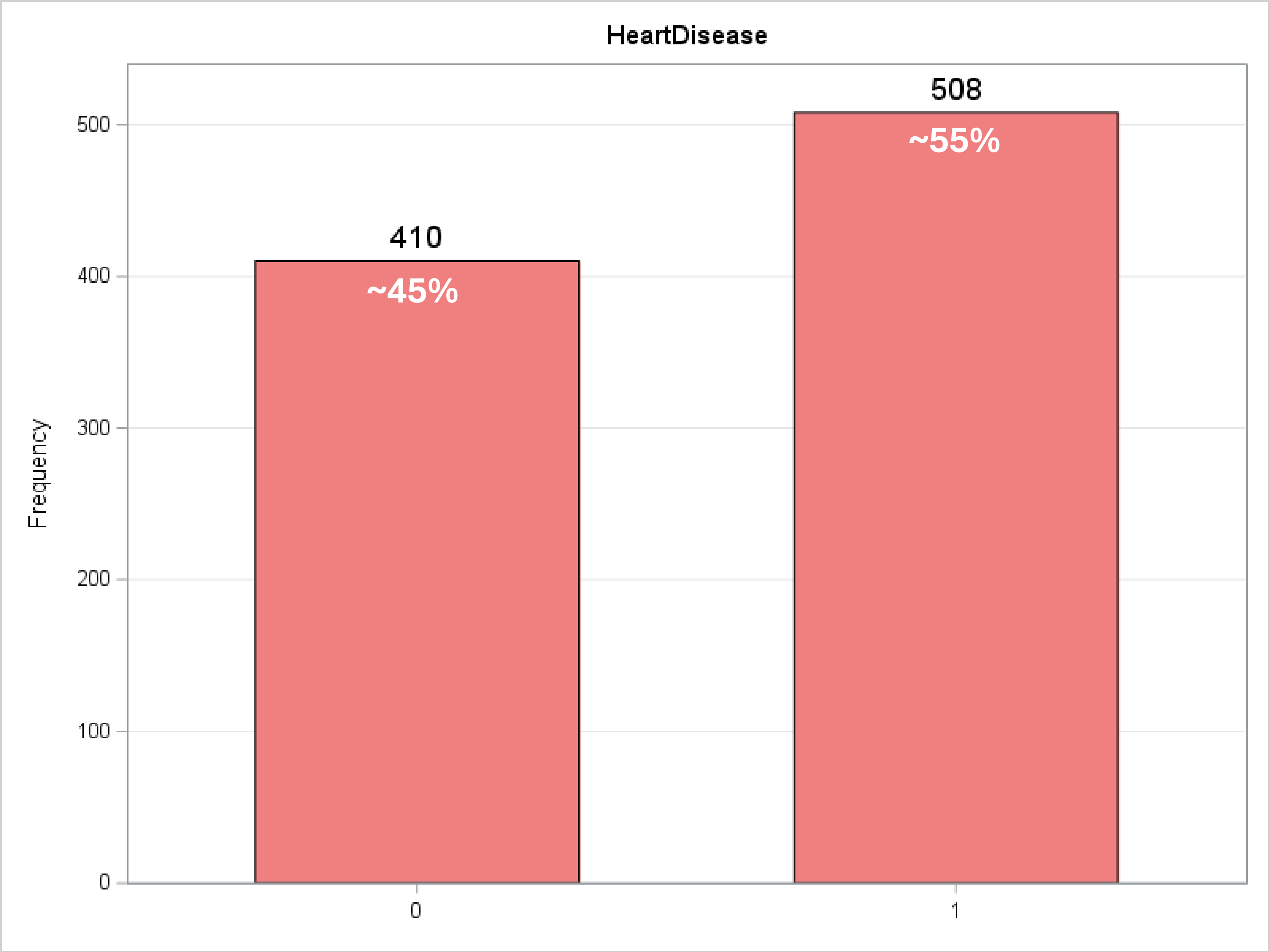


Univariate Analysis (Categorical): ST_Slope

The FREQ Procedure

ST_Slope	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Down	63	6.86	63	6.86
Flat	460	50.11	523	56.97
Up	395	43.03	918	100.00

TARGET VARIABLE



BIVARIATE ANALYSIS

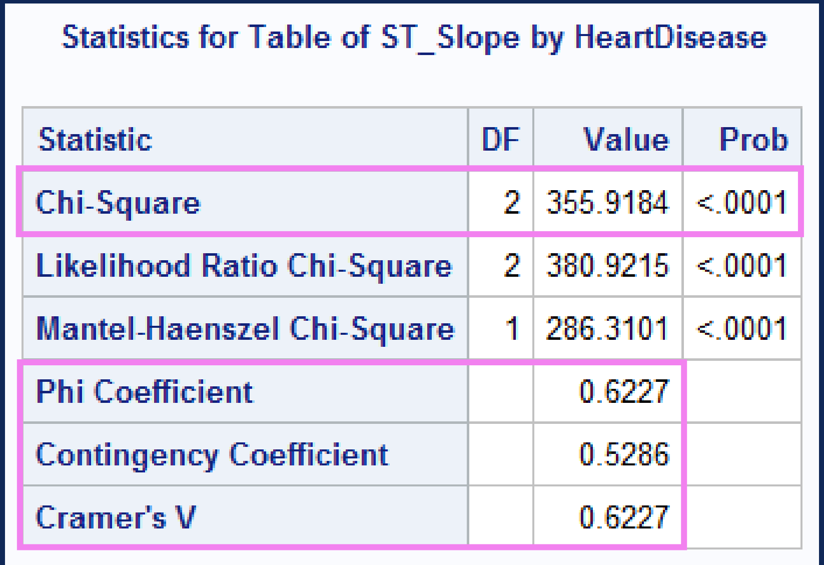
CATEGORICAL VARIABLES

VS

HEART DISEASE

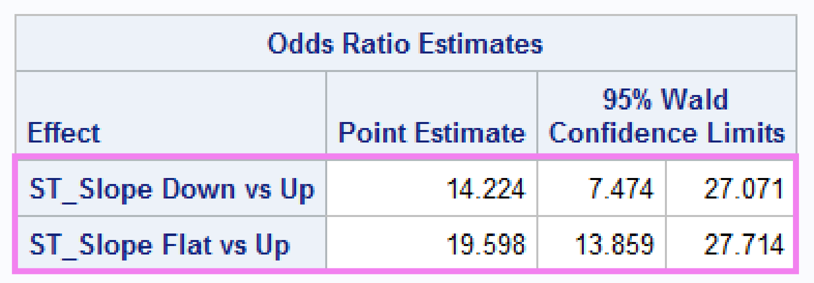
ST_SLOPE / HEART DISEASE

- VERY STRONG ASSOCIATION



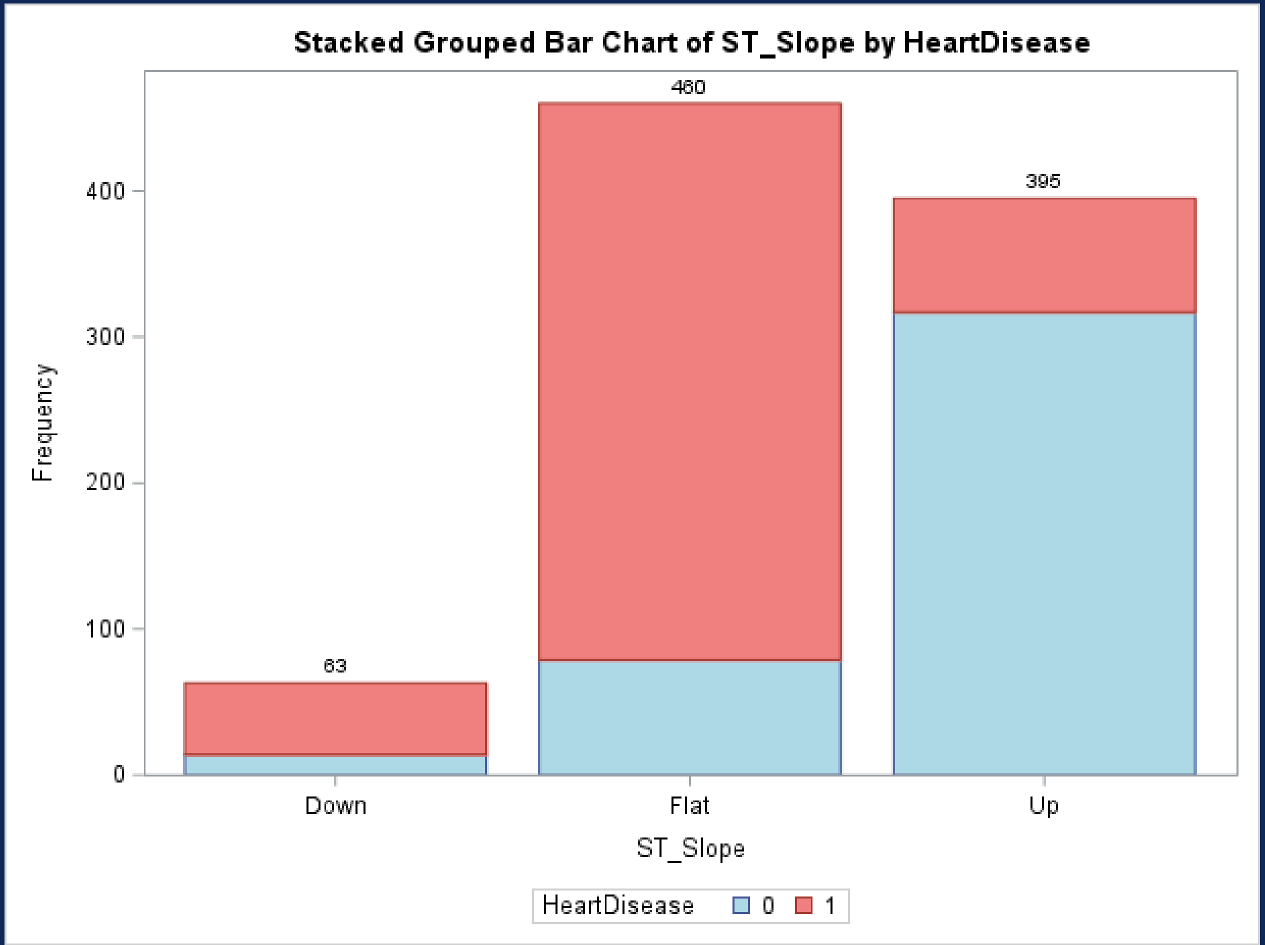
Statistics for Table of ST_Slope by HeartDisease

Statistic	DF	Value	Prob
Chi-Square	2	355.9184	<.0001
Likelihood Ratio Chi-Square	2	380.9215	<.0001
Mantel-Haenszel Chi-Square	1	286.3101	<.0001
Phi Coefficient		0.6227	
Contingency Coefficient		0.5286	
Cramer's V		0.6227	



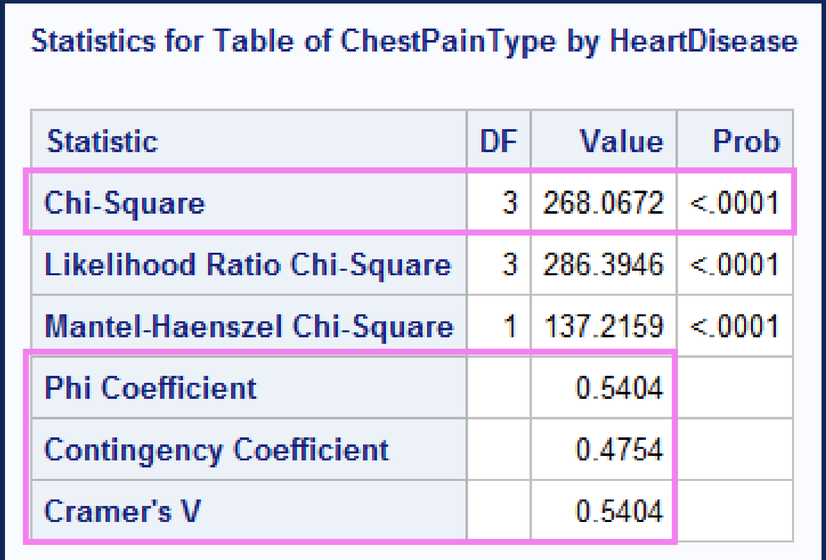
Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
ST_Slope Down vs Up	14.224	7.474	27.071
ST_Slope Flat vs Up	19.598	13.859	27.714



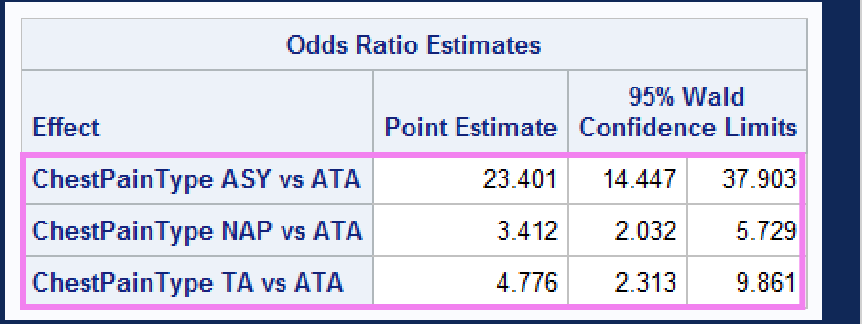
CHESTPAINTYPE / HEART DISEASE

- VERY STRONG ASSOCIATION



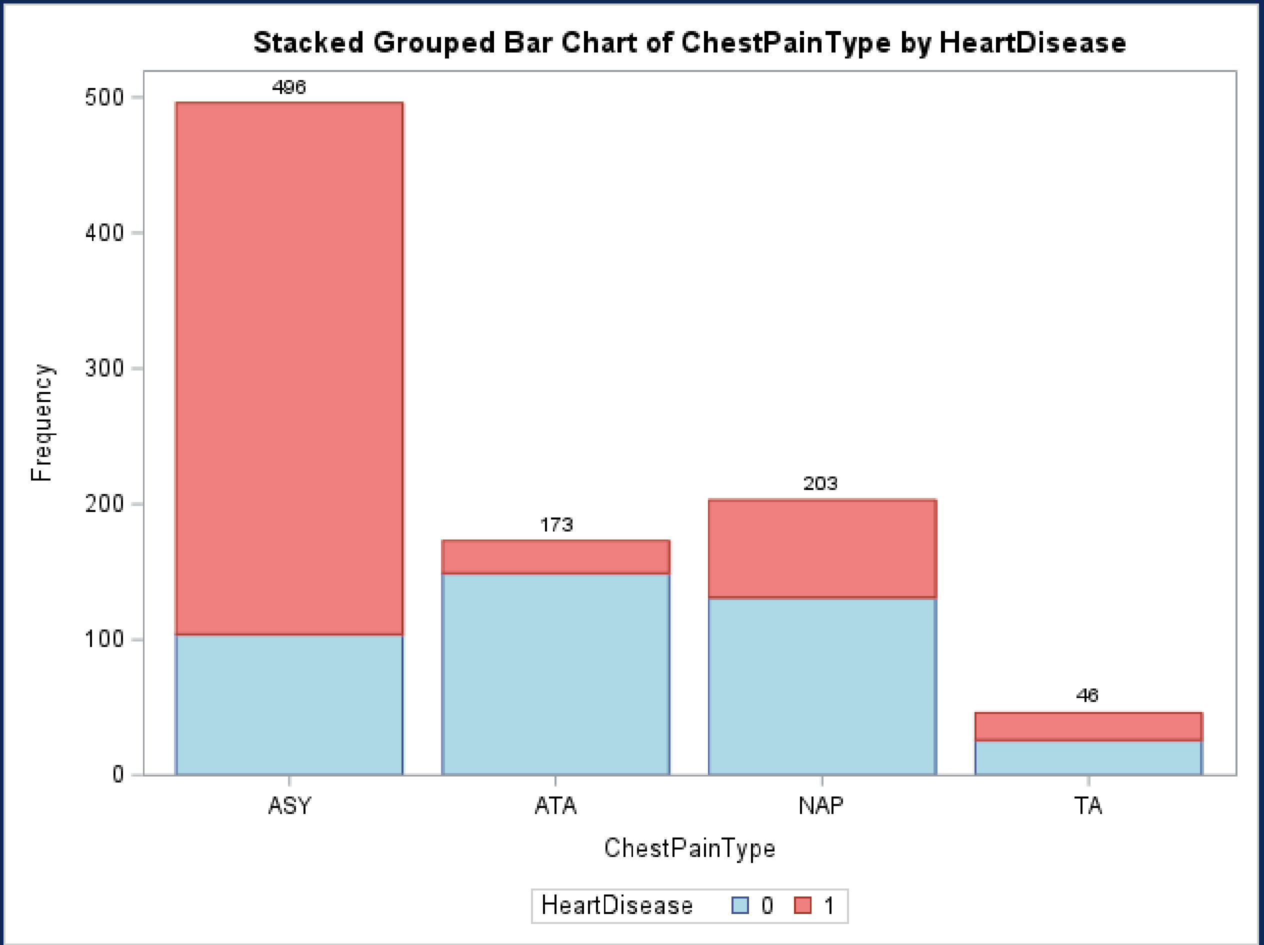
Statistics for Table of ChestPainType by HeartDisease

Statistic	DF	Value	Prob
Chi-Square	3	268.0672	<.0001
Likelihood Ratio Chi-Square	3	286.3946	<.0001
Mantel-Haenszel Chi-Square	1	137.2159	<.0001
Phi Coefficient		0.5404	
Contingency Coefficient		0.4754	
Cramer's V		0.5404	



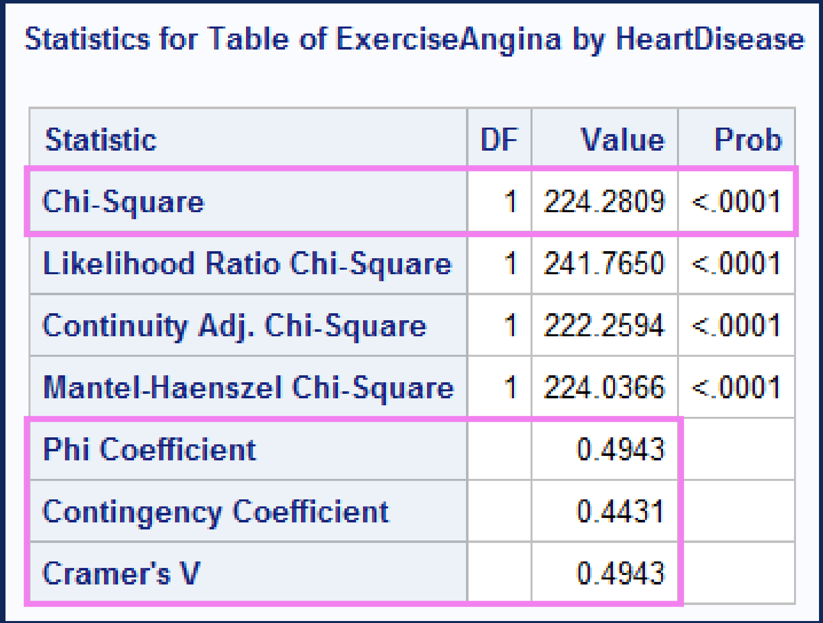
Odds Ratio Estimates

		95% Wald	
Effect	Point Estimate	Confidence Limits	
ChestPainType ASY vs ATA	23.401	14.447	37.903
ChestPainType NAP vs ATA	3.412	2.032	5.729
ChestPainType TA vs ATA	4.776	2.313	9.861



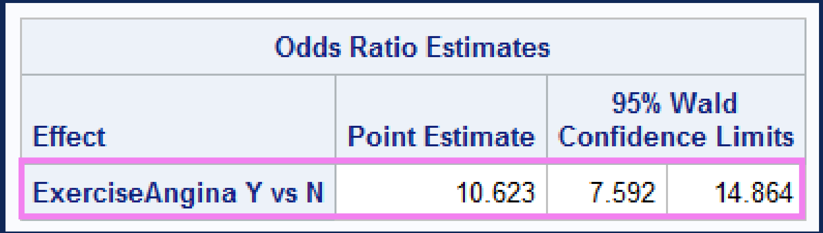
EXERCISEANGINA / HEART DISEASE

- STRONG ASSOCIATION



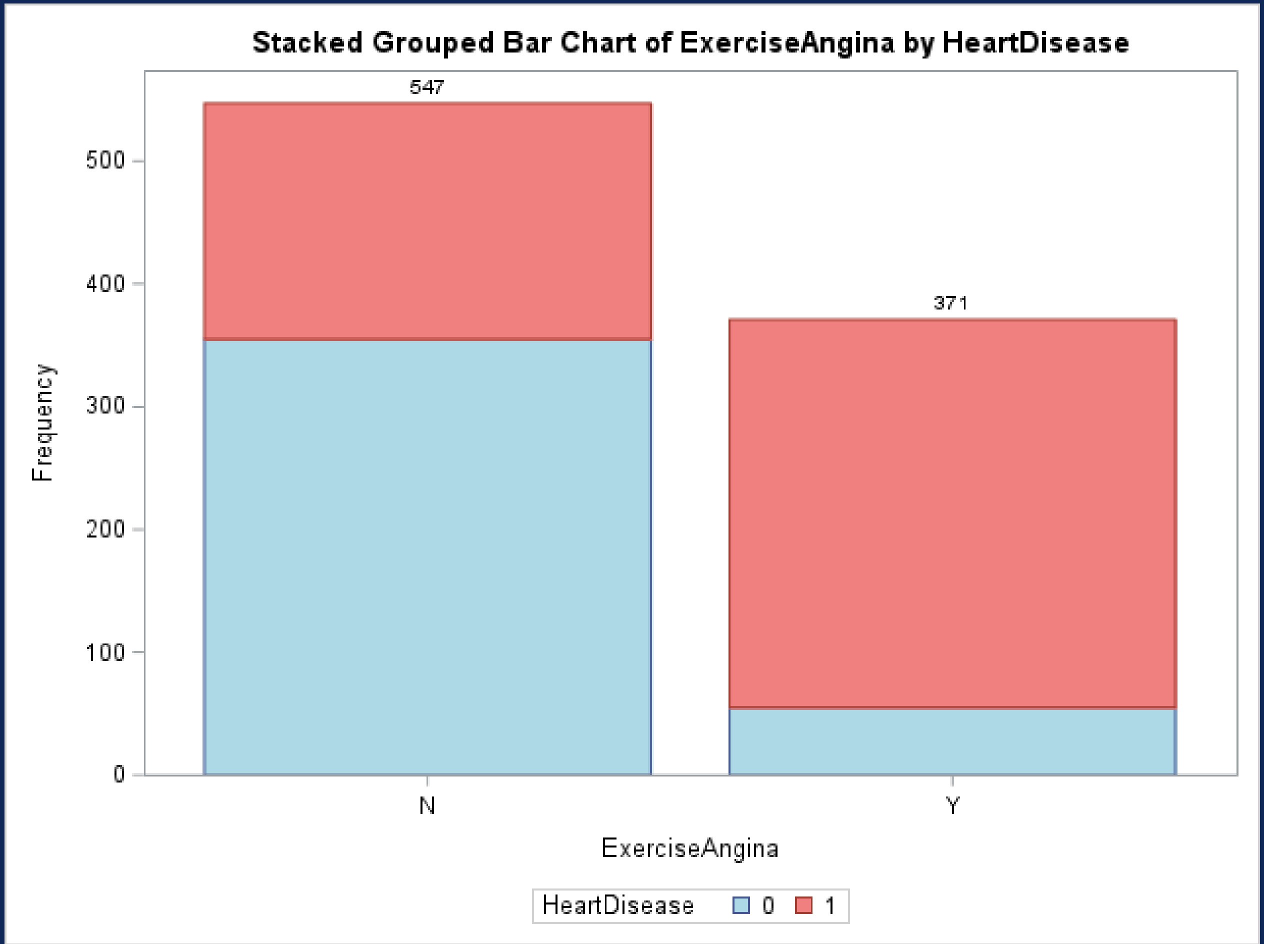
Statistics for Table of ExerciseAngina by HeartDisease

Statistic	DF	Value	Prob
Chi-Square	1	224.2809	<.0001
Likelihood Ratio Chi-Square	1	241.7650	<.0001
Continuity Adj. Chi-Square	1	222.2594	<.0001
Mantel-Haenszel Chi-Square	1	224.0366	<.0001
Phi Coefficient		0.4943	
Contingency Coefficient		0.4431	
Cramer's V		0.4943	



Odds Ratio Estimates

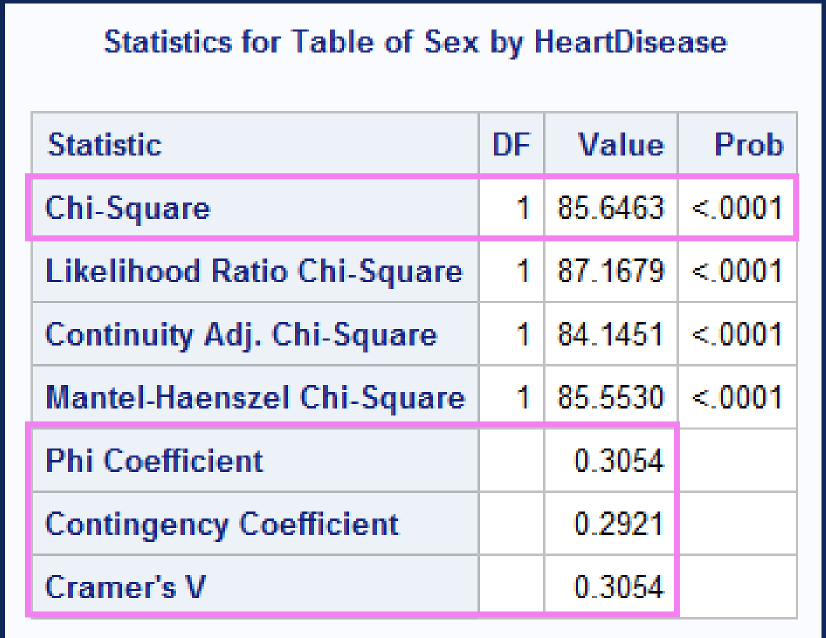
Effect	Point Estimate	95% Wald Confidence Limits	
ExerciseAngina Y vs N	10.623	7.592	14.864



Stacked Grouped Bar Chart of ExerciseAngina by HeartDisease

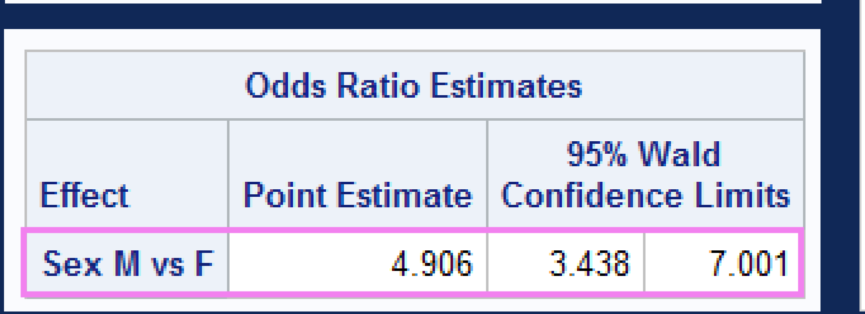
SEX / HEART DISEASE

- STRONG ASSOCIATION



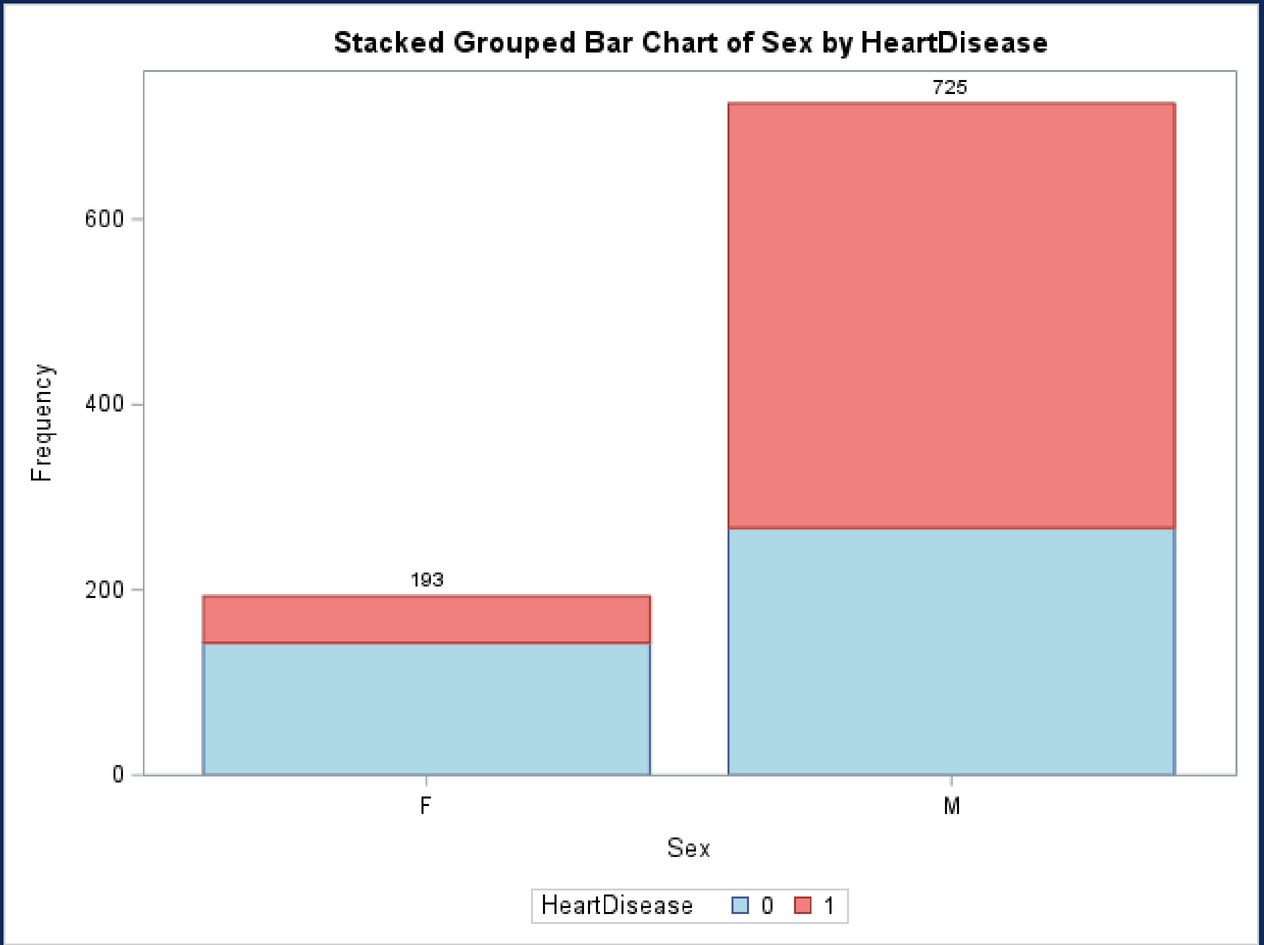
Statistics for Table of Sex by HeartDisease

Statistic	DF	Value	Prob
Chi-Square	1	85.6463	<.0001
Likelihood Ratio Chi-Square	1	87.1679	<.0001
Continuity Adj. Chi-Square	1	84.1451	<.0001
Mantel-Haenszel Chi-Square	1	85.5530	<.0001
Phi Coefficient		0.3054	
Contingency Coefficient		0.2921	
Cramer's V		0.3054	



Odds Ratio Estimates

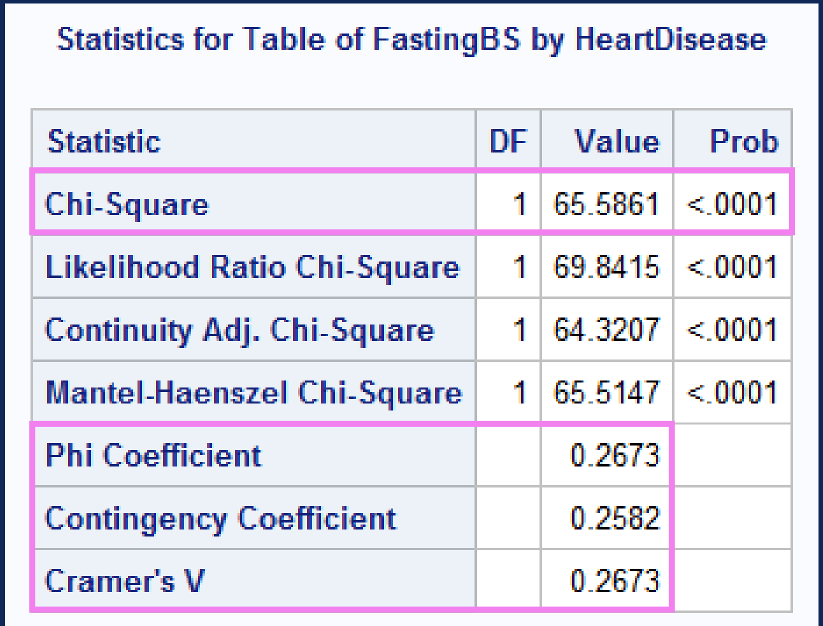
Effect	Point Estimate	95% Wald Confidence Limits	
Sex M vs F	4.906	3.438	7.001



Stacked Grouped Bar Chart of Sex by HeartDisease

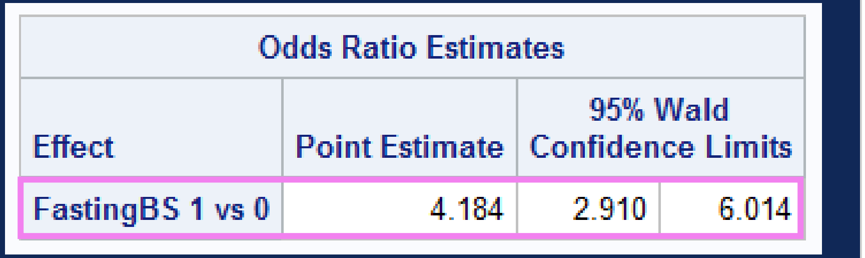
FASTINGBS / HEART DISEASE

- MODERATE ASSOCIATION

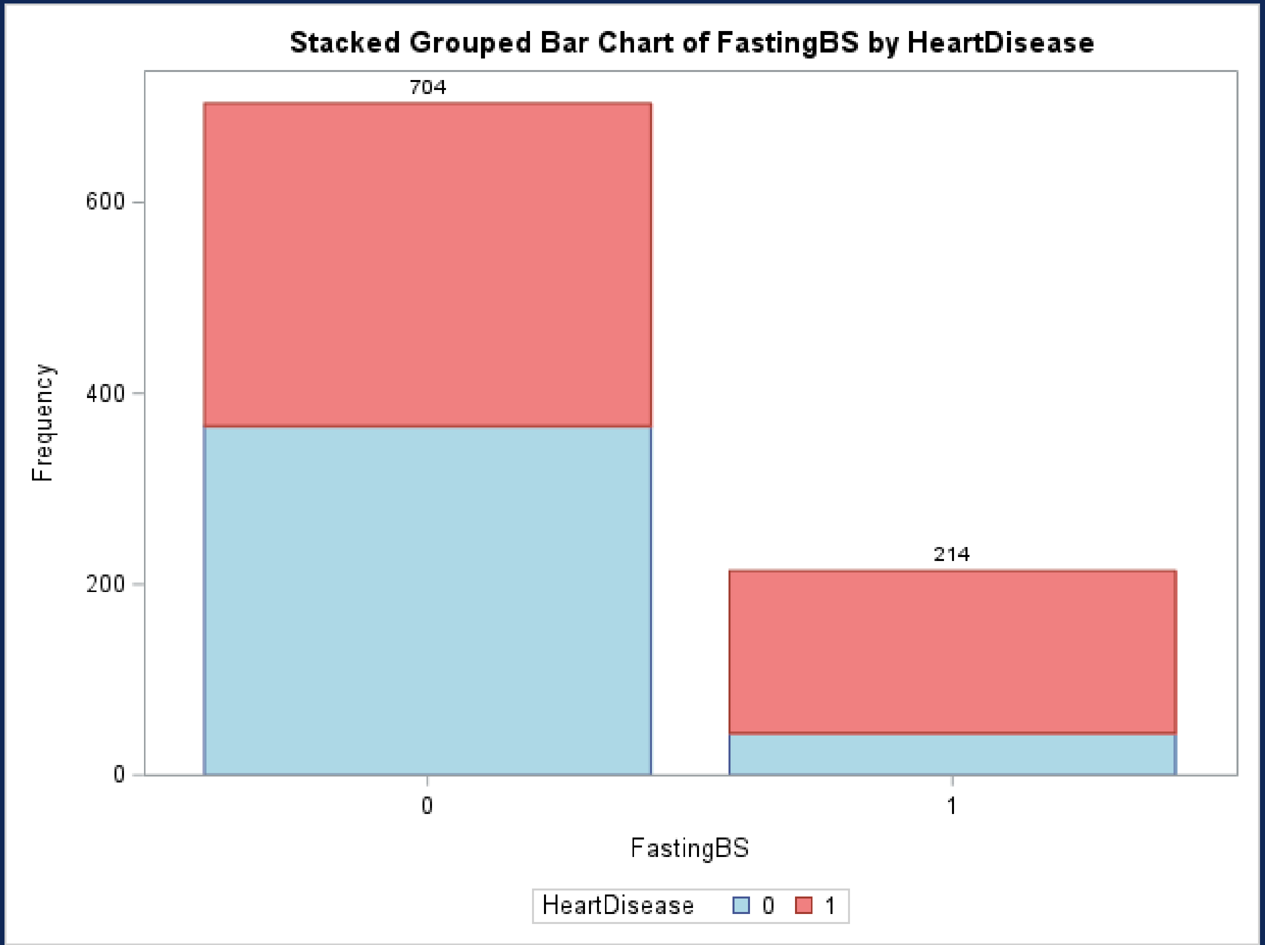


Statistics for Table of FastingBS by HeartDisease

Statistic	DF	Value	Prob
Chi-Square	1	65.5861	<.0001
Likelihood Ratio Chi-Square	1	69.8415	<.0001
Continuity Adj. Chi-Square	1	64.3207	<.0001
Mantel-Haenszel Chi-Square	1	65.5147	<.0001
Phi Coefficient		0.2673	
Contingency Coefficient		0.2582	
Cramer's V		0.2673	



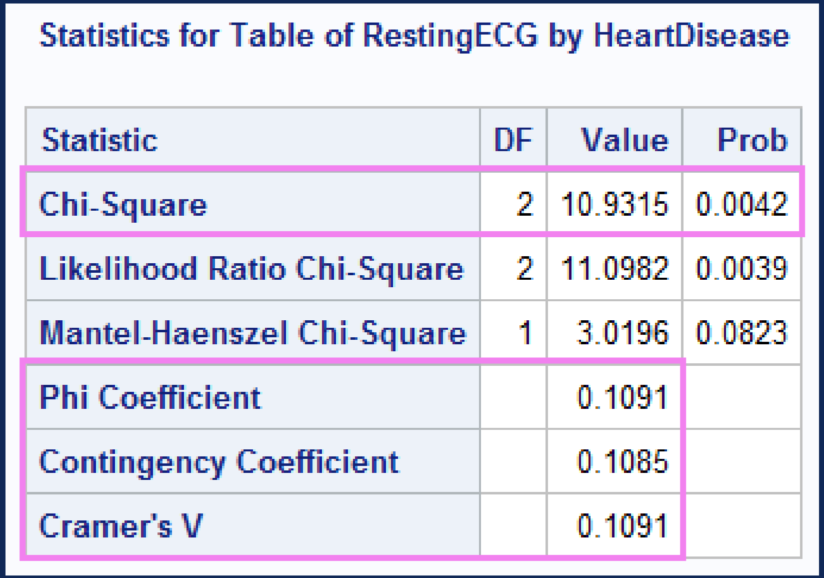
Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
FastingBS 1 vs 0	4.184	2.910	6.014



Stacked Grouped Bar Chart of FastingBS by HeartDisease

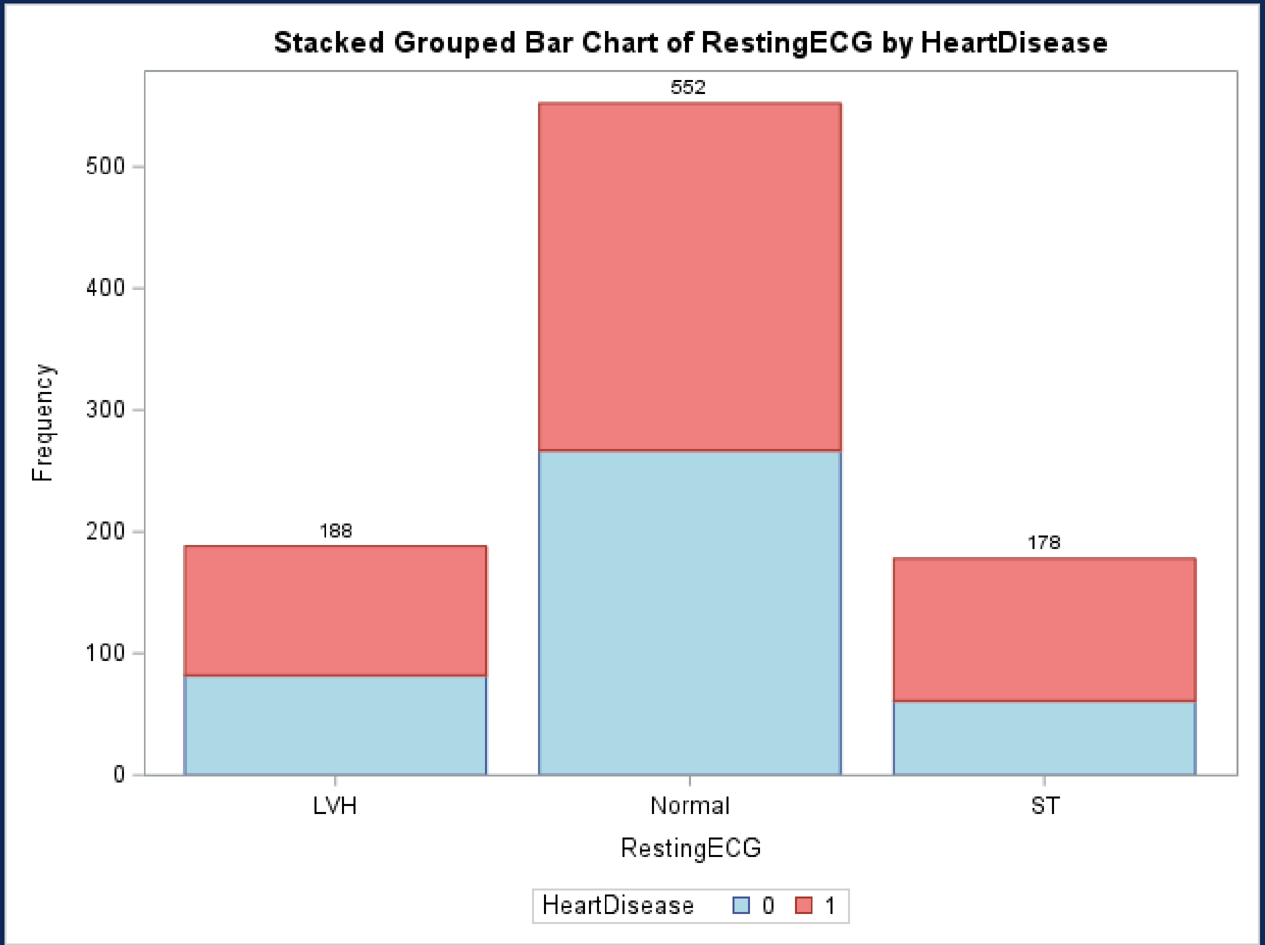
RESTINGECG / HEART DISEASE

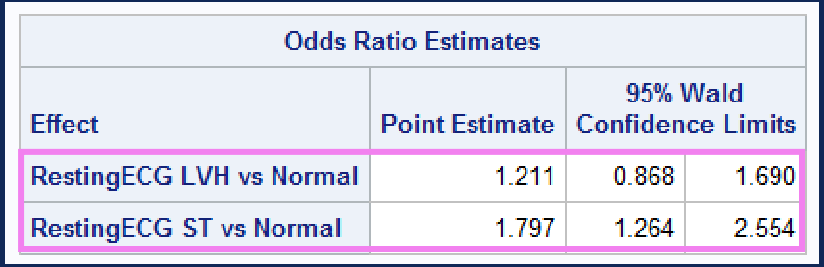
- WEAK ASSOCIATION



Statistics for Table of RestingECG by HeartDisease

Statistic	DF	Value	Prob
Chi-Square	2	10.9315	0.0042
Likelihood Ratio Chi-Square	2	11.0982	0.0039
Mantel-Haenszel Chi-Square	1	3.0196	0.0823
Phi Coefficient		0.1091	
Contingency Coefficient		0.1085	
Cramer's V		0.1091	





Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
RestingECG LVH vs Normal	1.211	0.868	1.690
RestingECG ST vs Normal	1.797	1.264	2.554

BIVARIATE ANALYSIS

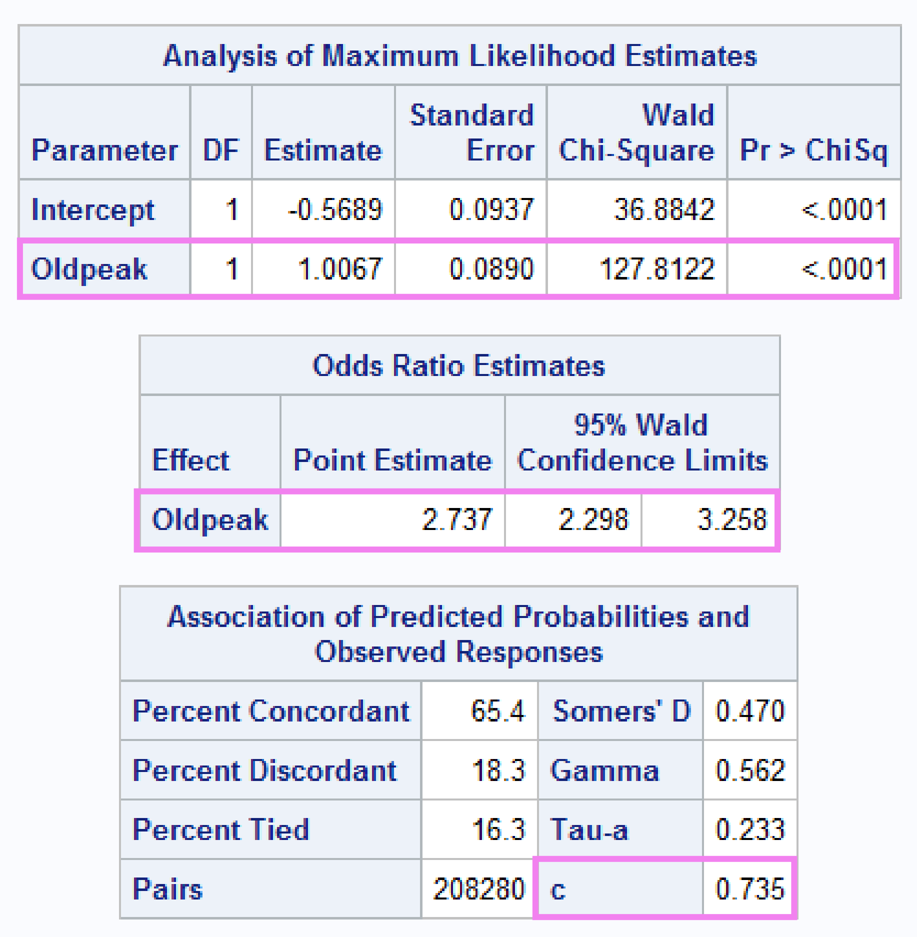
NUMERICAL VARIABLES

VS

HEART DISEASE

OLDPEAK / HEART DISEASE

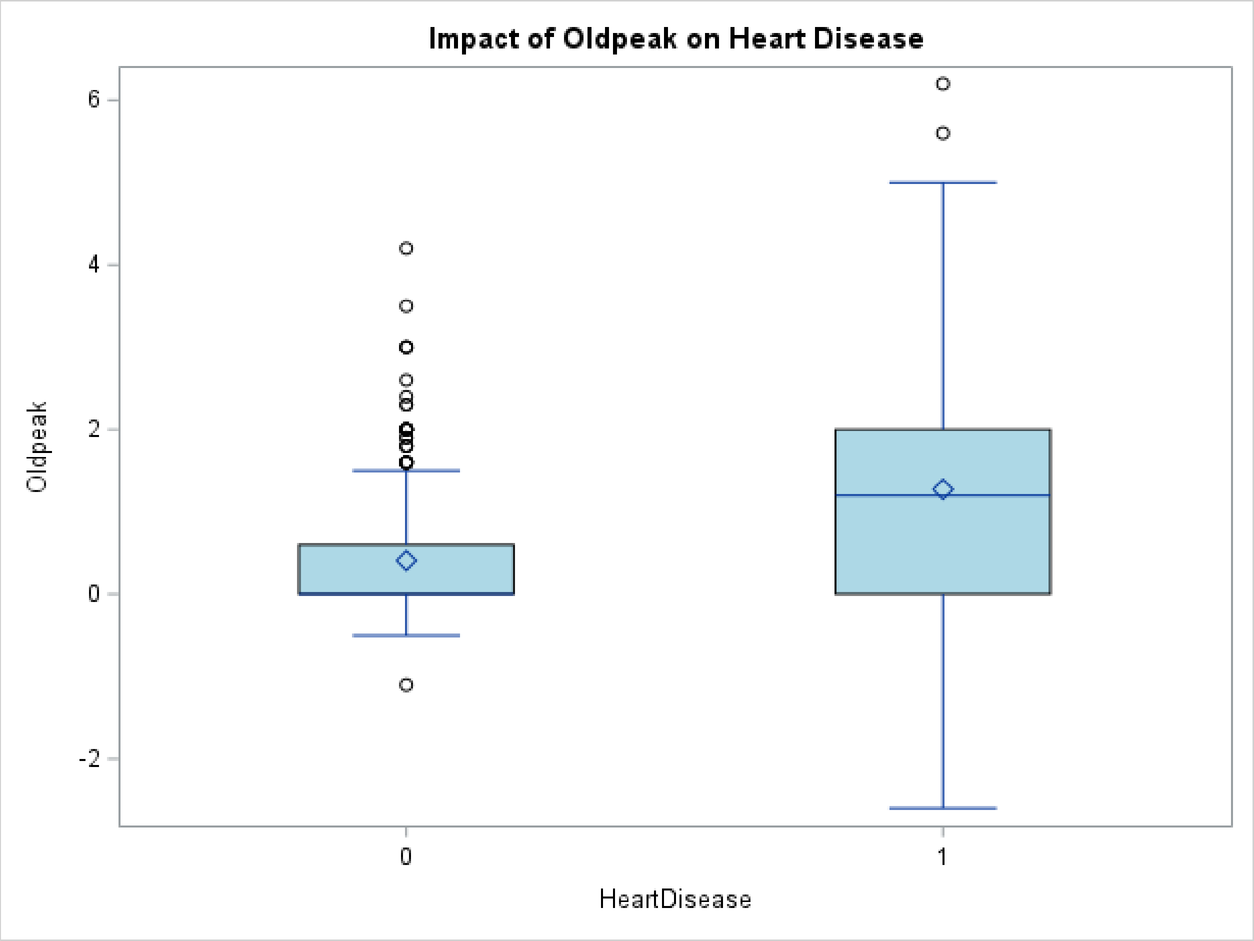
- STRONG POSITIVE RELATIONSHIP



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.5689	0.0937	36.8842	<.0001
Oldpeak	1	1.0067	0.0890	127.8122	<.0001

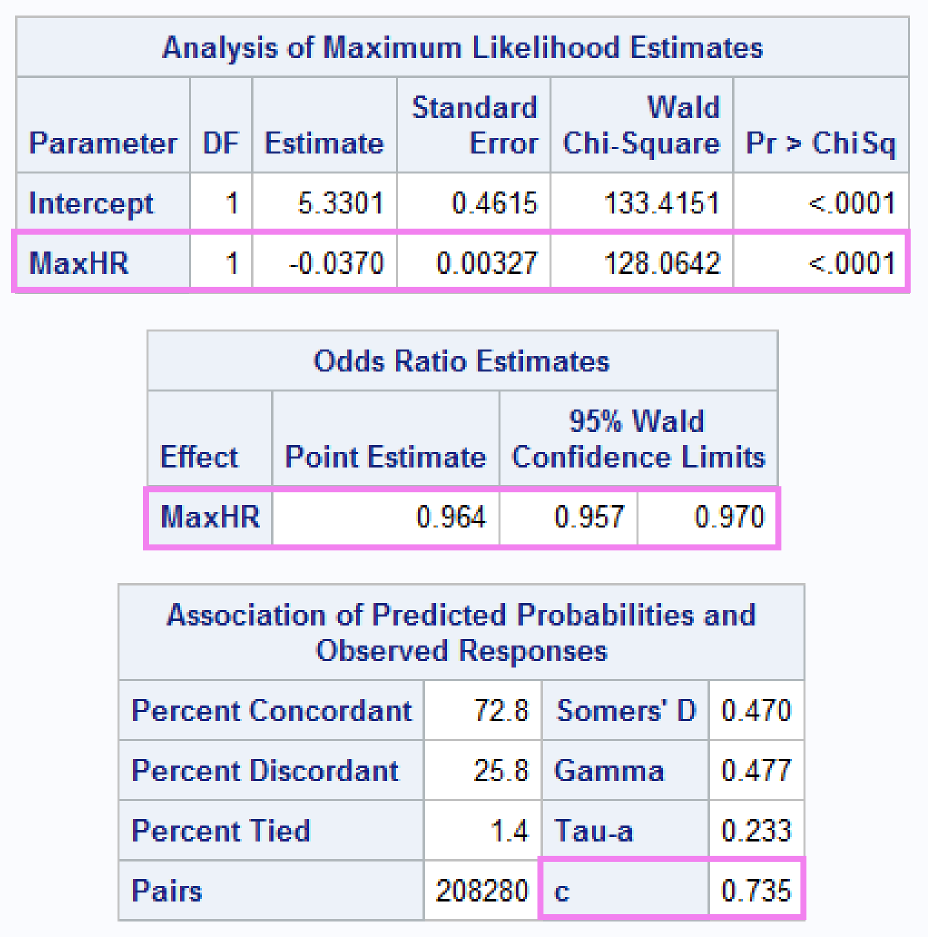
Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Oldpeak	2.737	2.298	3.258

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	65.4	Somers' D	0.470
Percent Discordant	18.3	Gamma	0.562
Percent Tied	16.3	Tau-a	0.233
Pairs	208280	c	0.735



MAXHR / HEART DISEASE

- STRONG NEGATIVE RELATIONSHIP



Analysis of Maximum Likelihood Estimates

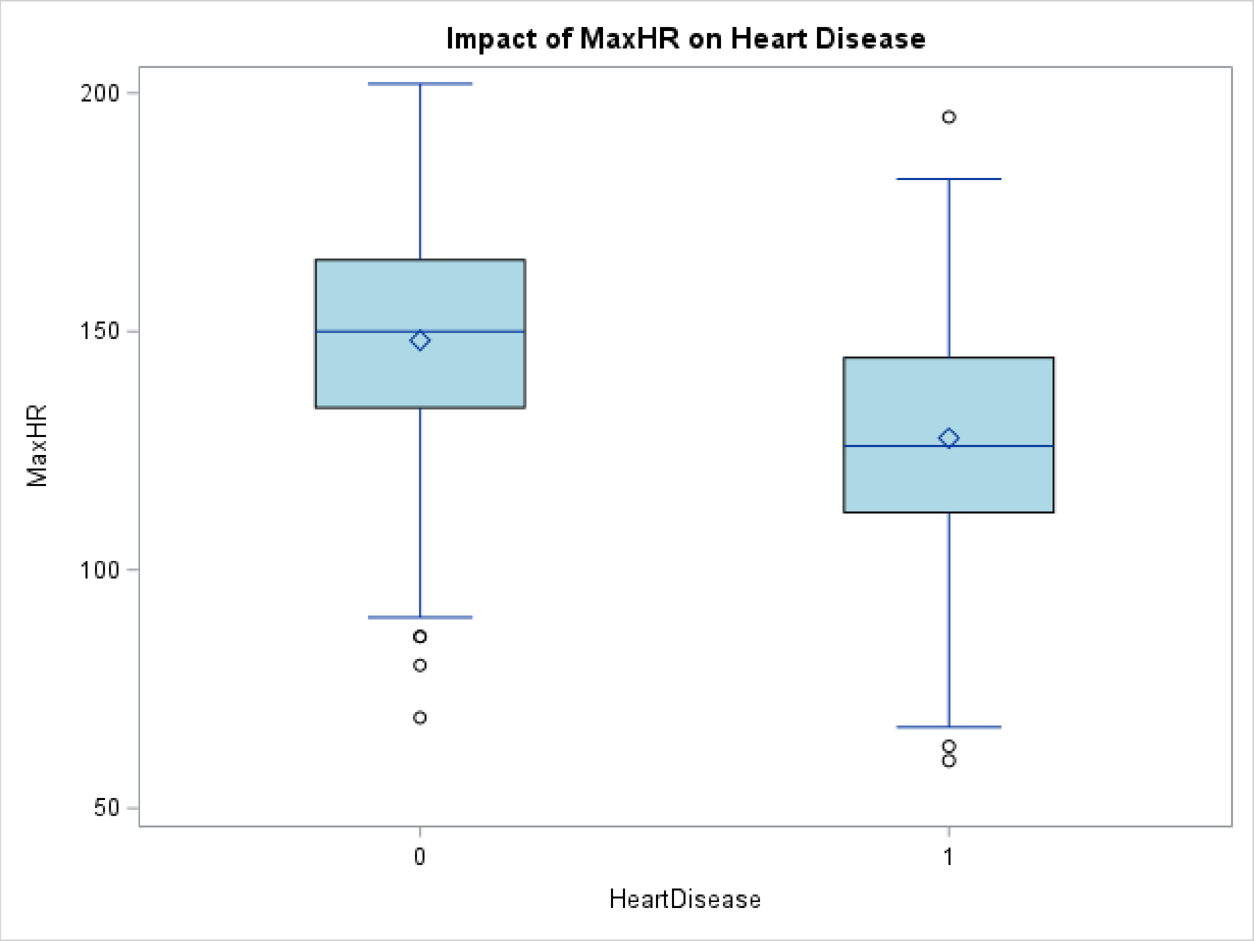
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	5.3301	0.4615	133.4151	<.0001
MaxHR	1	-0.0370	0.00327	128.0642	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
MaxHR	0.964	0.957	0.970

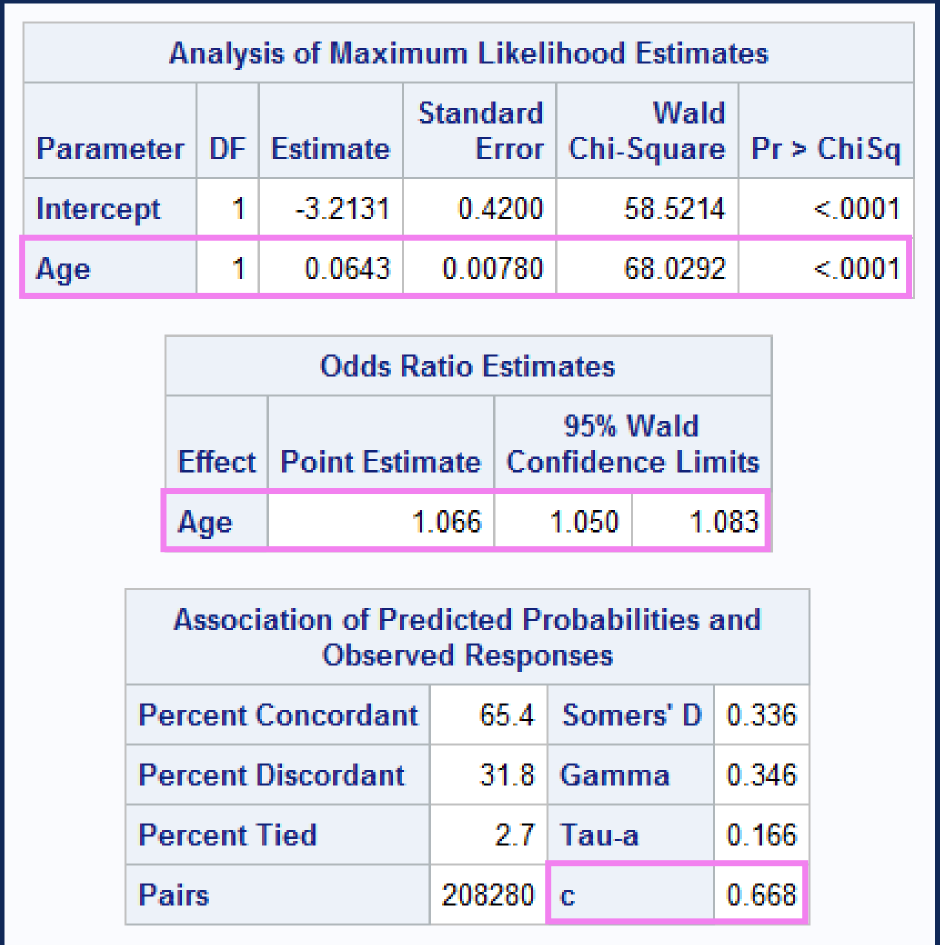
Association of Predicted Probabilities and Observed Responses

Percent Concordant	72.8	Somers' D	0.470
Percent Discordant	25.8	Gamma	0.477
Percent Tied	1.4	Tau-a	0.233
Pairs	208280	c	0.735



AGE / HEART DISEASE

- MODERATE POSITIVE RELATIONSHIP



Analysis of Maximum Likelihood Estimates

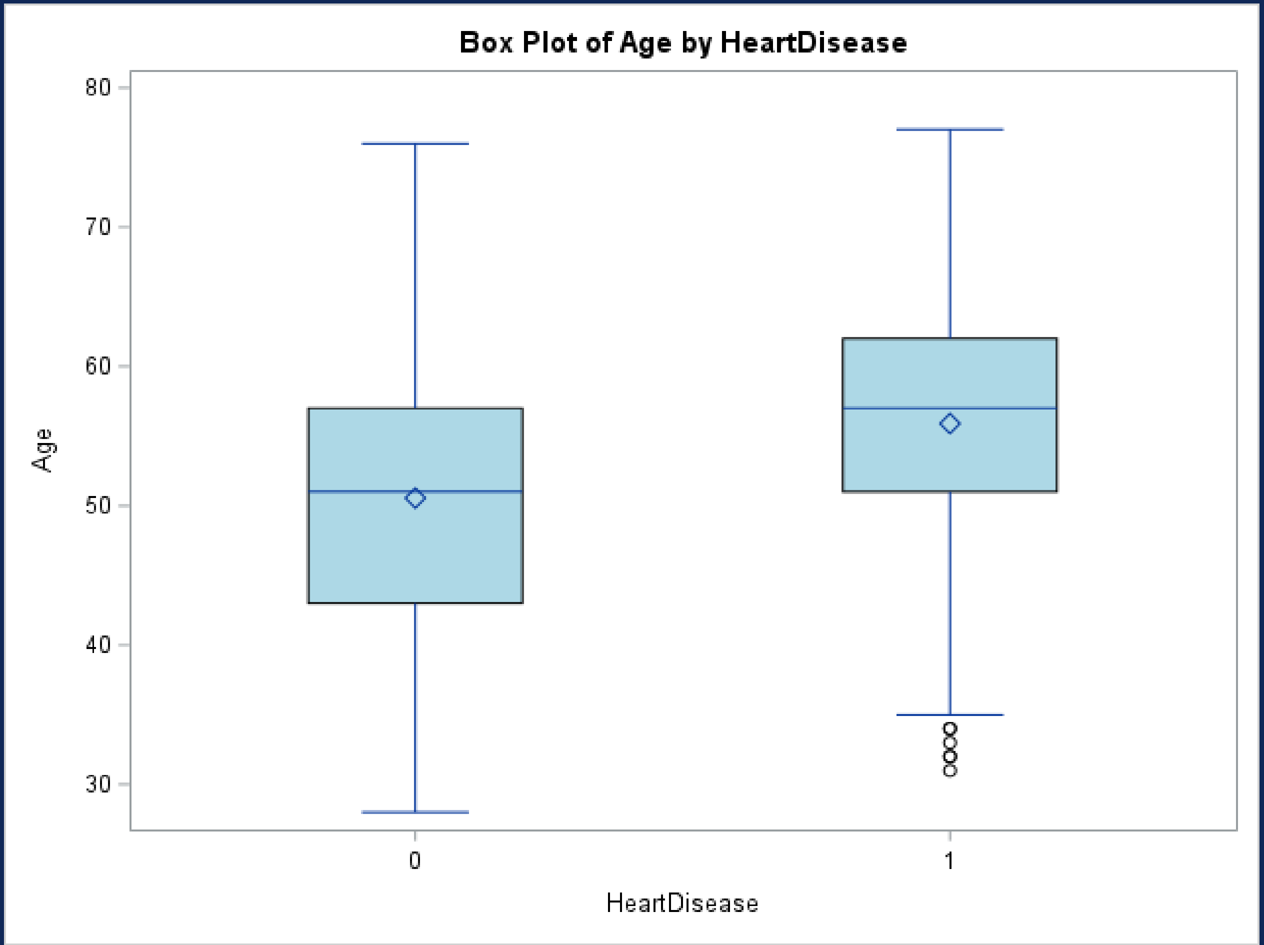
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.2131	0.4200	58.5214	<.0001
Age	1	0.0643	0.00780	68.0292	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Age	1.066	1.050	1.083

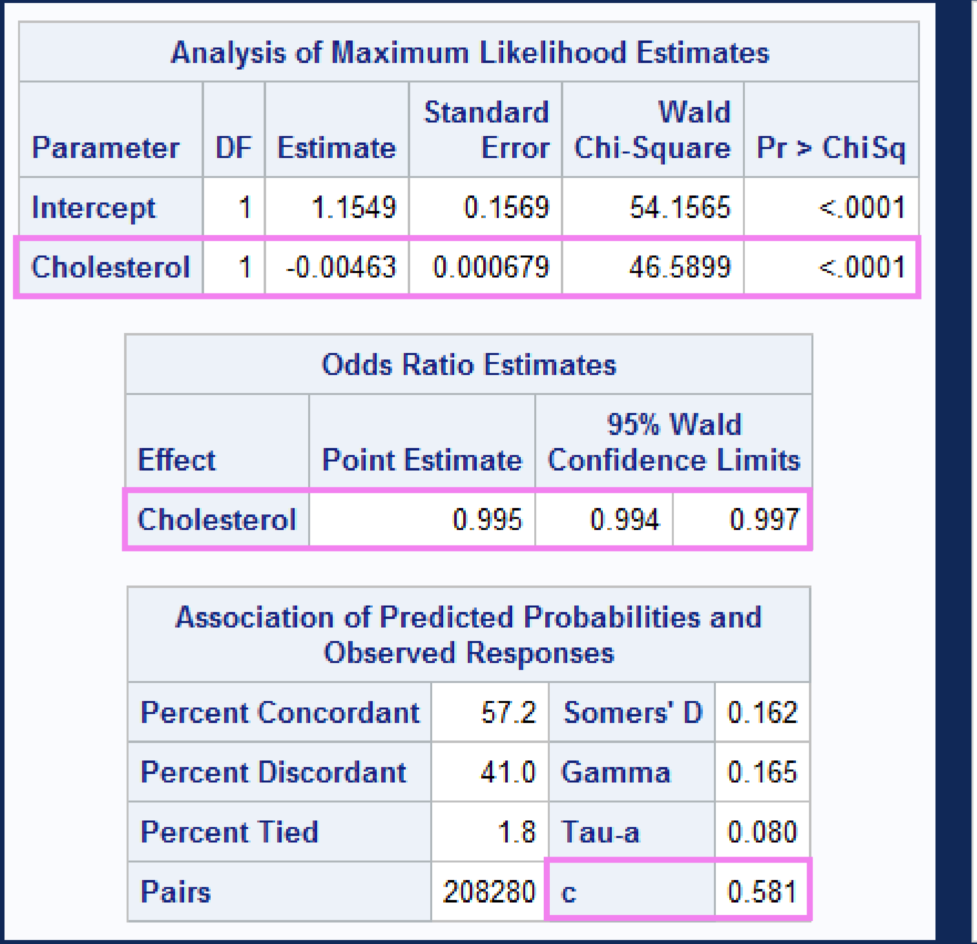
Association of Predicted Probabilities and Observed Responses

Percent Concordant	65.4	Somers' D	0.336
Percent Discordant	31.8	Gamma	0.346
Percent Tied	2.7	Tau-a	0.166
Pairs	208280	c	0.668



CHOLESTEROL / HEART DISEASE

- VERY WEAK NEGATIVE RELATIONSHIP



Analysis of Maximum Likelihood Estimates

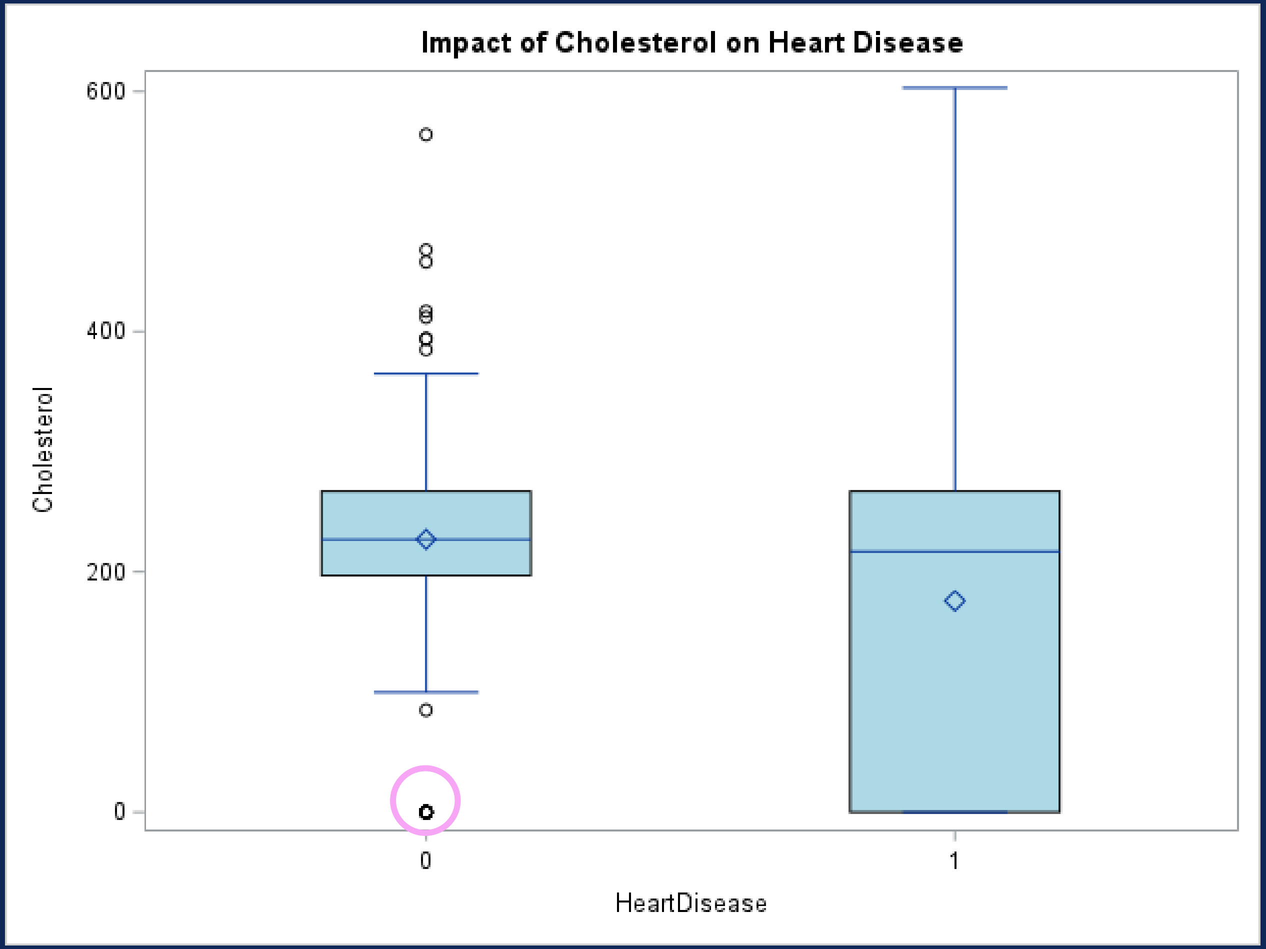
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.1549	0.1569	54.1565	<.0001
Cholesterol	1	-0.00463	0.000679	46.5899	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Cholesterol	0.995	0.994	0.997

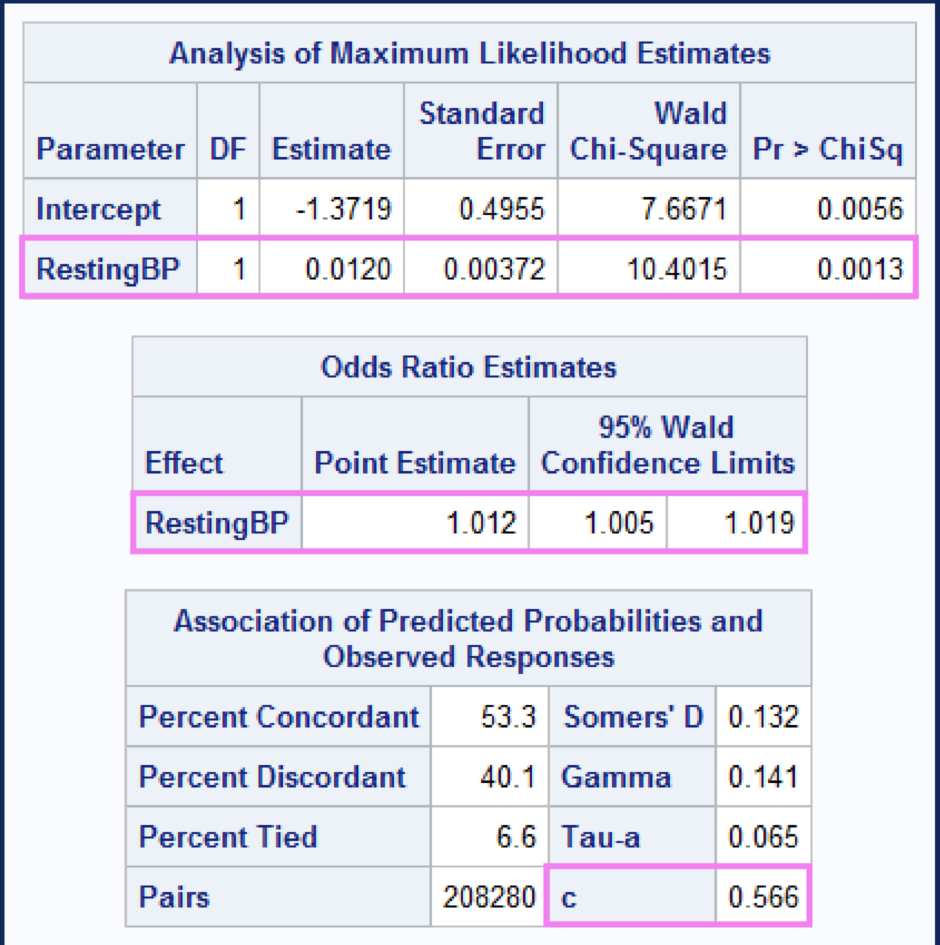
Association of Predicted Probabilities and Observed Responses

Percent Concordant	57.2	Somers' D	0.162
Percent Discordant	41.0	Gamma	0.165
Percent Tied	1.8	Tau-a	0.080
Pairs	208280	c	0.581



RESTINGBP / HEART DISEASE

- VERY WEAK POSITIVE RELATIONSHIP



Analysis of Maximum Likelihood Estimates

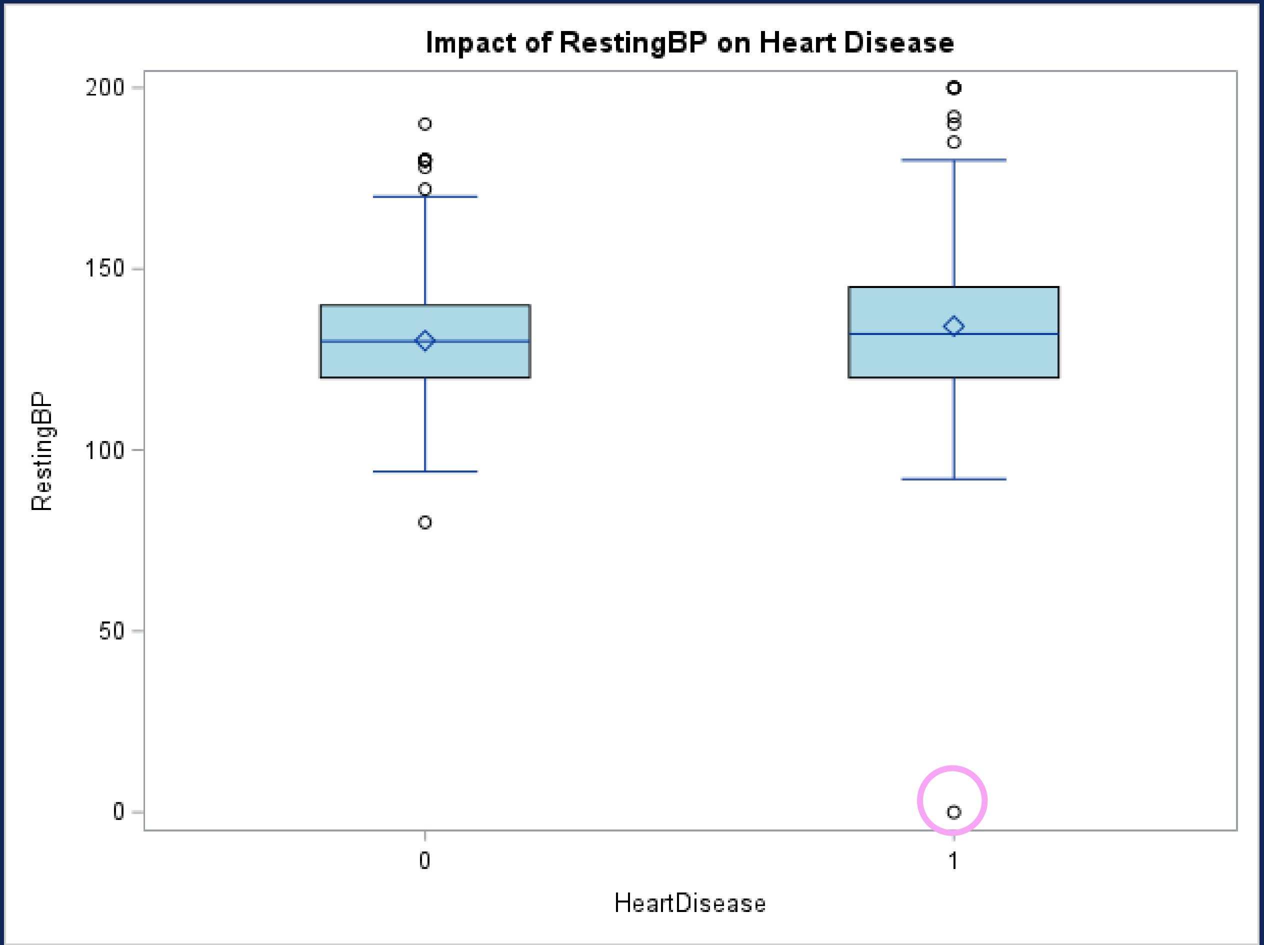
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.3719	0.4955	7.6671	0.0056
RestingBP	1	0.0120	0.00372	10.4015	0.0013

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
RestingBP	1.012	1.005	1.019

Association of Predicted Probabilities and Observed Responses

Percent Concordant	53.3	Somers' D	0.132
Percent Discordant	40.1	Gamma	0.141
Percent Tied	6.6	Tau-a	0.065
Pairs	208280	c	0.566



BIVARIATE ANALYSIS

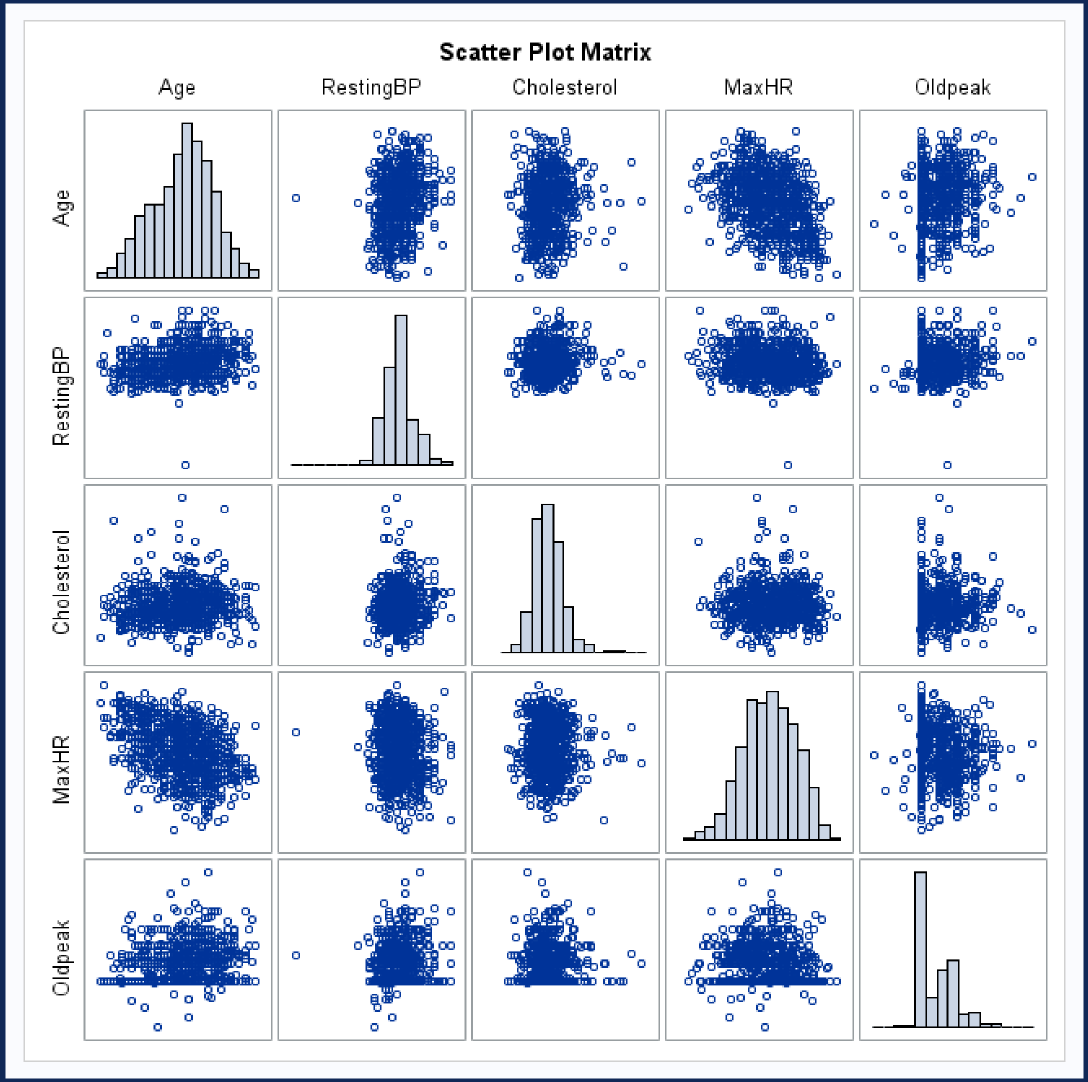
ALL FEATURES

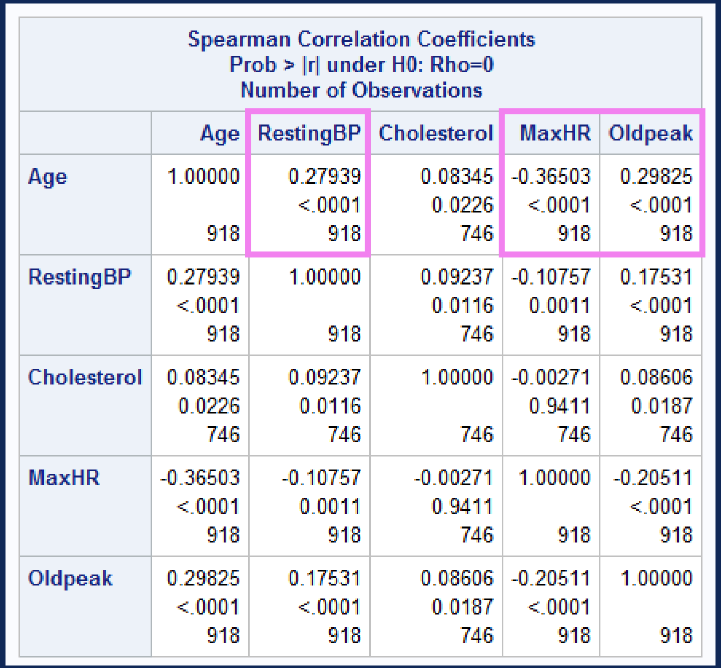
CONTINUOUS / CONTINUOUS

CATEGORICAL / CATEGORICAL

CONTINUOUS / CATEGORICAL

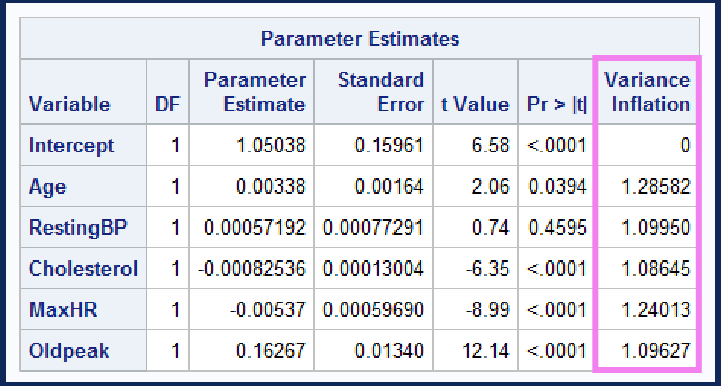
CONTINUOUS / CONTINUOUS





Spearman Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Age	RestingBP	Cholesterol	MaxHR	Oldpeak
Age	1.00000 918	0.27939 <.0001 918	0.08345 0.0226 746	-0.36503 <.0001 918	0.29825 <.0001 918
RestingBP	0.27939 <.0001 918	1.00000 918	0.09237 0.0116 746	-0.10757 0.0011 918	0.17531 <.0001 918
Cholesterol	0.08345 0.0226 746	0.09237 0.0116 746	1.00000 746	-0.00271 0.9411 746	0.08606 0.0187 746
MaxHR	-0.36503 <.0001 918	-0.10757 0.0011 918	-0.00271 0.9411 746	1.00000 918	-0.20511 <.0001 918
Oldpeak	0.29825 <.0001 918	0.17531 <.0001 918	0.08606 0.0187 746	-0.20511 <.0001 918	1.00000 918

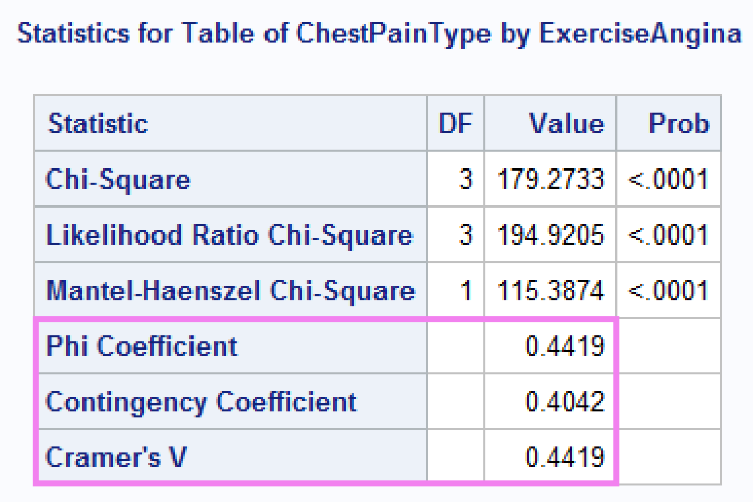


Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1.05038	0.15961	6.58	<.0001	0
Age	1	0.00338	0.00164	2.06	0.0394	1.28582
RestingBP	1	0.00057192	0.00077291	0.74	0.4595	1.09950
Cholesterol	1	-0.00082536	0.00013004	-6.35	<.0001	1.08645
MaxHR	1	-0.00537	0.00059690	-8.99	<.0001	1.24013
Oldpeak	1	0.16267	0.01340	12.14	<.0001	1.09627

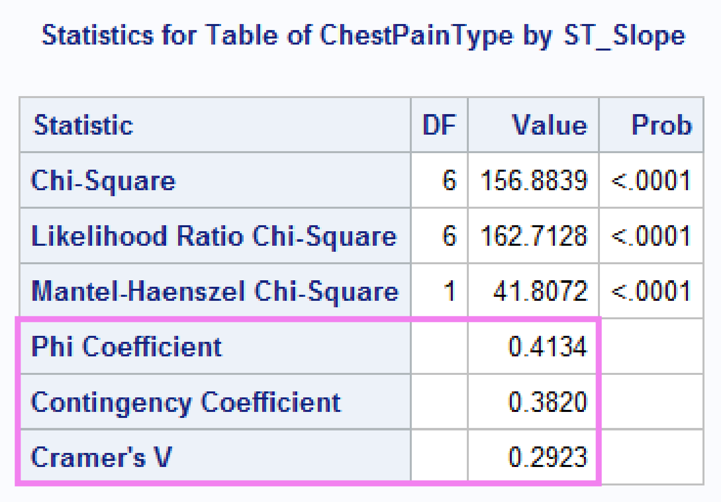
CHESTPAINTYPE / EXERCISEANGINA / ST_SLOPE

STRONG ASSOCIATIONS



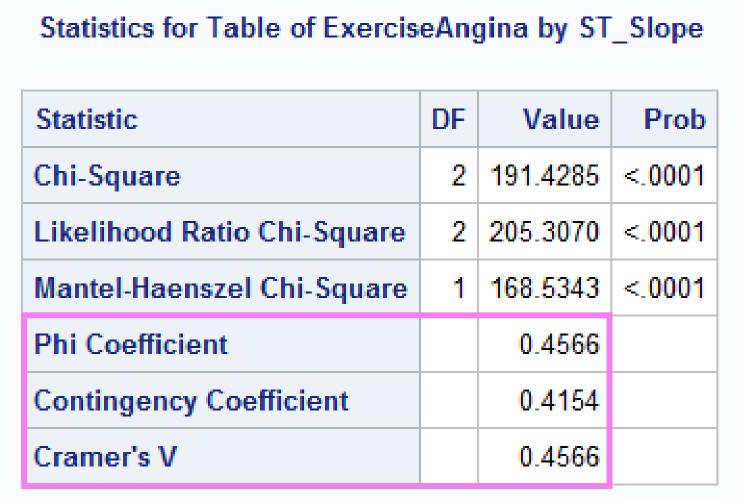
Statistics for Table of ChestPainType by ExerciseAngina

Statistic	DF	Value	Prob
Chi-Square	3	179.2733	<.0001
Likelihood Ratio Chi-Square	3	194.9205	<.0001
Mantel-Haenszel Chi-Square	1	115.3874	<.0001
Phi Coefficient		0.4419	
Contingency Coefficient		0.4042	
Cramer's V		0.4419	



Statistics for Table of ChestPainType by ST_Slope

Statistic	DF	Value	Prob
Chi-Square	6	156.8839	<.0001
Likelihood Ratio Chi-Square	6	162.7128	<.0001
Mantel-Haenszel Chi-Square	1	41.8072	<.0001
Phi Coefficient		0.4134	
Contingency Coefficient		0.3820	
Cramer's V		0.2923	



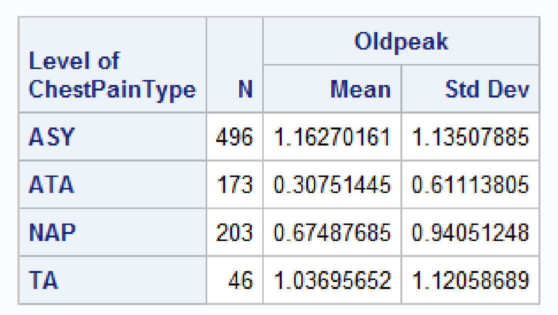
Statistics for Table of ExerciseAngina by ST_Slope

Statistic	DF	Value	Prob
Chi-Square	2	191.4285	<.0001
Likelihood Ratio Chi-Square	2	205.3070	<.0001
Mantel-Haenszel Chi-Square	1	168.5343	<.0001
Phi Coefficient		0.4566	
Contingency Coefficient		0.4154	
Cramer's V		0.4566	

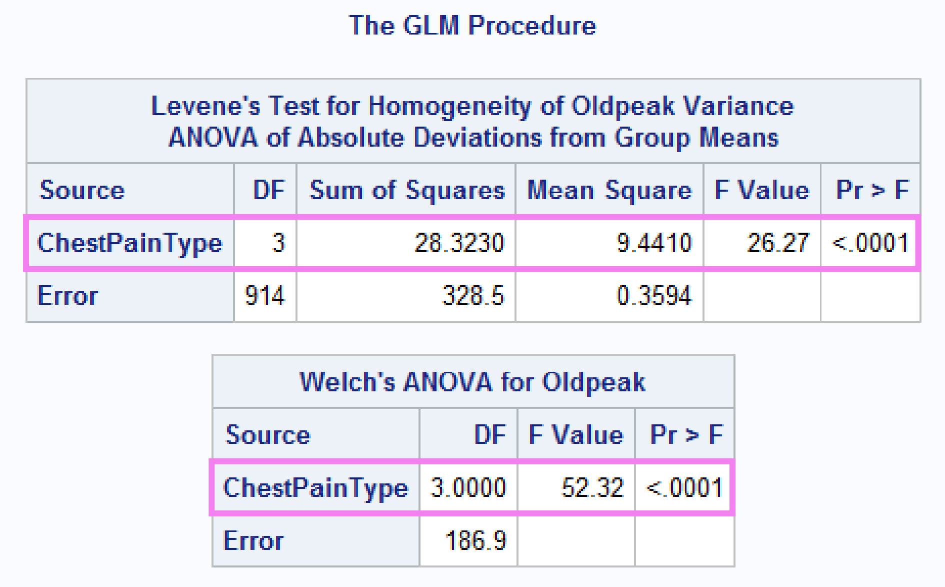
CATEGORICAL VS NUMERICAL

T-TEST
ANOVA TEST

CHESTPAINTYPE / OLDPEAK



Level of ChestPainType	N	Oldpeak	
		Mean	Std Dev
ASY	496	1.16270161	1.13507885
ATA	173	0.30751445	0.61113805
NAP	203	0.67487685	0.94051248
TA	46	1.03695652	1.12058689



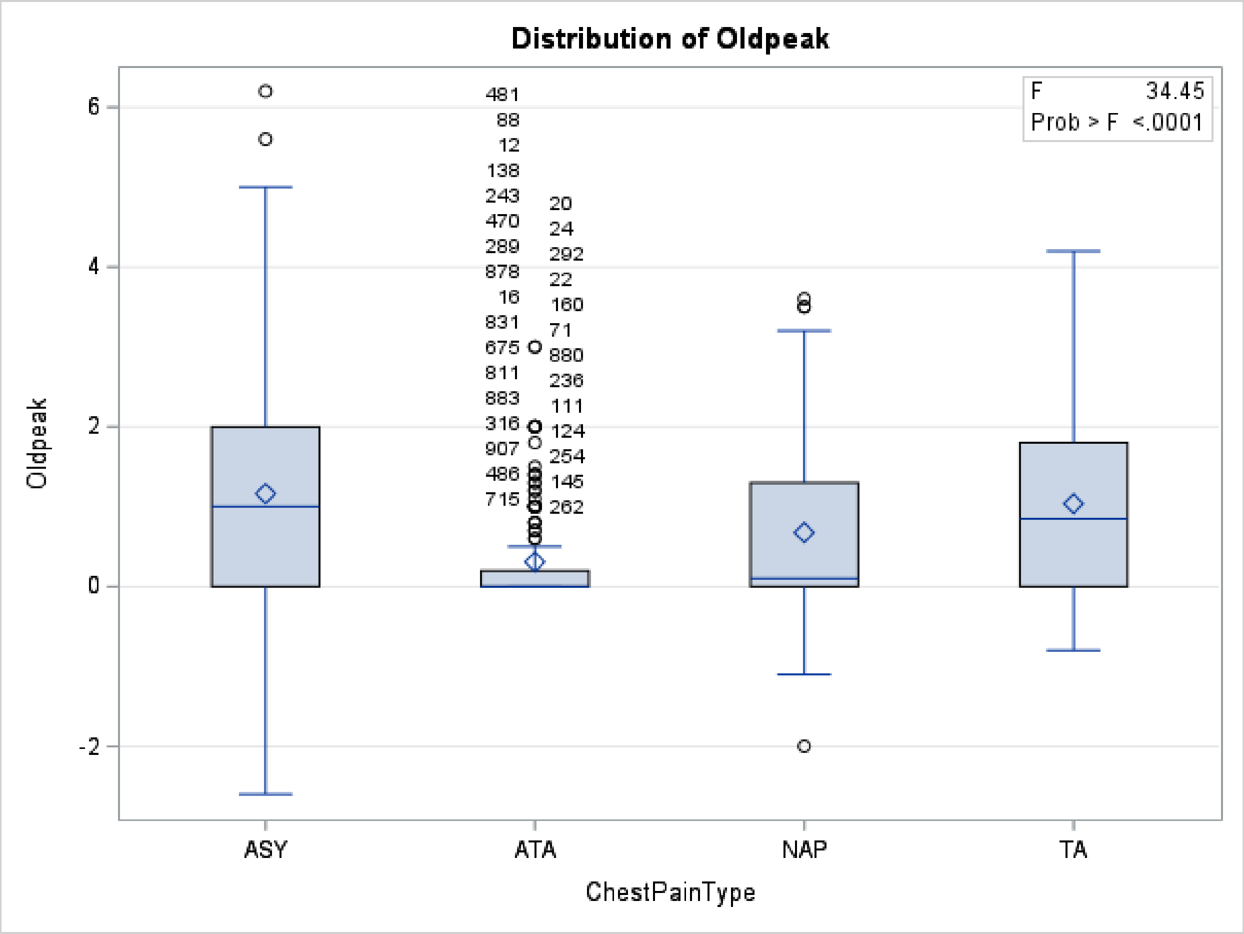
The GLM Procedure

Levene's Test for Homogeneity of Oldpeak Variance
ANOVA of Absolute Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
ChestPainType	3	28.3230	9.4410	26.27	<.0001
Error	914	328.5	0.3594		

Welch's ANOVA for Oldpeak

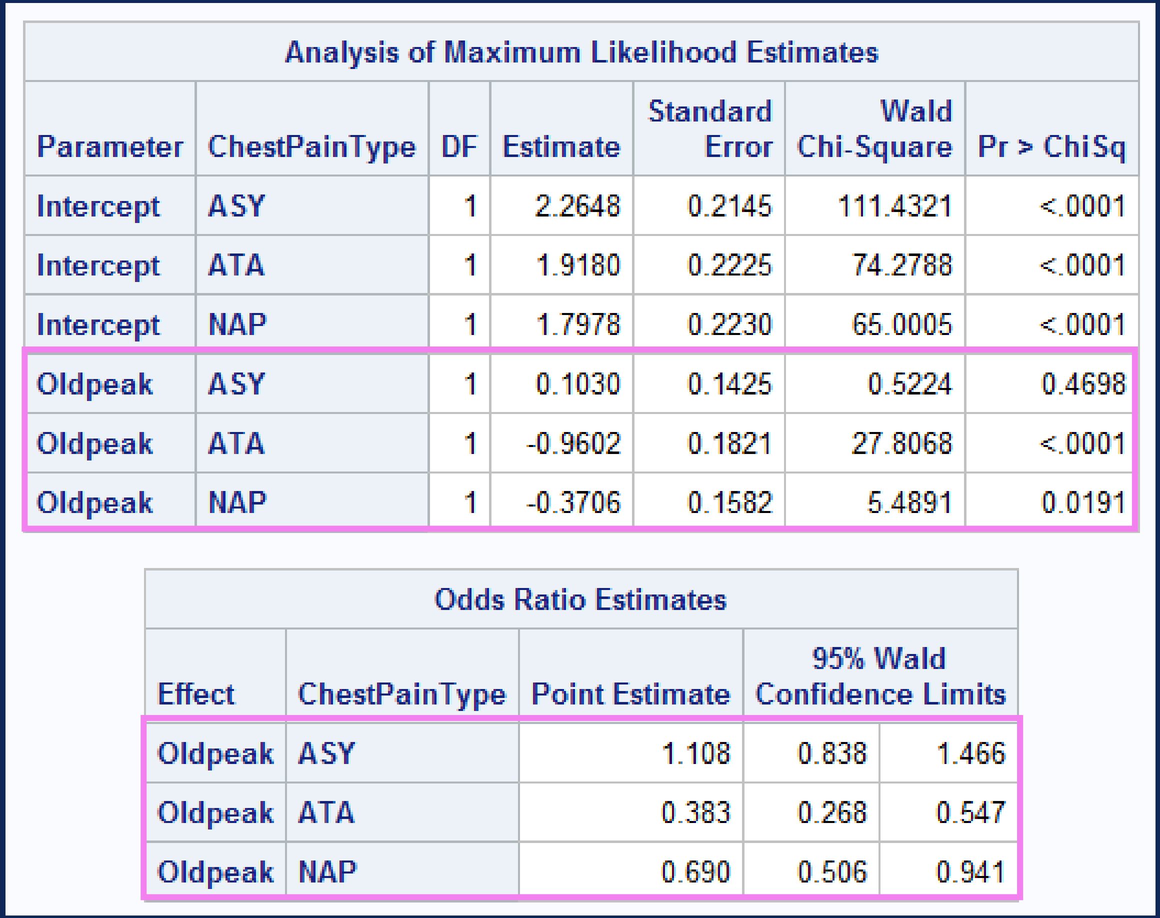
Source	DF	F Value	Pr > F
ChestPainType	3.0000	52.32	<.0001
Error	186.9		



Distribution of Oldpeak

CHESTPAINTYPE / OLDPEAK

- STRONG SIGNIFICANT RELATIONSHIP



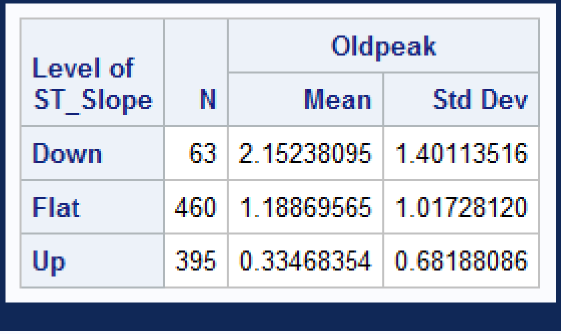
Analysis of Maximum Likelihood Estimates

Parameter	ChestPainType	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	ASY	1	2.2648	0.2145	111.4321	<.0001
Intercept	ATA	1	1.9180	0.2225	74.2788	<.0001
Intercept	NAP	1	1.7978	0.2230	65.0005	<.0001
Oldpeak	ASY	1	0.1030	0.1425	0.5224	0.4698
Oldpeak	ATA	1	-0.9602	0.1821	27.8068	<.0001
Oldpeak	NAP	1	-0.3706	0.1582	5.4891	0.0191

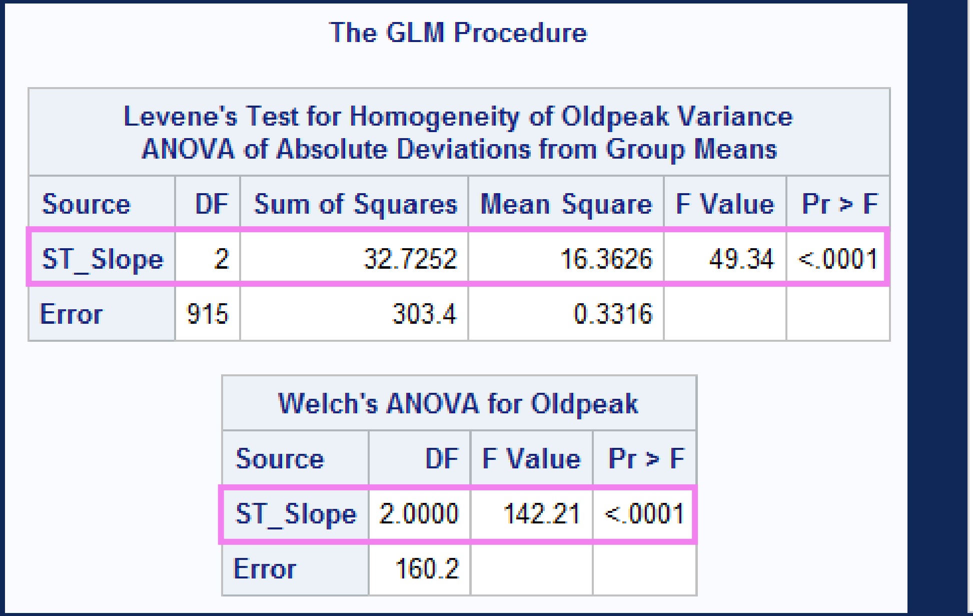
Odds Ratio Estimates

Effect	ChestPainType	Point Estimate	95% Wald Confidence Limits	
Oldpeak	ASY	1.108	0.838	1.466
Oldpeak	ATA	0.383	0.268	0.547
Oldpeak	NAP	0.690	0.506	0.941

ST_SLOPE / OLDPEAK



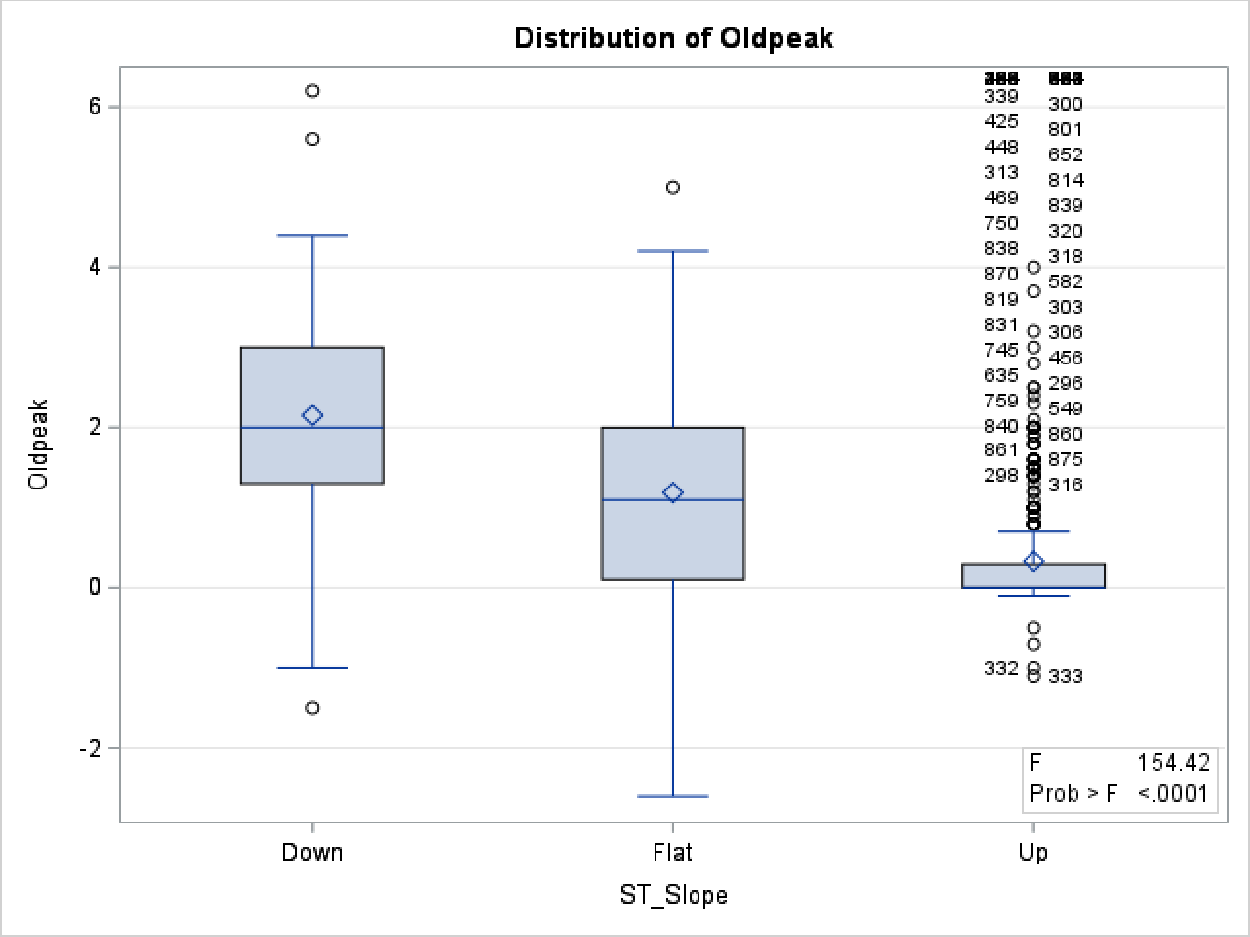
Level of ST_Slope	N	Oldpeak	
		Mean	Std Dev
Down	63	2.15238095	1.40113516
Flat	460	1.18869565	1.01728120
Up	395	0.33468354	0.68188086



The GLM Procedure

Levene's Test for Homogeneity of Oldpeak Variance ANOVA of Absolute Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
ST_Slope	2	32.7252	16.3626	49.34	<.0001
Error	915	303.4	0.3316		

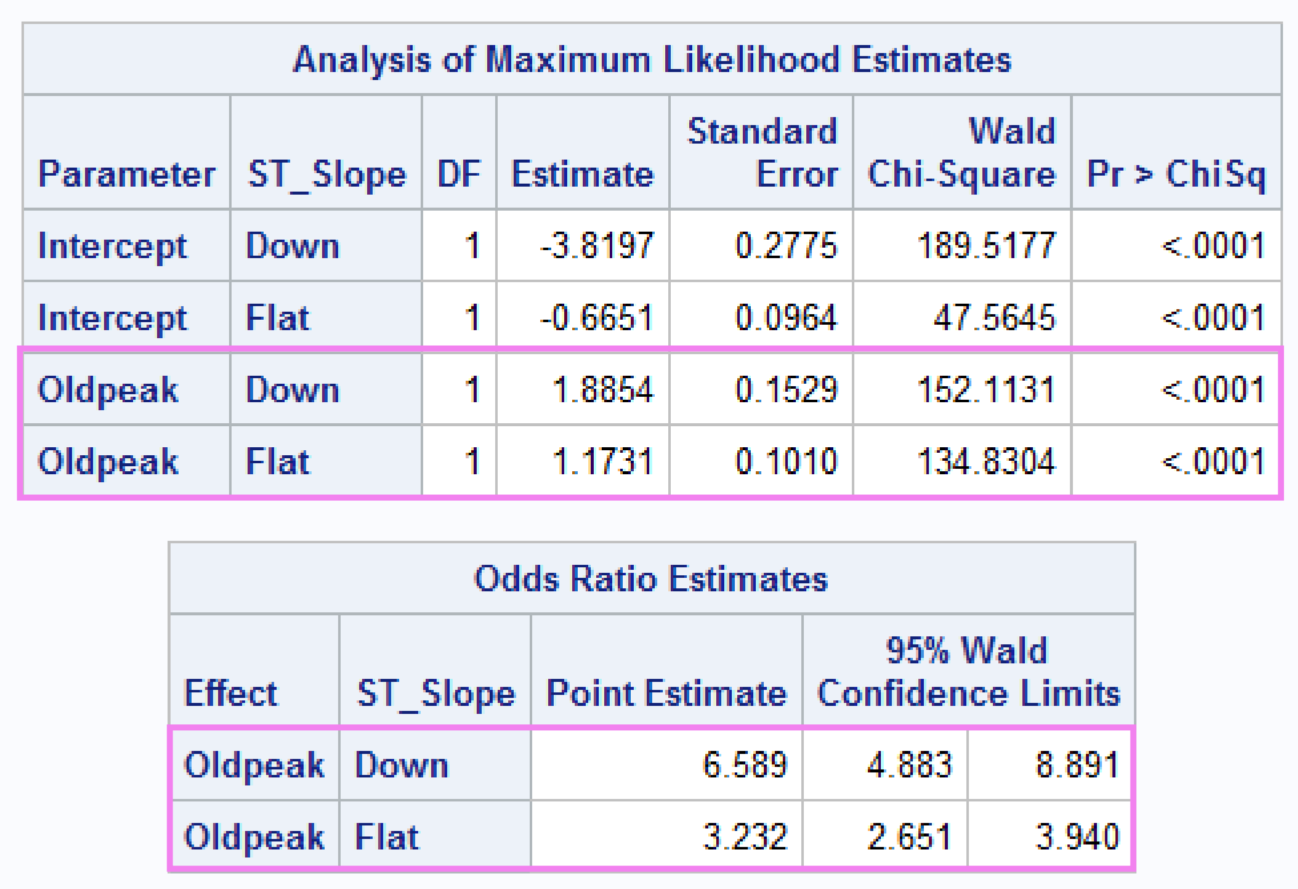
Welch's ANOVA for Oldpeak			
Source	DF	F Value	Pr > F
ST_Slope	2.0000	142.21	<.0001
Error	160.2		



Distribution of Oldpeak

ST_SLOPE / OLDPEAK

- STRONG POSITIVE RELATIONSHIP



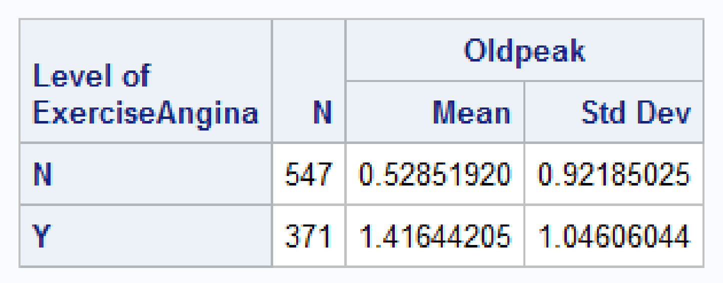
Analysis of Maximum Likelihood Estimates

Parameter	ST_Slope	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	Down	1	-3.8197	0.2775	189.5177	<.0001
Intercept	Flat	1	-0.6651	0.0964	47.5645	<.0001
Oldpeak	Down	1	1.8854	0.1529	152.1131	<.0001
Oldpeak	Flat	1	1.1731	0.1010	134.8304	<.0001

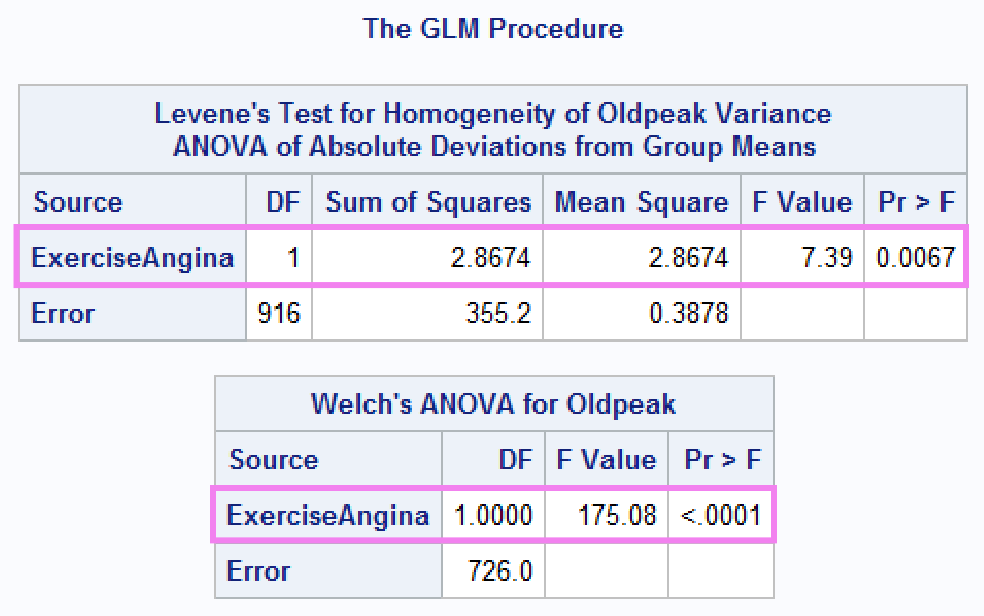
Odds Ratio Estimates

Effect	ST_Slope	Point Estimate	95% Wald Confidence Limits	
Oldpeak	Down	6.589	4.883	8.891
Oldpeak	Flat	3.232	2.651	3.940

EXERCISEANGINA / OLDPEAK



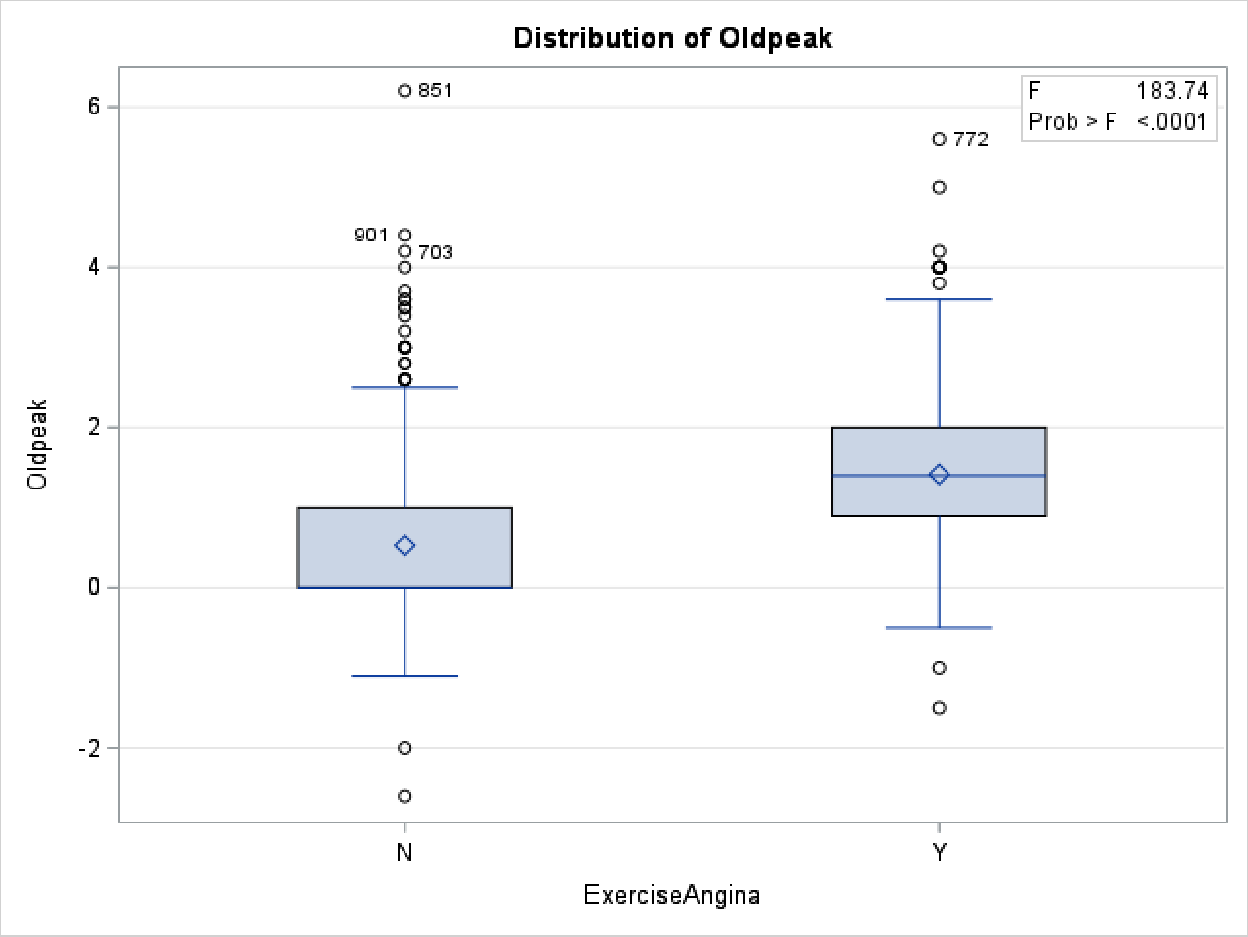
Level of ExerciseAngina	N	Oldpeak	
		Mean	Std Dev
N	547	0.52851920	0.92185025
Y	371	1.41644205	1.04606044



The GLM Procedure

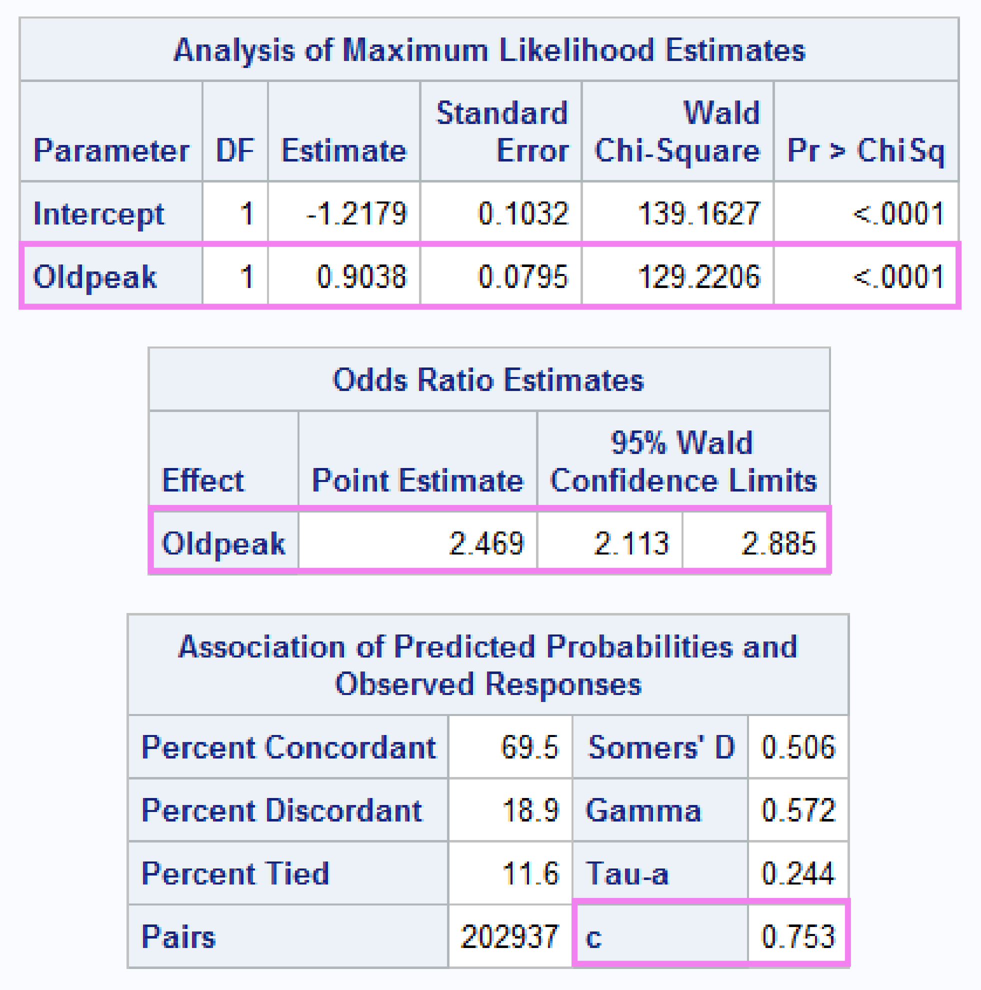
Levene's Test for Homogeneity of Oldpeak Variance ANOVA of Absolute Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
ExerciseAngina	1	2.8674	2.8674	7.39	0.0067
Error	916	355.2	0.3878		

Welch's ANOVA for Oldpeak			
Source	DF	F Value	Pr > F
ExerciseAngina	1.0000	175.08	<.0001
Error	726.0		



EXERCISEANGINA / OLDPEAK

- MODERATE POSITIVE RELATIONSHIP

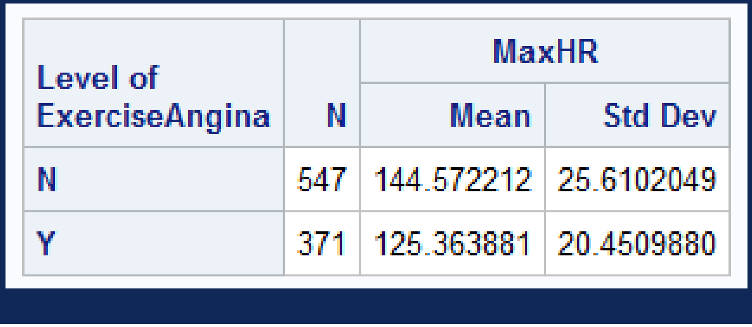


Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.2179	0.1032	139.1627	<.0001
Oldpeak	1	0.9038	0.0795	129.2206	<.0001

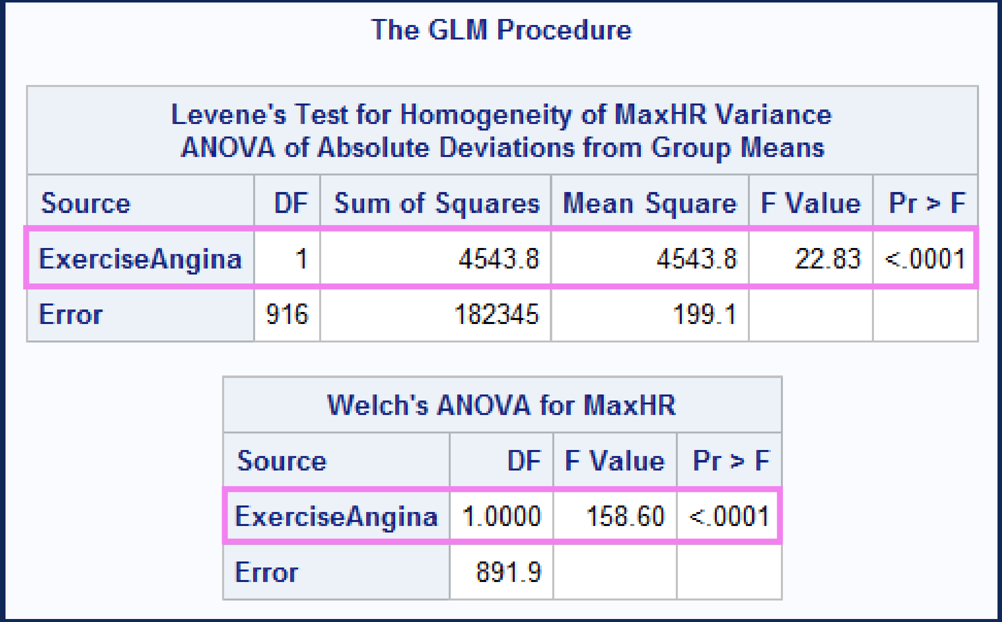
Odds Ratio Estimates			
		95% Wald	
Effect	Point Estimate	Confidence Limits	
Oldpeak	2.469	2.113	2.885

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	69.5	Somers' D	0.506
Percent Discordant	18.9	Gamma	0.572
Percent Tied	11.6	Tau-a	0.244
Pairs	202937	c	0.753

EXERCISEANGINA / MAXHR



Level of ExerciseAngina	N	MaxHR	
		Mean	Std Dev
N	547	144.572212	25.6102049
Y	371	125.363881	20.4509880



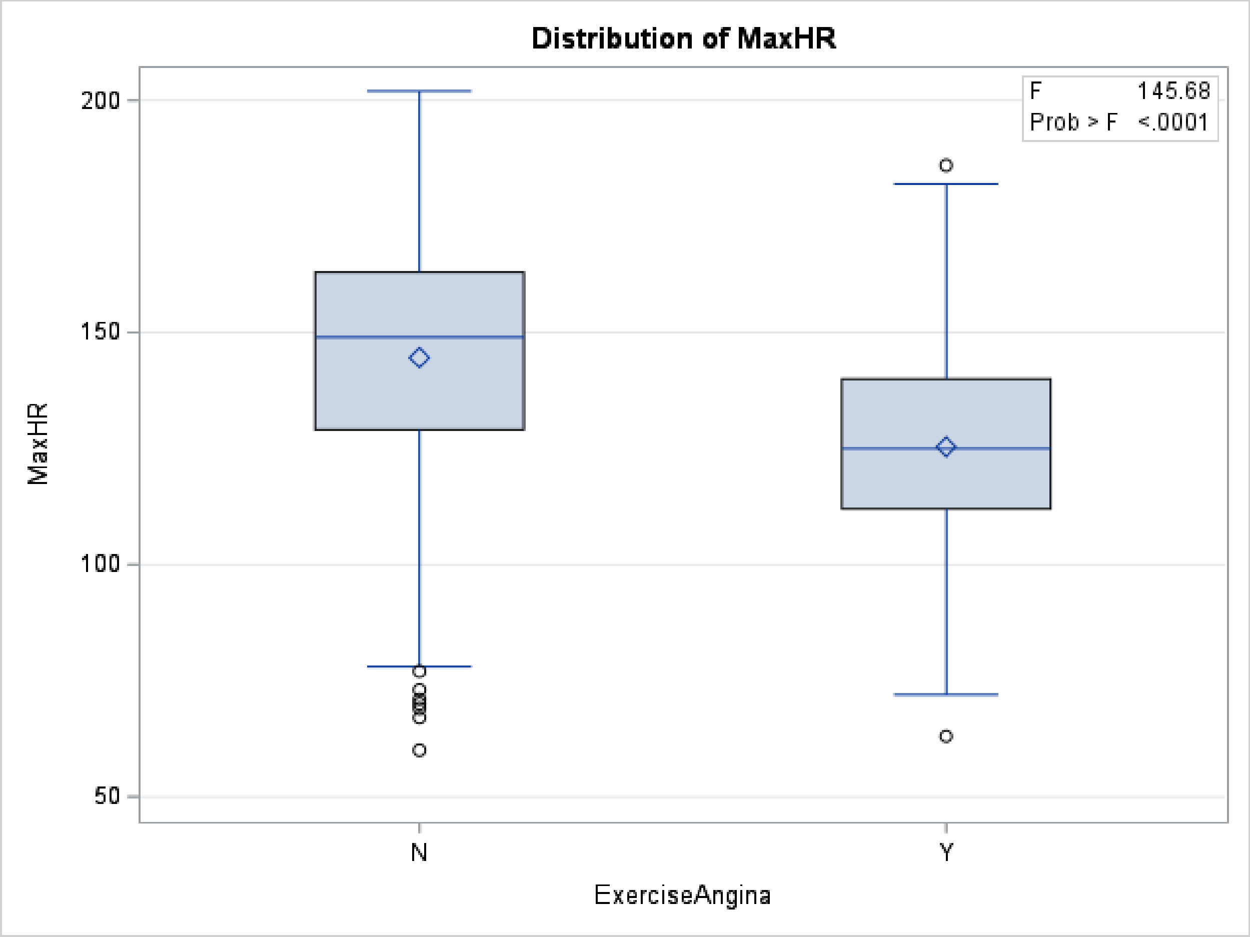
The GLM Procedure

Levene's Test for Homogeneity of MaxHR Variance
ANOVA of Absolute Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
ExerciseAngina	1	4543.8	4543.8	22.83	<.0001
Error	916	182345	199.1		

Welch's ANOVA for MaxHR

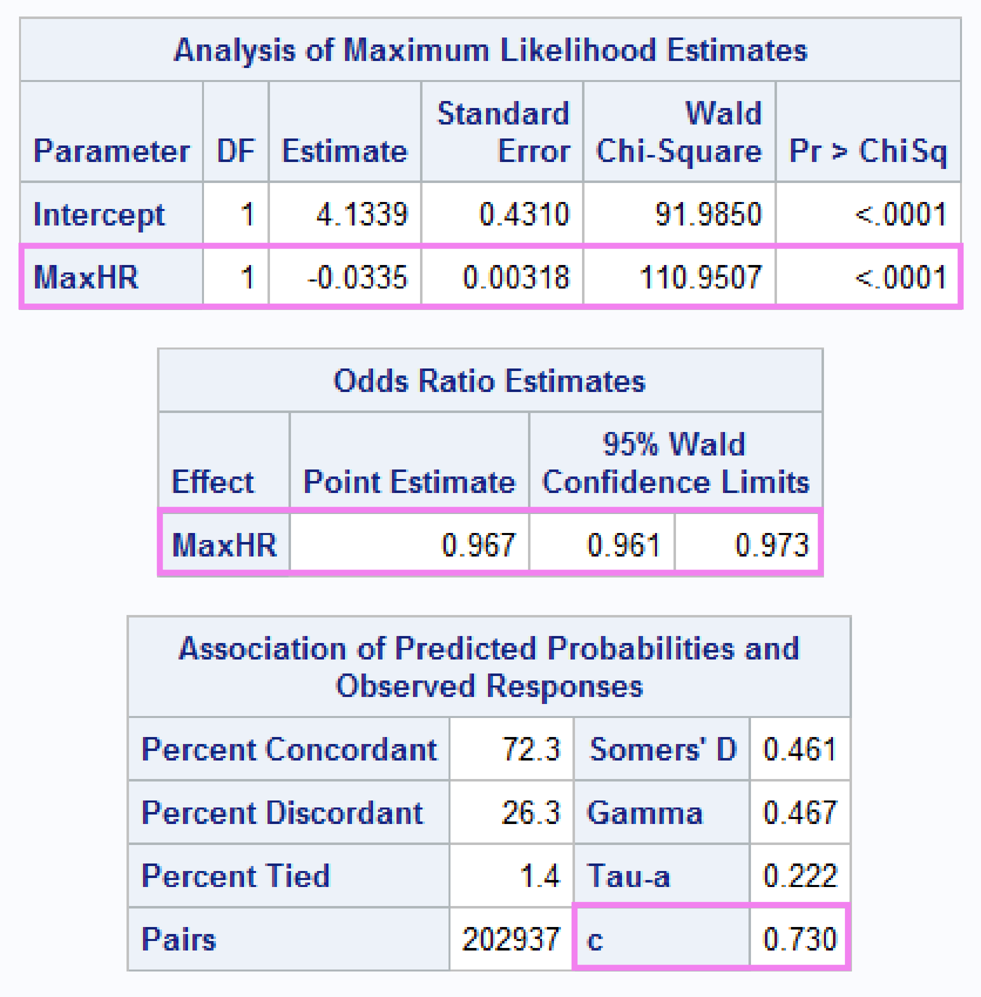
Source	DF	F Value	Pr > F
ExerciseAngina	1.0000	158.60	<.0001
Error	891.9		



Distribution of MaxHR

EXERCISEANGINA / MAXHR

- STRONG NEGATIVE RELATIONSHIP



Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	4.1339	0.4310	91.9850	<.0001
MaxHR	1	-0.0335	0.00318	110.9507	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
MaxHR	0.967	0.961	0.973

Association of Predicted Probabilities and Observed Responses

Percent Concordant	72.3	Somers' D	0.461
Percent Discordant	26.3	Gamma	0.467
Percent Tied	1.4	Tau-a	0.222
Pairs	202937	c	0.730

SIGNIFICANT RELATIONSHIPS

ChestPianType

- Cholesterol
- MaxHR
- **Oldpeak +**
- Age

ExerciseAngina

- **MaxHR -**
- **Oldpeak +**
- RestingBP
- Age

ST_Slope

- Cholesterol
- MaxHR
- **Oldpeak +**
- RestingBP
- Age

FastingBS

- Cholesterol
- MaxHR
- RestingBP
- Age

RestingECG

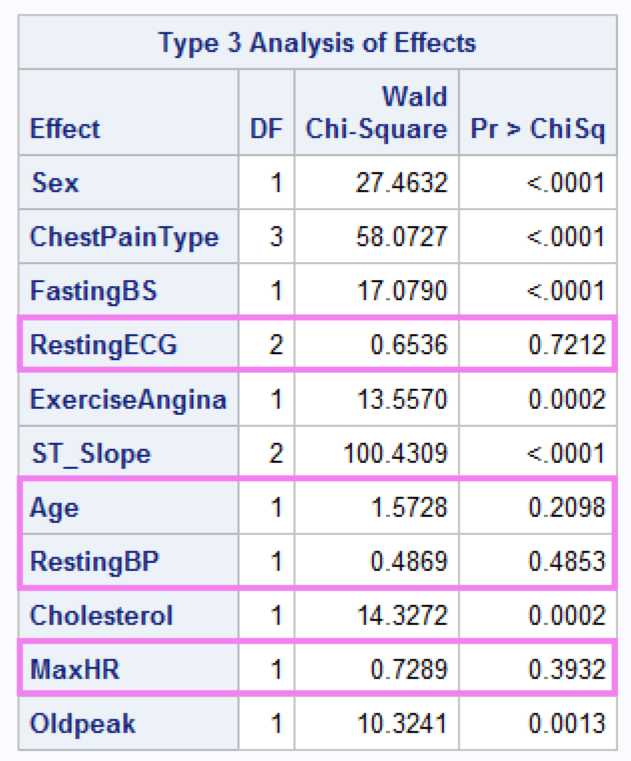
- Cholesterol
- MaxHR
- Oldpeak
- RestingBP
- Age

Sex

- Cholesterol
- MaxHR
- Oldpeak

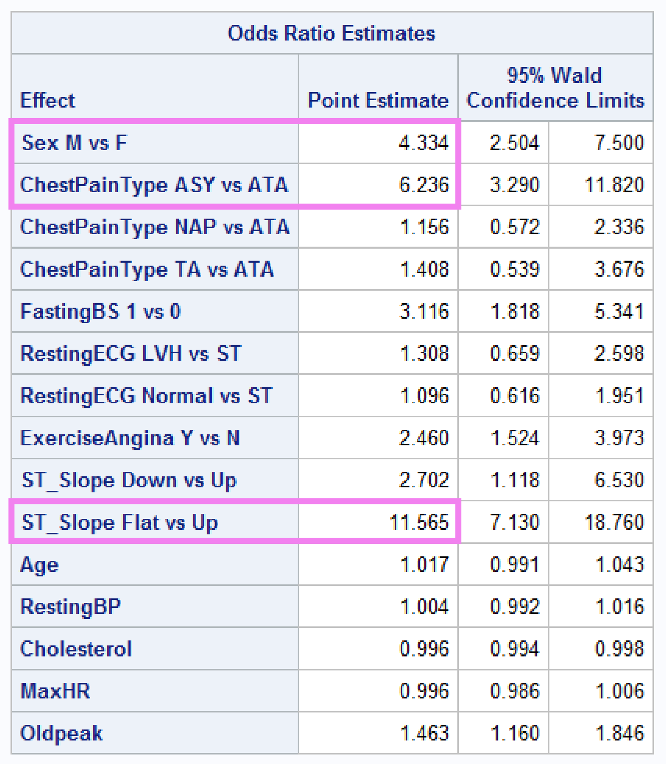
SUMMARY

ST_Slope, ChestPainType, Sex



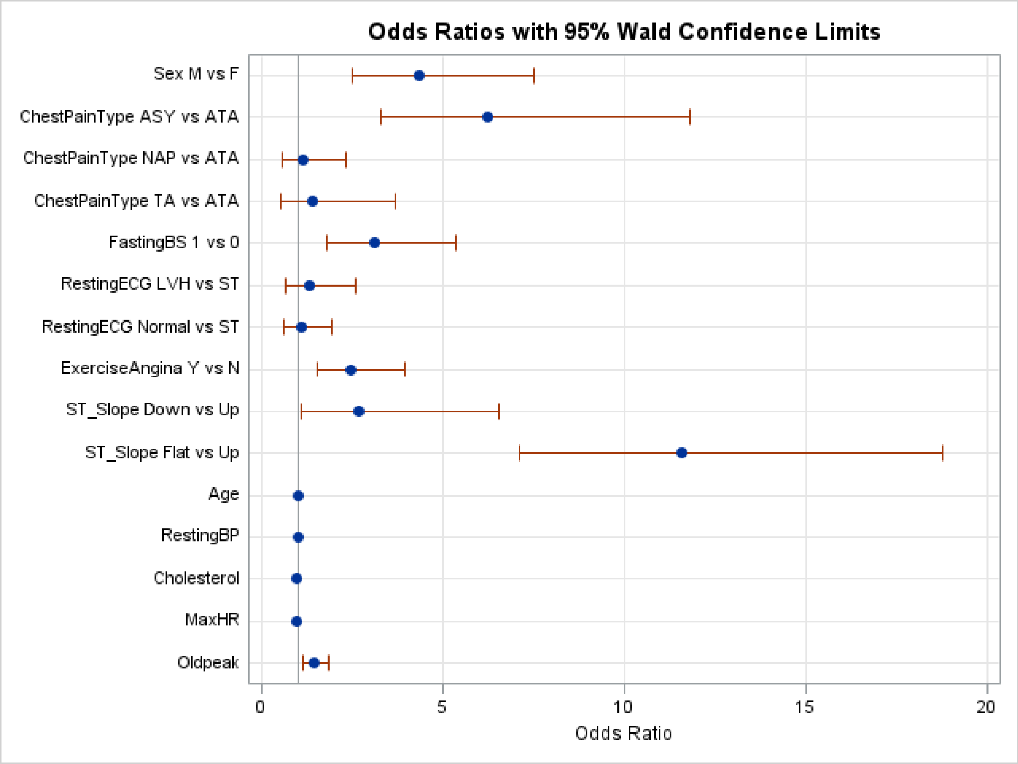
Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Sex	1	27.4632	<.0001
ChestPainType	3	58.0727	<.0001
FastingBS	1	17.0790	<.0001
RestingECG	2	0.6536	0.7212
ExerciseAngina	1	13.5570	0.0002
ST_Slope	2	100.4309	<.0001
Age	1	1.5728	0.2098
RestingBP	1	0.4869	0.4853
Cholesterol	1	14.3272	0.0002
MaxHR	1	0.7289	0.3932
Oldpeak	1	10.3241	0.0013



Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Sex M vs F	4.334	2.504	7.500
ChestPainType ASY vs ATA	6.236	3.290	11.820
ChestPainType NAP vs ATA	1.156	0.572	2.336
ChestPainType TA vs ATA	1.408	0.539	3.676
FastingBS 1 vs 0	3.116	1.818	5.341
RestingECG LVH vs ST	1.308	0.659	2.598
RestingECG Normal vs ST	1.096	0.616	1.951
ExerciseAngina Y vs N	2.460	1.524	3.973
ST_Slope Down vs Up	2.702	1.118	6.530
ST_Slope Flat vs Up	11.565	7.130	18.760
Age	1.017	0.991	1.043
RestingBP	1.004	0.992	1.016
Cholesterol	0.996	0.994	0.998
MaxHR	0.996	0.986	1.006
Oldpeak	1.463	1.160	1.846



FEATURE ENGINEERING

RestingBP

1 missing value

MEDIAN

Cholesterol

172 missing values

18% of dataset

PROC MI

ChestPainType

SEGMENTATION

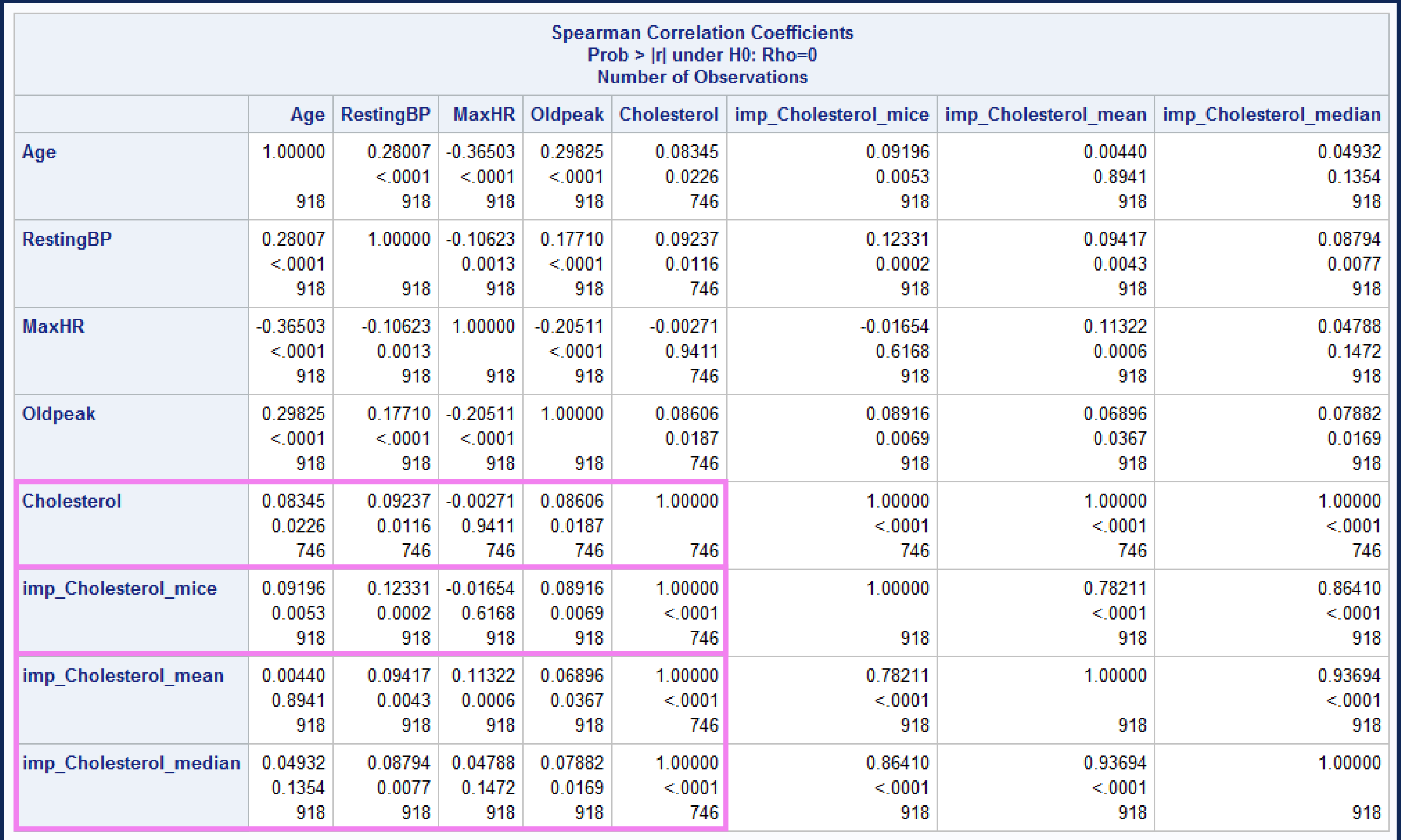
ASY, ATA, NAP+TA

ST_Slope

SEGMENTATION

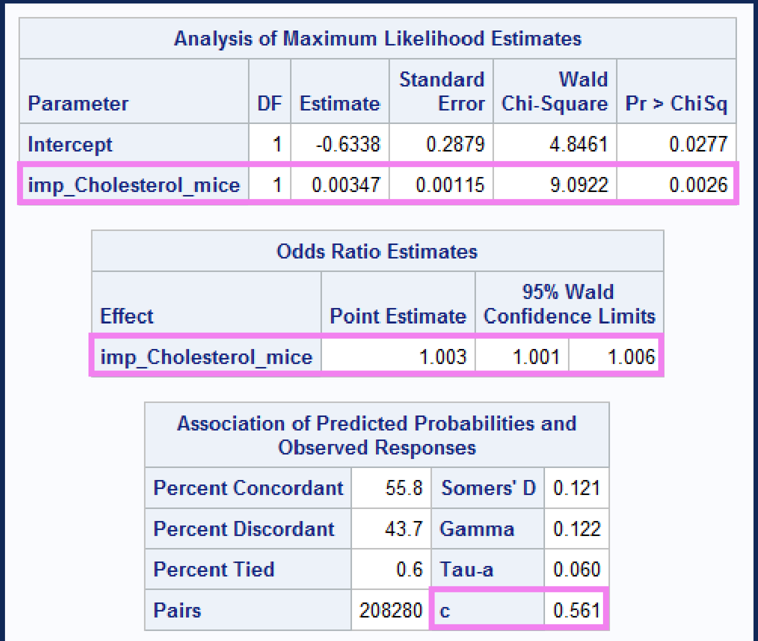
Up, Down+Flat

CHOLESTEROL



Spearman Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Age	RestingBP	MaxHR	Oldpeak	Cholesterol	imp_Cholesterol_mice	imp_Cholesterol_mean	imp_Cholesterol_median
Age	1.00000 918	0.28007 <.0001 918	-0.36503 <.0001 918	0.29825 <.0001 918	0.08345 0.0226 746	0.09196 0.0053 918	0.00440 0.8941 918	0.04932 0.1354 918
RestingBP	0.28007 <.0001 918	1.00000 918	-0.10623 0.0013 918	0.17710 <.0001 918	0.09237 0.0116 746	0.12331 0.0002 918	0.09417 0.0043 918	0.08794 0.0077 918
MaxHR	-0.36503 <.0001 918	-0.10623 0.0013 918	1.00000 918	-0.20511 <.0001 918	-0.00271 0.9411 746	-0.01654 0.6168 918	0.11322 0.0006 918	0.04788 0.1472 918
Oldpeak	0.29825 <.0001 918	0.17710 <.0001 918	-0.20511 <.0001 918	1.00000 918	0.08606 0.0187 746	0.08916 0.0069 918	0.06896 0.0367 918	0.07882 0.0169 918
Cholesterol	0.08345 0.0226 746	0.09237 0.0116 746	-0.00271 0.9411 746	0.08606 0.0187 746	1.00000 746	1.00000 <.0001 746	1.00000 <.0001 746	1.00000 <.0001 746
imp_Cholesterol_mice	0.09196 0.0053 918	0.12331 0.0002 918	-0.01654 0.6168 918	0.08916 0.0069 918	1.00000 <.0001 746	1.00000 918	0.78211 <.0001 918	0.86410 <.0001 918
imp_Cholesterol_mean	0.00440 0.8941 918	0.09417 0.0043 918	0.11322 0.0006 918	0.06896 0.0367 918	1.00000 <.0001 746	0.78211 <.0001 918	1.00000 918	0.93694 <.0001 918
imp_Cholesterol_median	0.04932 0.1354 918	0.08794 0.0077 918	0.04788 0.1472 918	0.07882 0.0169 918	1.00000 <.0001 746	0.86410 <.0001 918	0.93694 <.0001 918	1.00000 918



Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.6338	0.2879	4.8461	0.0277
imp_Cholesterol_mice	1	0.00347	0.00115	9.0922	0.0026

Odds Ratio Estimates

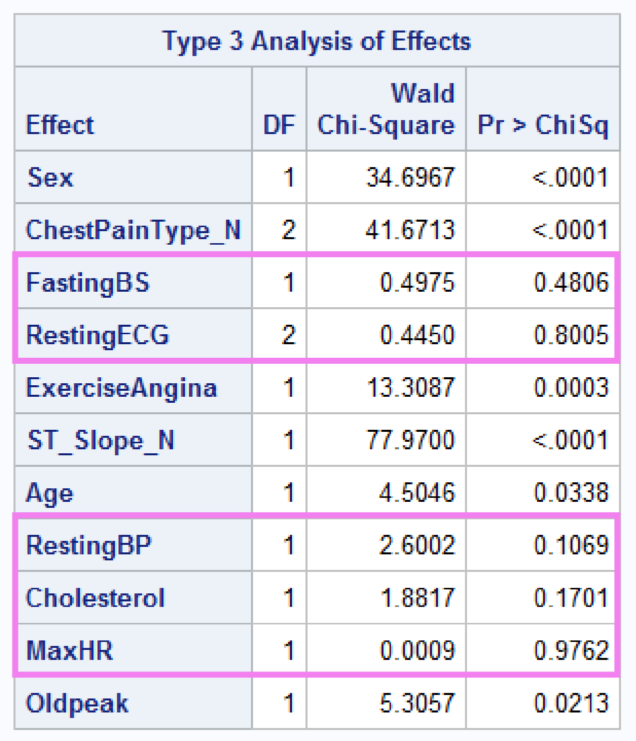
Effect	Point Estimate	95% Wald Confidence Limits	
imp_Cholesterol_mice	1.003	1.001	1.006

Association of Predicted Probabilities and Observed Responses

Percent Concordant	55.8	Somers' D	0.121
Percent Discordant	43.7	Gamma	0.122
Percent Tied	0.6	Tau-a	0.060
Pairs	208280	c	0.561

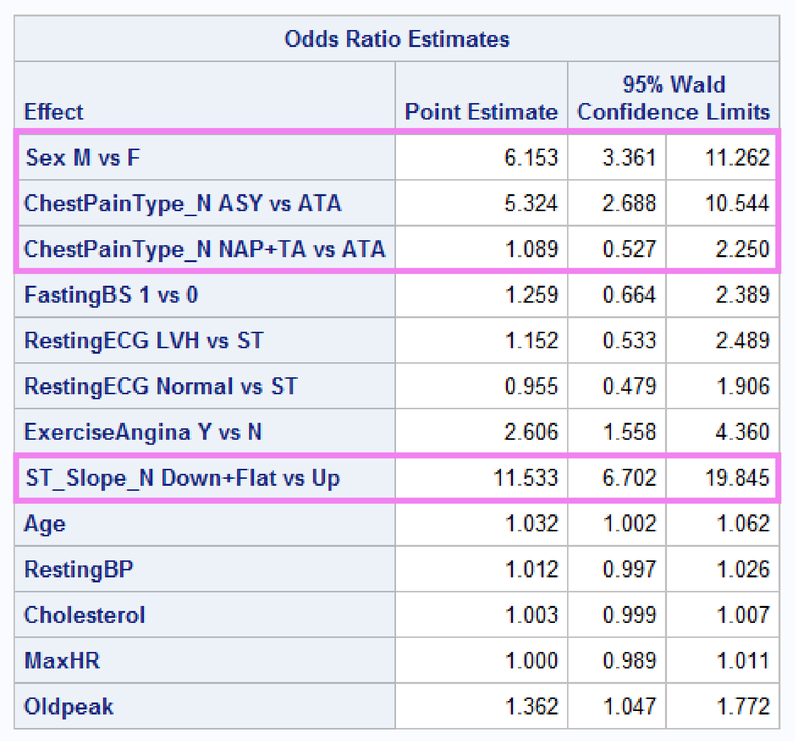
SUMMARY

ST_Slope, Sex, ChestPainType



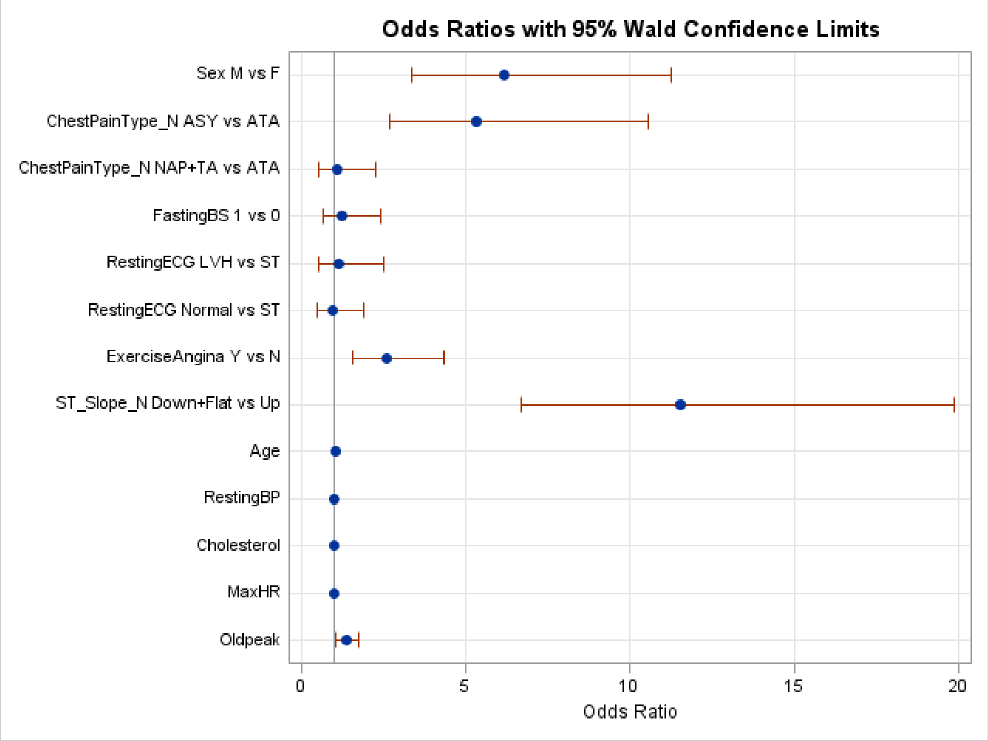
Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Sex	1	34.6967	<.0001
ChestPainType_N	2	41.6713	<.0001
FastingBS	1	0.4975	0.4806
RestingECG	2	0.4450	0.8005
ExerciseAngina	1	13.3087	0.0003
ST_Slope_N	1	77.9700	<.0001
Age	1	4.5046	0.0338
RestingBP	1	2.6002	0.1069
Cholesterol	1	1.8817	0.1701
MaxHR	1	0.0009	0.9762
Oldpeak	1	5.3057	0.0213



Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Sex M vs F	6.153	3.361	11.262
ChestPainType_N ASY vs ATA	5.324	2.688	10.544
ChestPainType_N NAP+TA vs ATA	1.089	0.527	2.250
FastingBS 1 vs 0	1.259	0.664	2.389
RestingECG LVH vs ST	1.152	0.533	2.489
RestingECG Normal vs ST	0.955	0.479	1.906
ExerciseAngina Y vs N	2.606	1.558	4.360
ST_Slope_N Down+Flat vs Up	11.533	6.702	19.845
Age	1.032	1.002	1.062
RestingBP	1.012	0.997	1.026
Cholesterol	1.003	0.999	1.007
MaxHR	1.000	0.989	1.011
Oldpeak	1.362	1.047	1.772



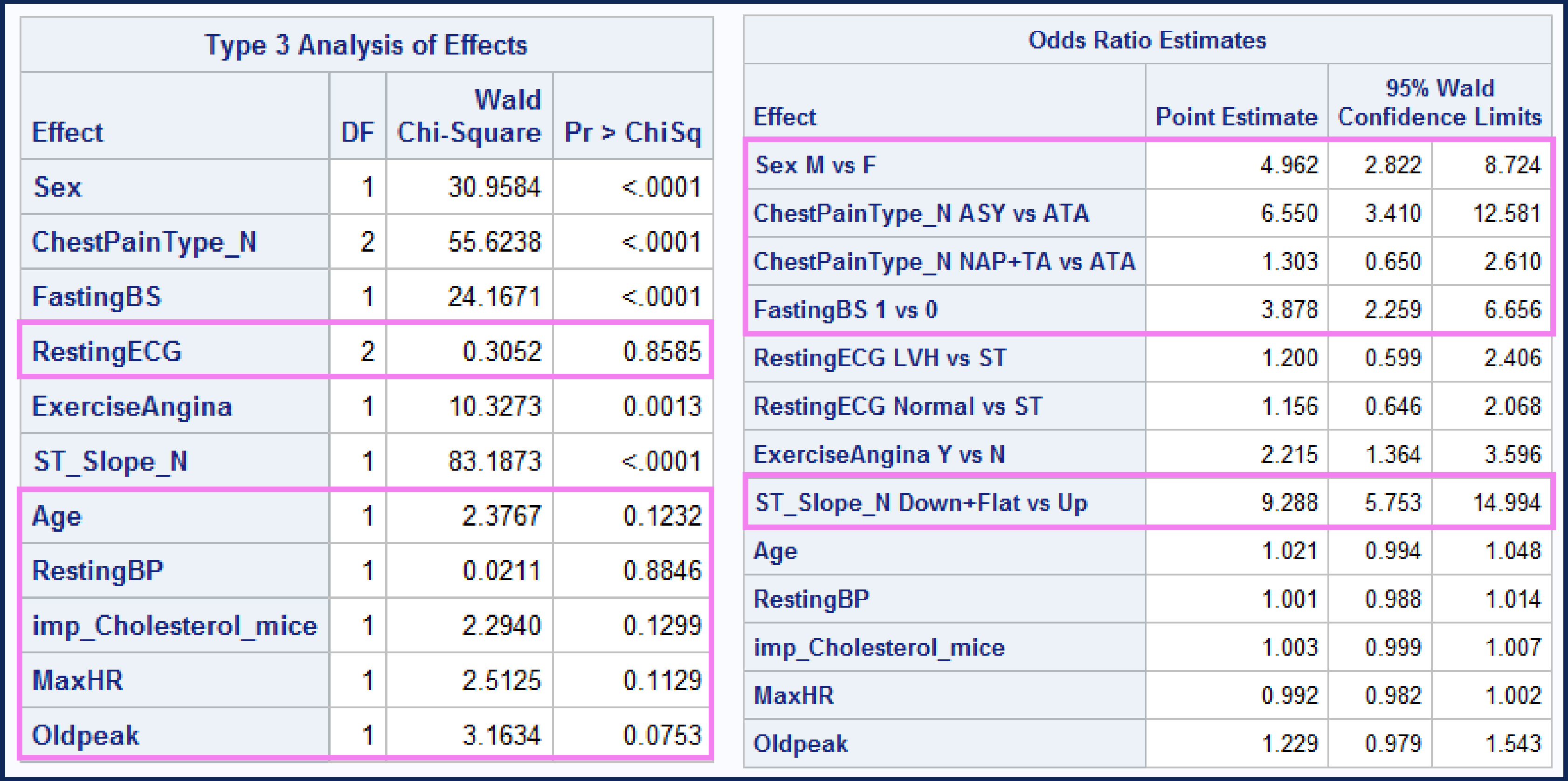
Odds Ratios with 95% Wald Confidence Limits

PREDICTIVE MODELING

LOGISTIC REGRESSION
DECISION TREE

LOGISTIC REGRESSION

ST_SLOPE, CHESTPAINTYPE, SEX, FASTINGBS



Type 3 Analysis of Effects

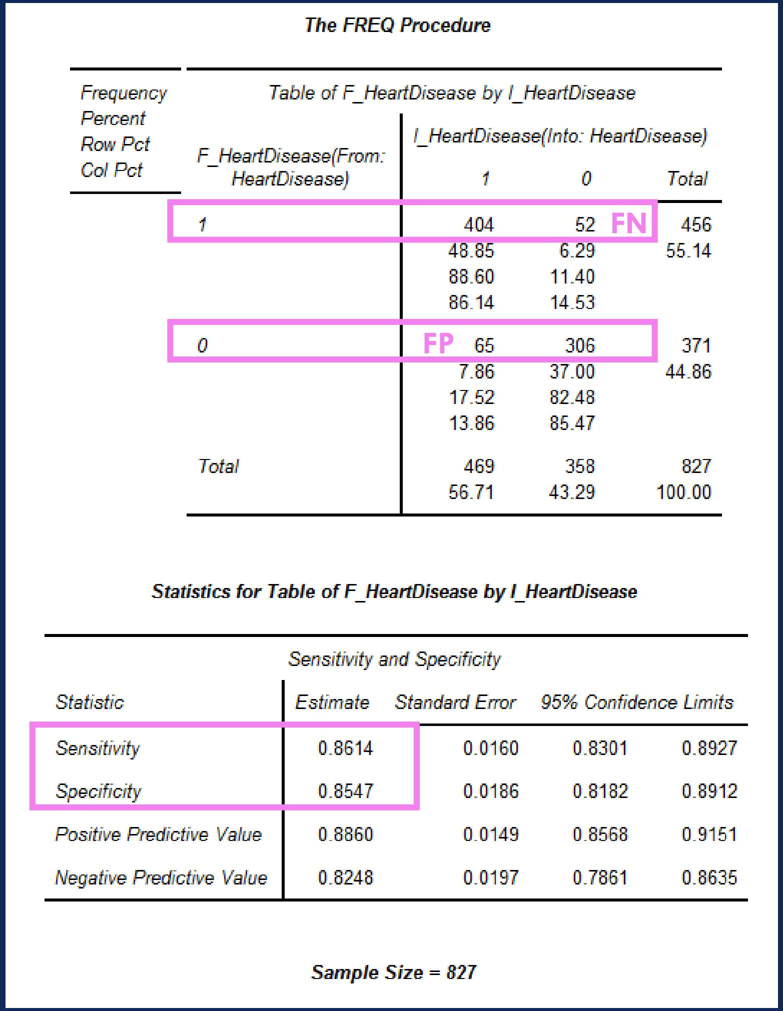
Effect	DF	Wald Chi-Square	Pr > ChiSq
Sex	1	30.9584	<.0001
ChestPainType_N	2	55.6238	<.0001
FastingBS	1	24.1671	<.0001
RestingECG	2	0.3052	0.8585
ExerciseAngina	1	10.3273	0.0013
ST_Slope_N	1	83.1873	<.0001
Age	1	2.3767	0.1232
RestingBP	1	0.0211	0.8846
imp_Cholesterol_mice	1	2.2940	0.1299
MaxHR	1	2.5125	0.1129
Oldpeak	1	3.1634	0.0753

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Sex M vs F	4.962	2.822	8.724
ChestPainType_N ASY vs ATA	6.550	3.410	12.581
ChestPainType_N NAP+TA vs ATA	1.303	0.650	2.610
FastingBS 1 vs 0	3.878	2.259	6.656
RestingECG LVH vs ST	1.200	0.599	2.406
RestingECG Normal vs ST	1.156	0.646	2.068
ExerciseAngina Y vs N	2.215	1.364	3.596
ST_Slope_N Down+Flat vs Up	9.288	5.753	14.994
Age	1.021	0.994	1.048
RestingBP	1.001	0.988	1.014
imp_Cholesterol_mice	1.003	0.999	1.007
MaxHR	0.992	0.982	1.002
Oldpeak	1.229	0.979	1.543

LOGISTIC REGRESSION

TRAIN SET (AUC 0.9273)



The FREQ Procedure

Table of F_HeartDisease by I_HeartDisease

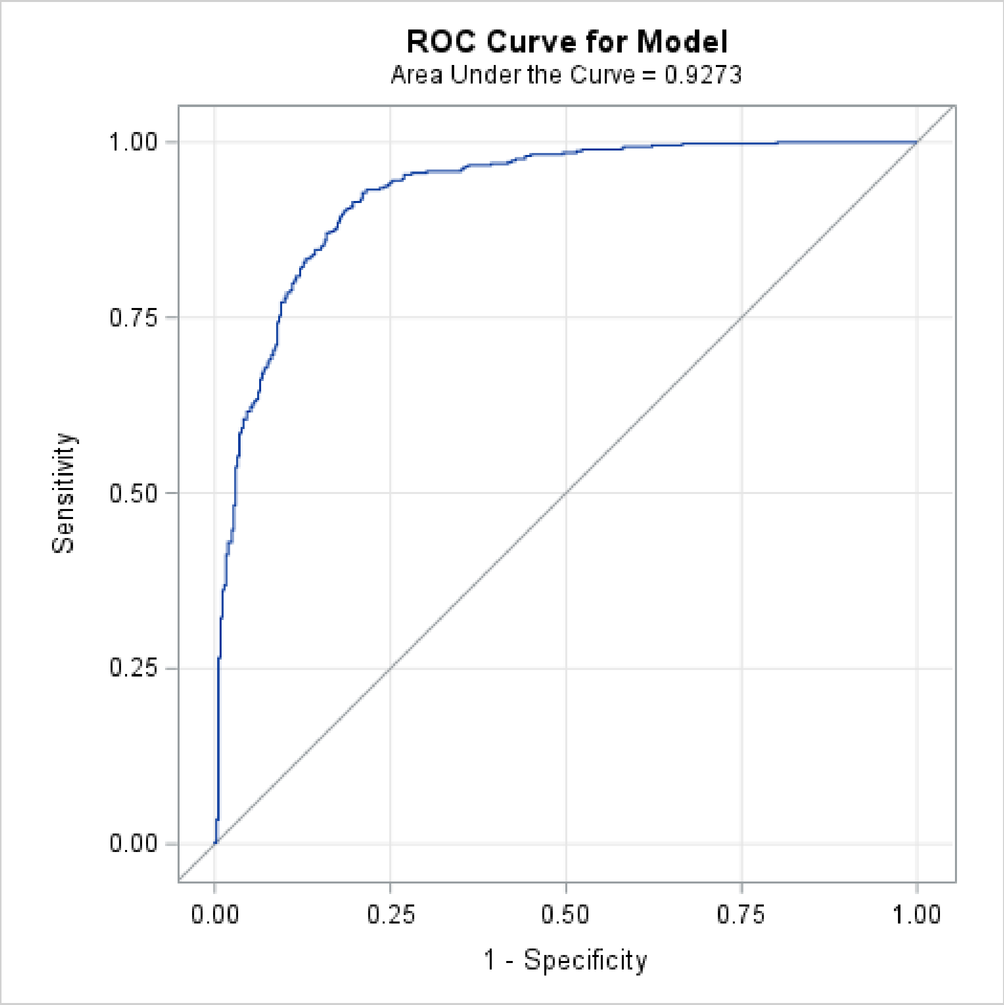
Frequency *Percent* *Row Pct* *Col Pct*		*I_HeartDisease(Into: HeartDisease)*		
	F_HeartDisease(From: HeartDisease)	*1*	*0*	*Total*
	1	404 48.85 88.60 86.14	52 FN 6.29 11.40 14.53	456 55.14
	0	FP 65 7.86 17.52 13.86	306 37.00 82.48 85.47	371 44.86
	Total	469 56.71	358 43.29	827 100.00

Statistics for Table of F_HeartDisease by I_HeartDisease

Sensitivity and Specificity

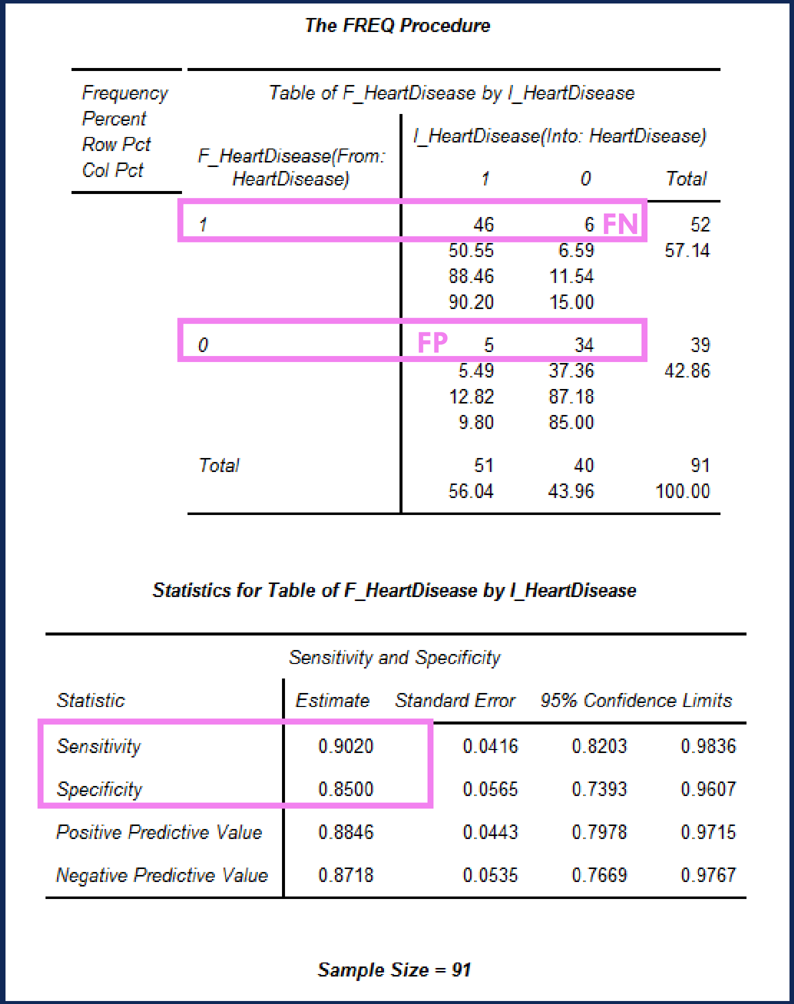
Statistic	*Estimate*	*Standard Error*	*95% Confidence Limits*	
Sensitivity	0.8614	0.0160	0.8301	0.8927
Specificity	0.8547	0.0186	0.8182	0.8912
Positive Predictive Value	0.8860	0.0149	0.8568	0.9151
Negative Predictive Value	0.8248	0.0197	0.7861	0.8635

Sample Size = 827



LOGISTIC REGRESSION

TEST SET (AUC 0.9497)



The FREQ Procedure

Table of F_HeartDisease by I_HeartDisease

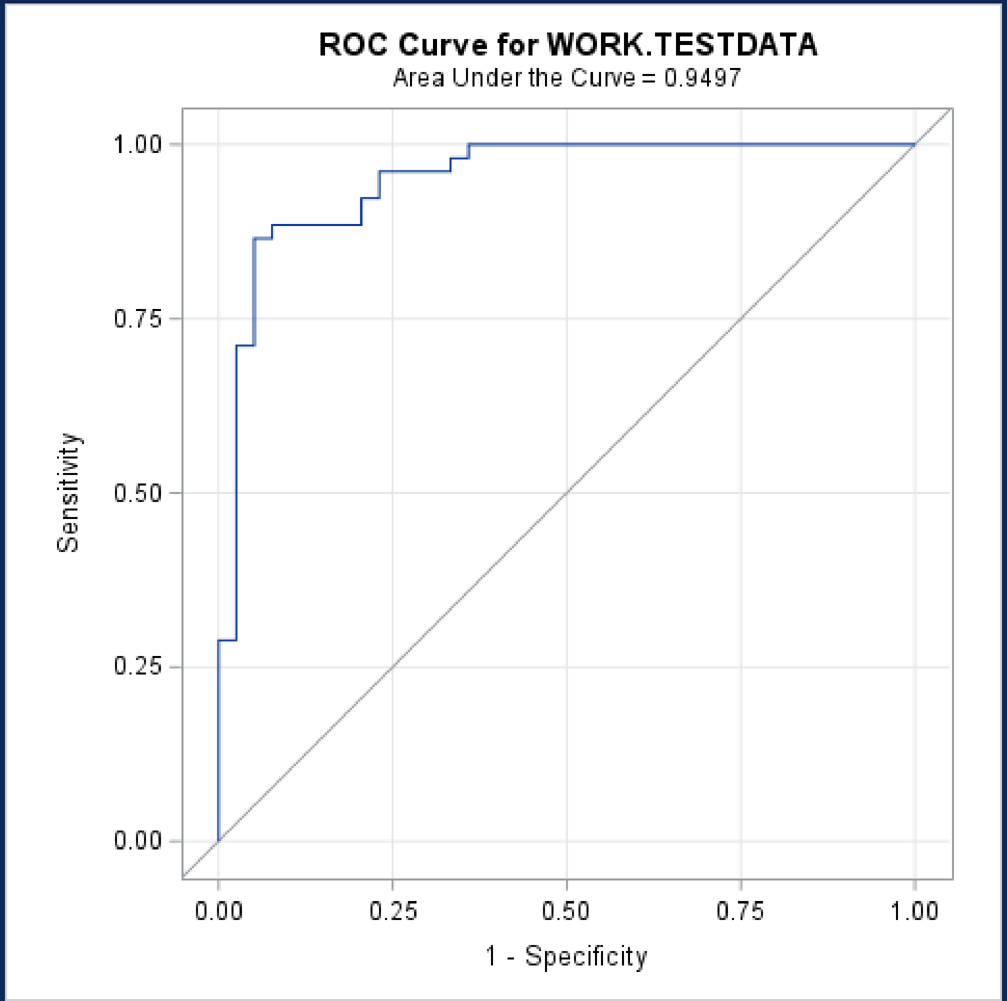
Frequency Percent Row Pct Col Pct	I_HeartDisease(Into: HeartDisease)		
F_HeartDisease(From: HeartDisease)	1	0	Total
1	46 50.55 88.46 90.20	6 FN 6.59 11.54 15.00	52 57.14
0	FP 5 5.49 12.82 9.80	34 37.36 87.18 85.00	39 42.86
Total	51 56.04	40 43.96	91 100.00

Statistics for Table of F_HeartDisease by I_HeartDisease

Sensitivity and Specificity

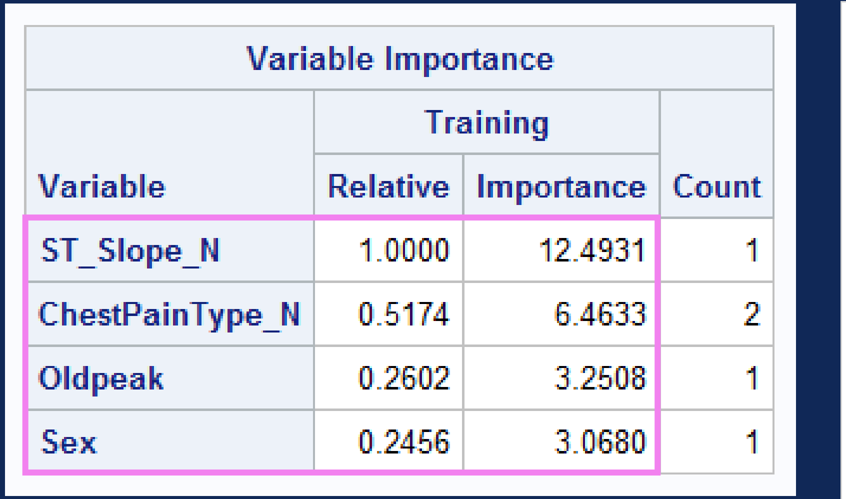
Statistic	Estimate	Standard Error	95% Confidence Limits	
Sensitivity	0.9020	0.0416	0.8203	0.9836
Specificity	0.8500	0.0565	0.7393	0.9607
Positive Predictive Value	0.8846	0.0443	0.7978	0.9715
Negative Predictive Value	0.8718	0.0535	0.7669	0.9767

Sample Size = 91



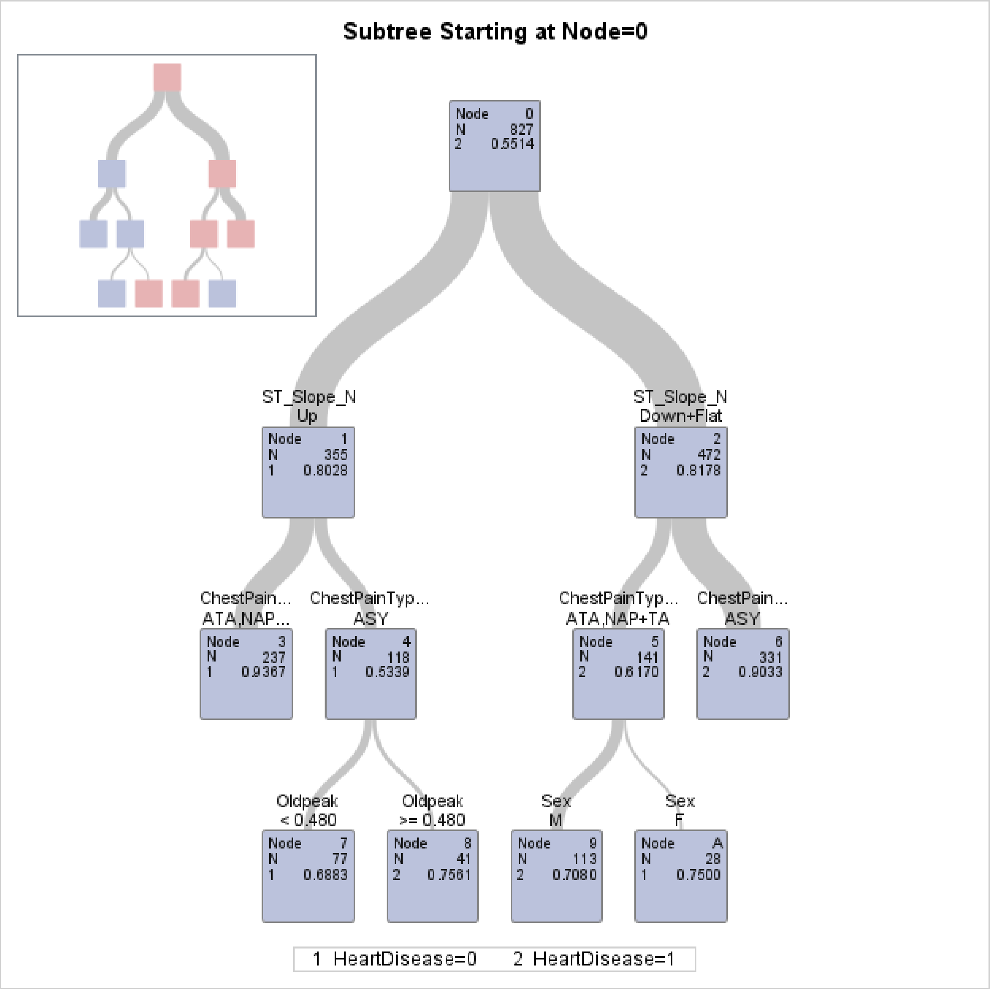
DECISION TREE

ST_SLOPE, CHESTPAINTYPE, OLDPEAK, SEX



Variable Importance

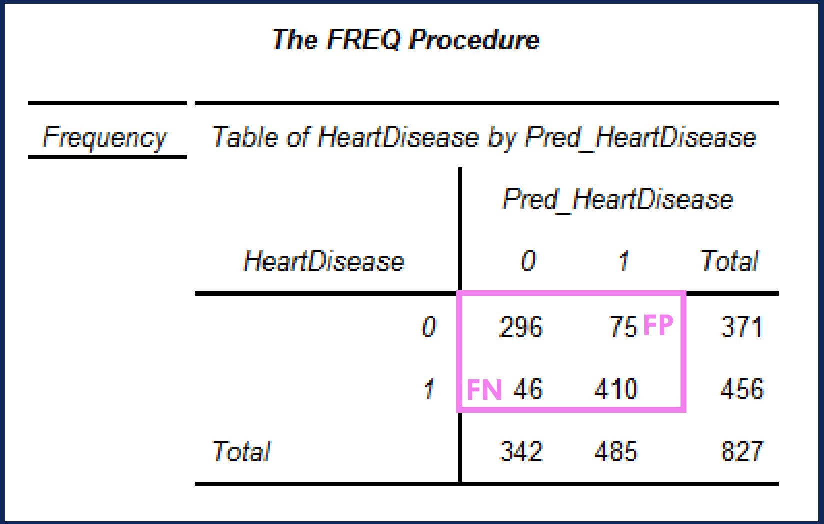
Variable	Training		Count
	Relative	Importance	
ST_Slope_N	1.0000	12.4931	1
ChestPainType_N	0.5174	6.4633	2
Oldpeak	0.2602	3.2508	1
Sex	0.2456	3.0680	1



Subtree Starting at Node=0

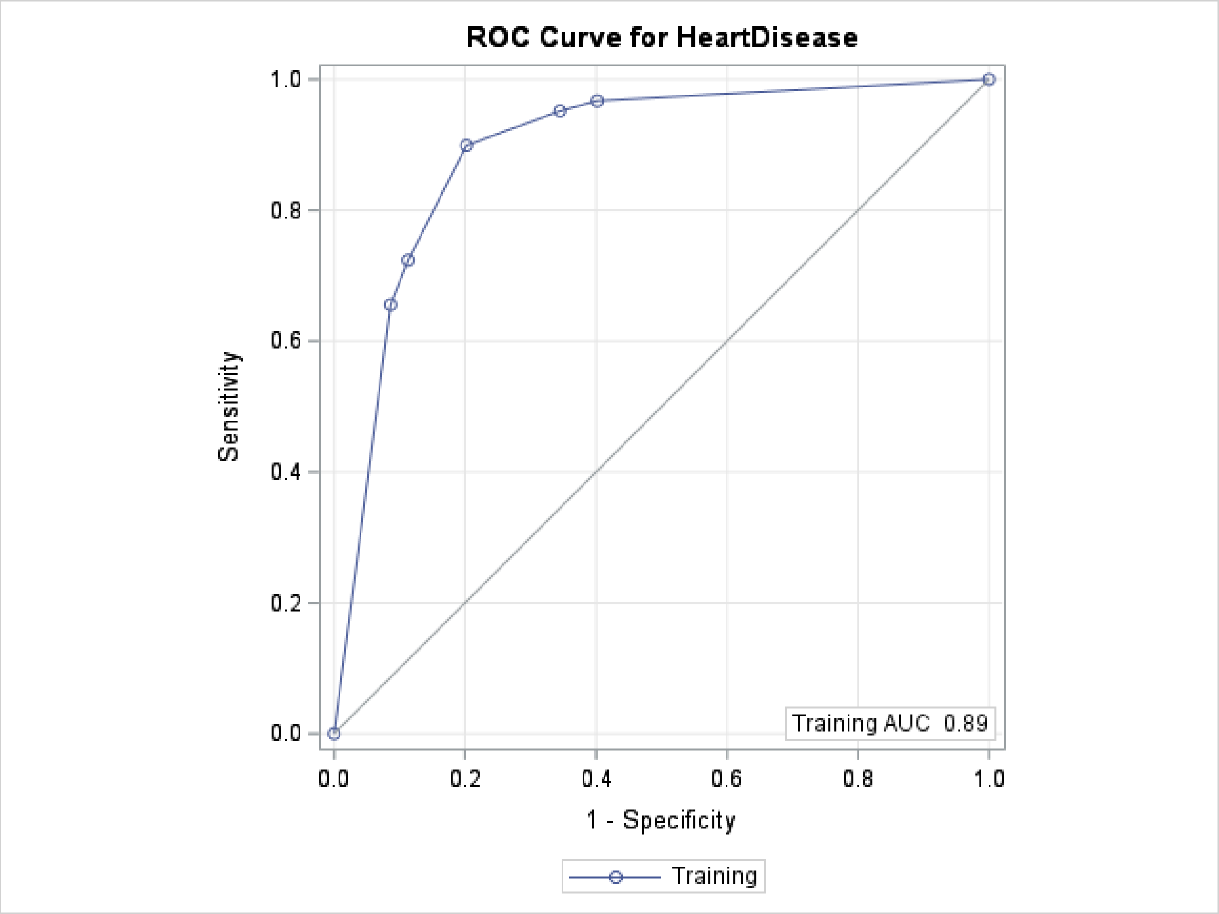
DECISION TREE

TRAIN SET (AUC 0.89)



The FREQ Procedure

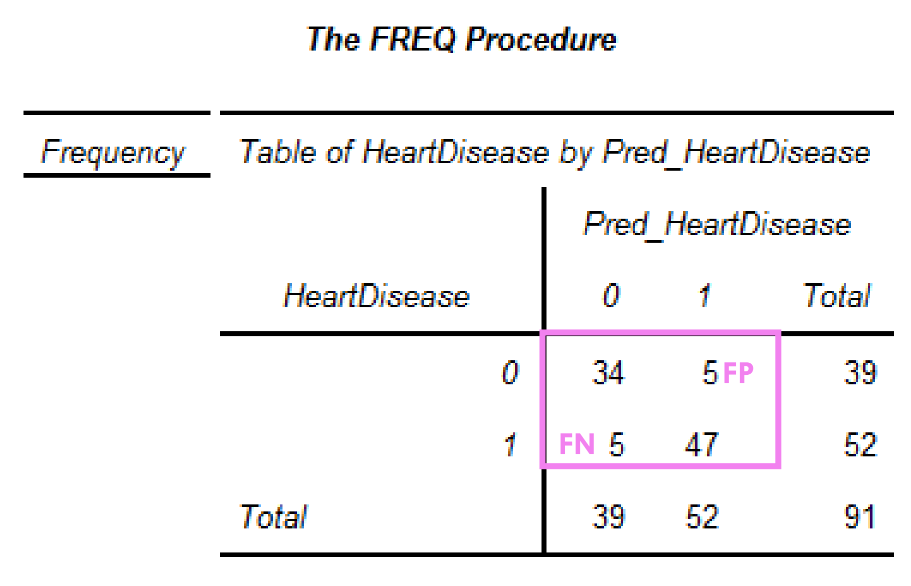
Frequency	*Table of HeartDisease by Pred_HeartDisease*		
	Pred_HeartDisease		
HeartDisease	*0*	*1*	*Total*
0	296	75 FP	371
1	FN 46	410	456
Total	342	485	827



ROC Curve for HeartDisease

DECISION TREE

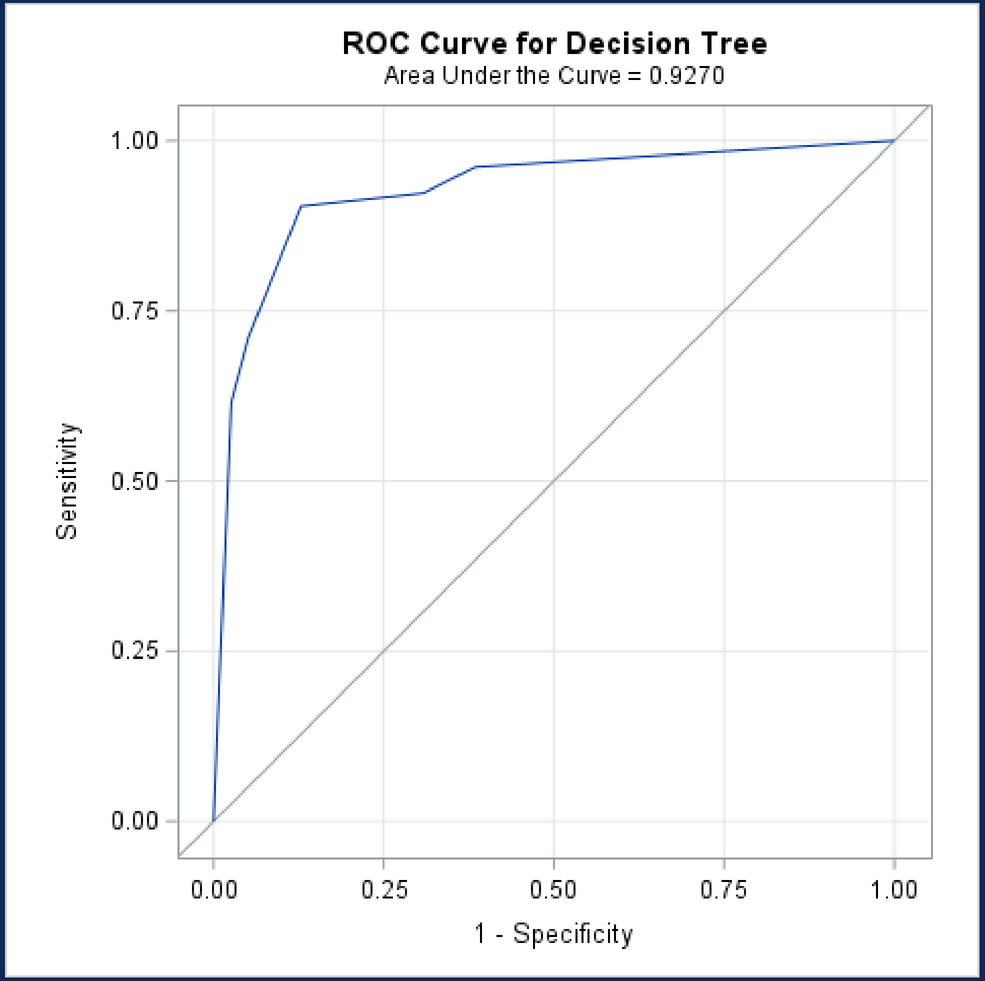
TEST SET (AUC 0.9270)



The FREQ Procedure

Frequency — *Table of HeartDisease by Pred_HeartDisease*

	Pred_HeartDisease		
HeartDisease	*0*	*1*	*Total*
0	34	5 FP	39
1	FN 5	47	52
Total	39	52	91



ROC Curve for Decision Tree

Area Under the Curve = 0.9270

KEY FINDINGS

- Logistic Regression – ST_Slope, Chest Pain Type, Sex, and FastingBS are the strongest predictors of heart disease. Downward/flat **ST_Slope (OR: 9.288)**, **asymptomatic chest pain (OR: 6.550)**, **male sex (OR: 4.962)**, and **high fasting blood sugar (OR: 3.878)** significantly increase the risk.

- Decision Tree - ST_Slope, Chest Pain Type, Oldpeak, and Sex are the most important predictors of heart disease. **ST_Slope (Importance: 12.4931)** is the strongest factor, followed by **Chest Pain Type (6.4633)**, **Oldpeak (3.2508)**, and **Sex (3.0680)**, indicating their significant role in classification.

CONCLUSIONS

- Model Performance Comparison – Logistic regression exhibited overfitting, while the decision tree provided better generalization with fewer false negatives. However, logistic regression still demonstrated strong overall performance.
- Data Enhancement – Increasing the dataset size and incorporating more diverse samples could improve model robustness and accuracy, reducing potential biases.
- Feature Importance – ST_Slope, ChestPainType, and Sex emerged as critical predictors of heart disease. Further investigation into their clinical significance could refine model effectiveness.
- Alternative Models – Exploring ensemble methods like Random Forest or XGBoost may further optimize prediction accuracy and mitigate overfitting issues.

THANK YOU

Li Wu

Instructor: Sun Makosso-Kallyth