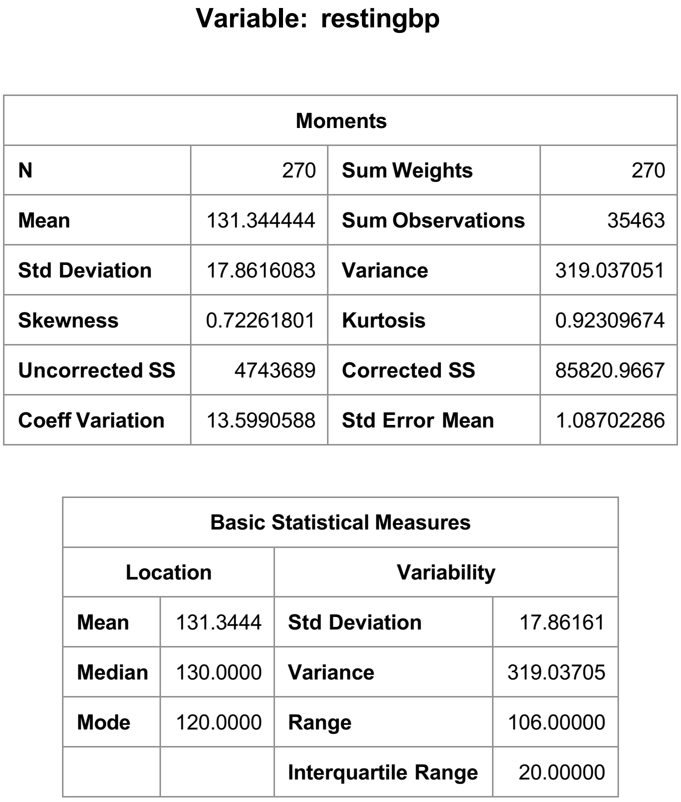
Exercise 1:

a) basic descriptive statistics for **restingbp** for all of the data



Variable: restingbp

Moments			
N	270	**Sum Weights**	270
Mean	131.344444	**Sum Observations**	35463
Std Deviation	17.8616083	**Variance**	319.037051
Skewness	0.72261801	**Kurtosis**	0.92309674
Uncorrected SS	4743689	**Corrected SS**	85820.9667
Coeff Variation	13.5990588	**Std Error Mean**	1.08702286

Basic Statistical Measures			
Location		**Variability**	
Mean	131.3444	**Std Deviation**	17.86161
Median	130.0000	**Variance**	319.03705
Mode	120.0000	**Range**	106.00000
		Interquartile Range	20.00000

While the mean and median are somewhat close at 132.34 and 130, there is a noticeable positive skew of 0.72. The positive skew indicates a tail to the right and the median and IQR should be used to describe location and spread.

Half of resting blood pressure values are expected to be above 130 and half below based on this data, and the difference between 75th and 25th percentiles is expected to be 20.

b) The same analysis as in part a) by **heartdisease variable.**

heartdisease=absence

Moments			
N	150	**Sum Weights**	150
Mean	128.866667	**Sum Observations**	19330
Std Deviation	16.4576604	**Variance**	270.854586
Skewness	0.41352706	**Kurtosis**	0.26136044
Uncorrected SS	2531350	**Corrected SS**	40357.3333
Coeff Variation	12.7710764	**Std Error Mean**	1.34376235

Basic Statistical Measures			
Location		Variability	
Mean	128.8667	**Std Deviation**	16.45766
Median	130.0000	**Variance**	270.85459
Mode	120.0000	**Range**	86.00000
		Interquartile Range	20.00000

In the absence of heartdisease we see that the mean and median are somewhat close at 128.89 and 130 (same as before), there is a some positive skew of 0.41. The positive skew indicates a tail to the right and the median and IQR should be used to describe location and spread.

Half of resting blood pressure values are expected to be above 130 and half below based on this data, and the difference between 75th and 25th percentiles is expected to be 20 (same as before).

heartdisease=presence

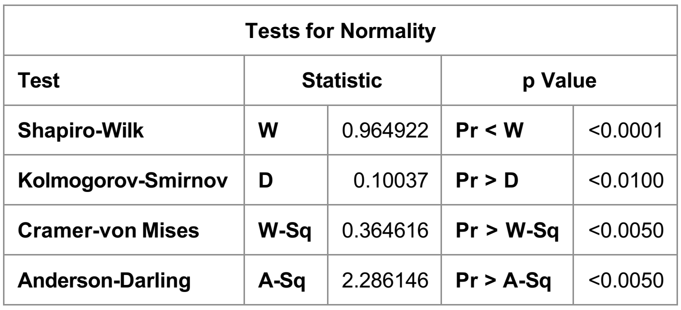
Moments			
N	120	**Sum Weights**	120
Mean	134.441667	**Sum Observations**	16133
Std Deviation	19.0954242	**Variance**	364.635224
Skewness	0.8886235	**Kurtosis**	0.96280272
Uncorrected SS	2212339	**Corrected SS**	43391.5917
Coeff Variation	14.2035015	**Std Error Mean**	1.74316576

Basic Statistical Measures			
Location		**Variability**	
Mean	134.4417	**Std Deviation**	19.09542
Median	130.0000	**Variance**	364.63522
Mode	120.0000	**Range**	100.00000
		Interquartile Range	25.00000

Note: The mode displayed is the smallest of 2 modes with a count of 13.

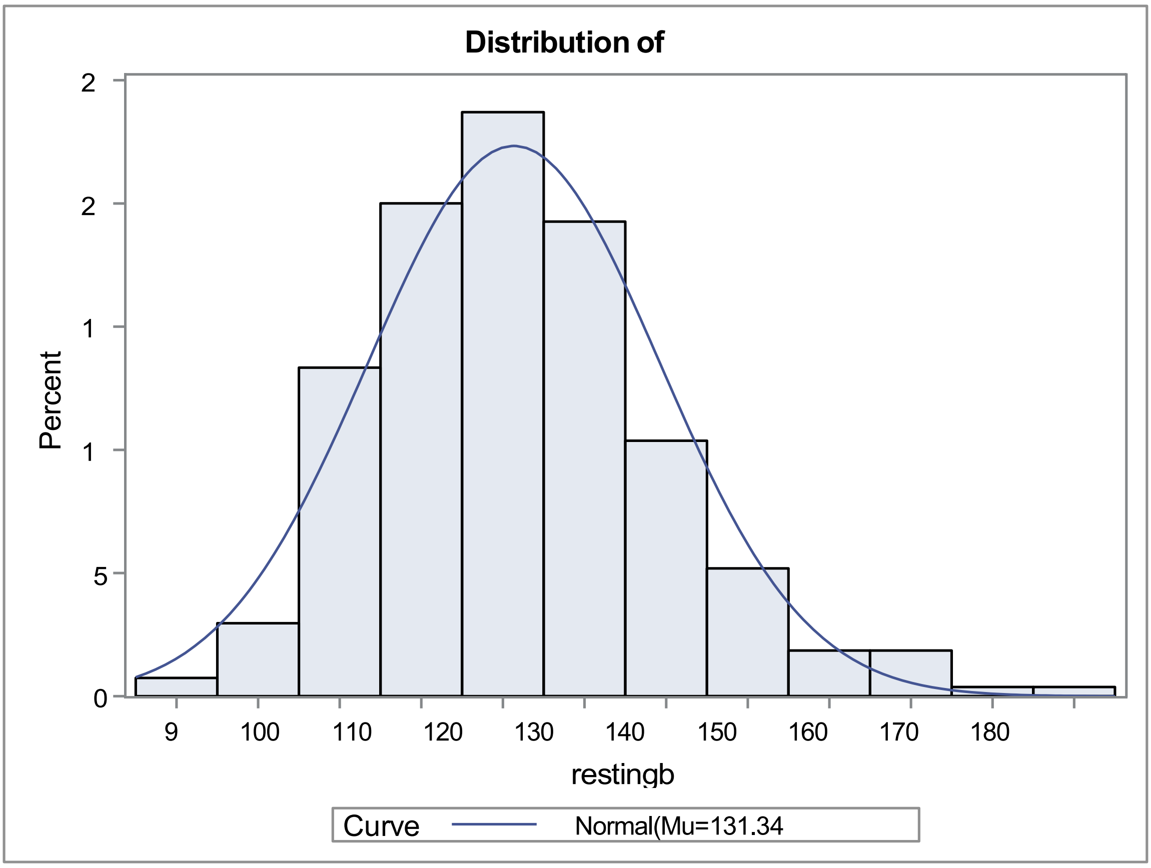
In the presence of **heartdisease** we see that the mean and median are somewhat close at 134.44 (similar to part a) and 130 (same as before), there is a high positive skew of 0.889. The positive skew indicates a tail to the right and the median and IQR should be used to describe location and spread.
Half of resting blood pressure values are expected to be above 130 and half below based on this data, and the difference between 75th and 25th percentiles is expected to be 25.

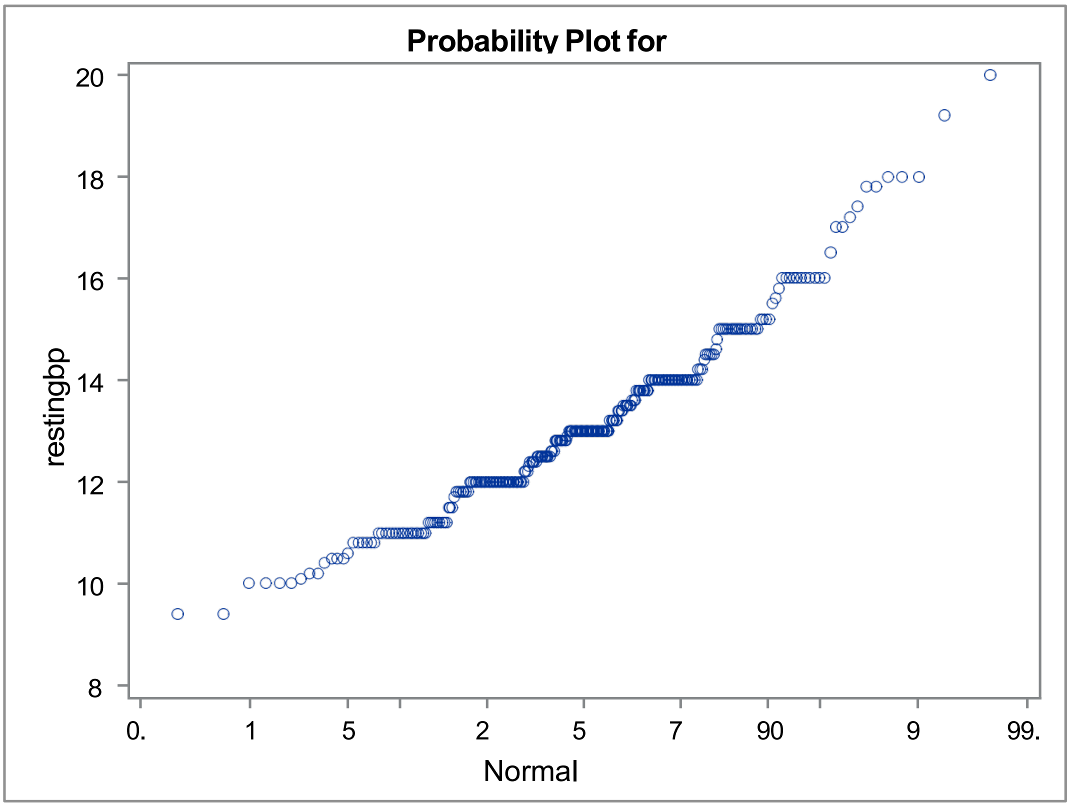
Exercise 2:



Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.964922	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.10037	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.364616	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	2.286146	**Pr > A-Sq**	<0.0050

a) Resting blood pressure in general appear to be far from normal. All the hypothesis tests also reject a null normal distribution. However, the histogram looks bell-shaped, and the probability appears to be not so straight. So, an assumption of normality is not reasonable for resting blood pressure in general.



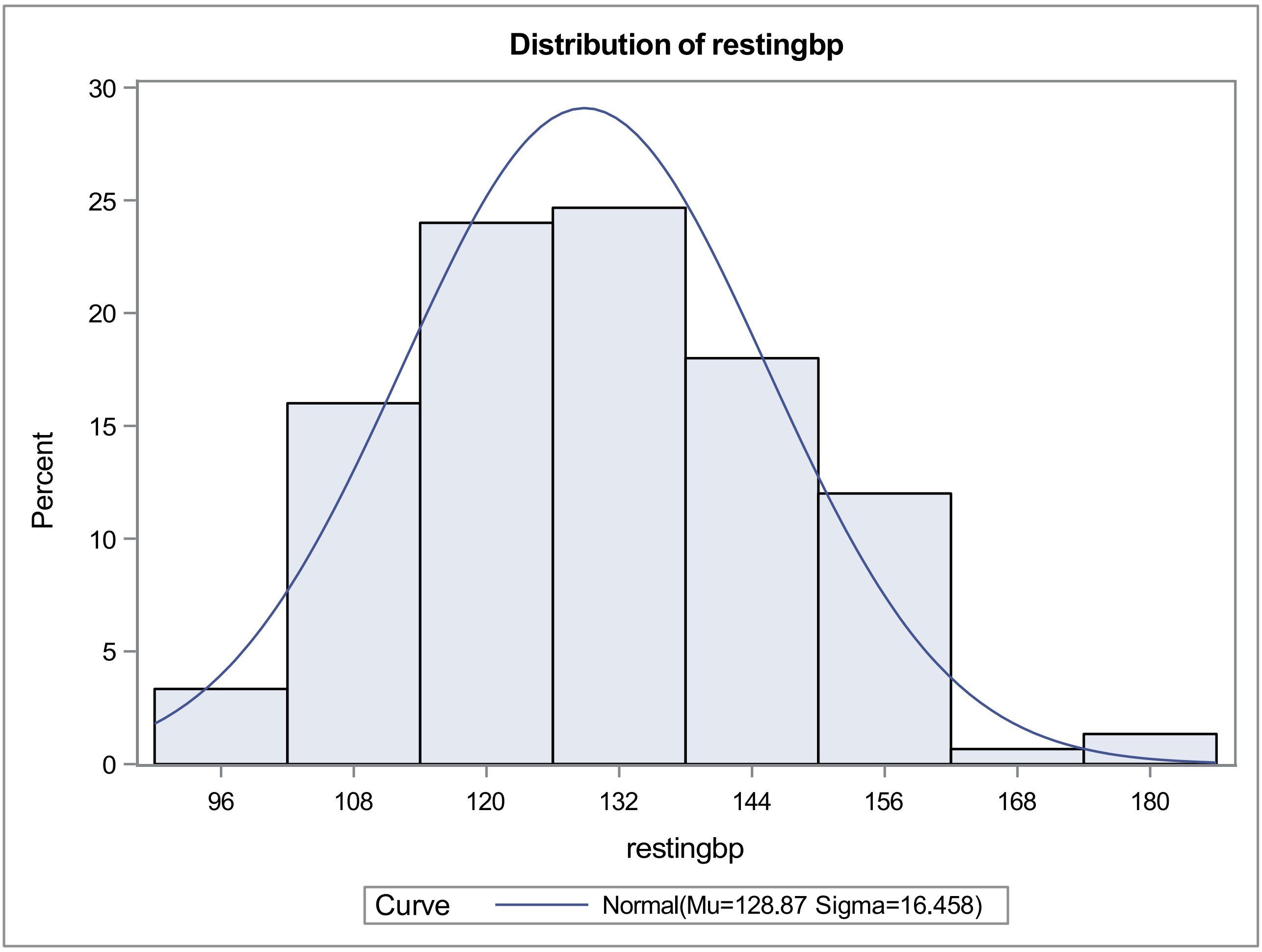


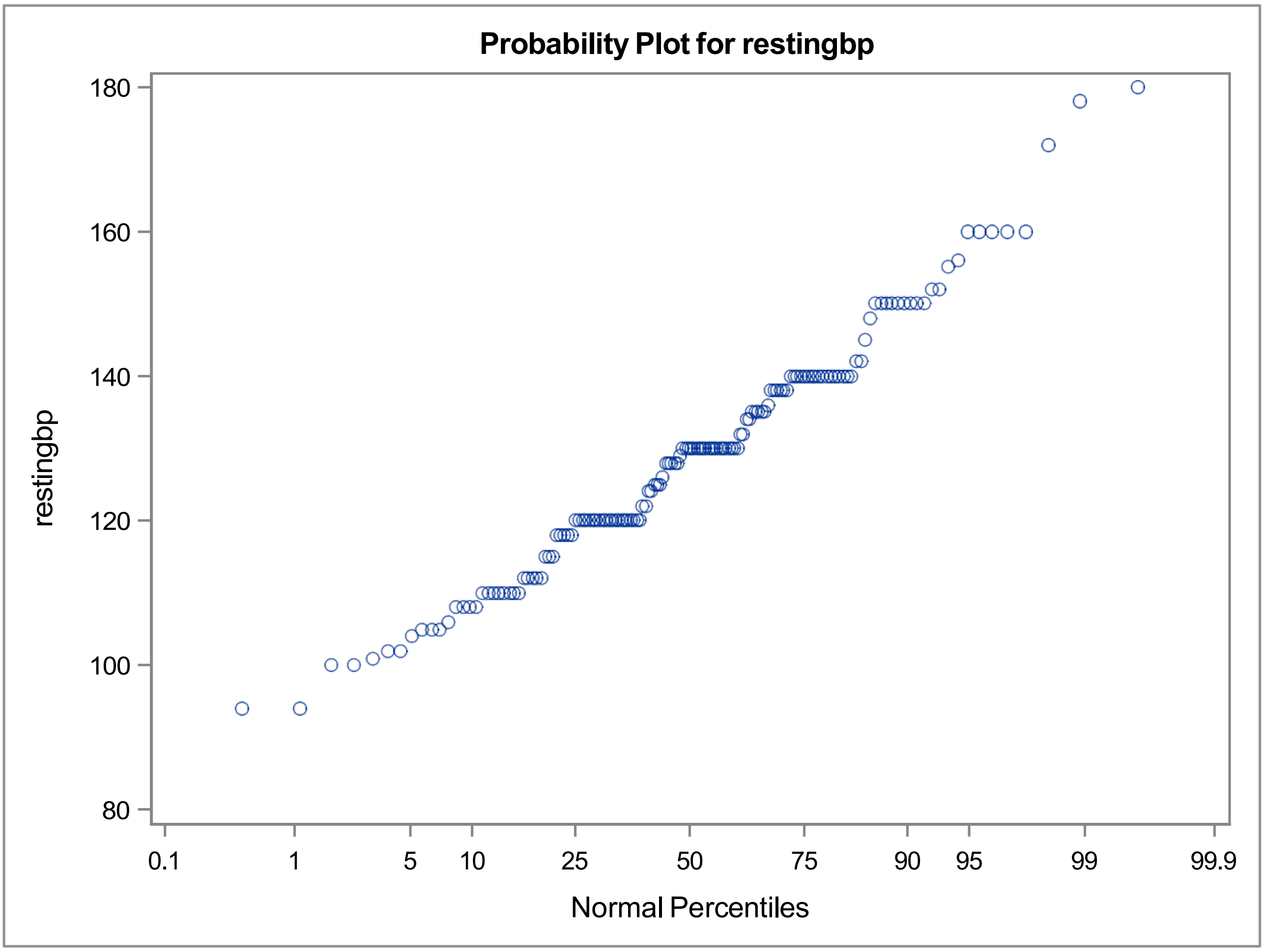
b) In the absence of heart disease, we see that resting blood pressure is also not normal. The histogram does not appear bell-shaped, while the probability plot appears to be straight. The normality tests also all fail to reject an assumption of normality. All the tests imply that the distribution is not normal, especially if we take Kolmogorov-Smirnov for out sample (n=270).

Variable: restingbp

heartdisease=absence

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.978975	**Pr < W**	0.0213
Kolmogorov-Smirnov	**D**	0.091639	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.152709	**Pr > W-Sq**	0.0224
Anderson-Darling	**A-Sq**	0.899307	**Pr > A-Sq**	0.0221

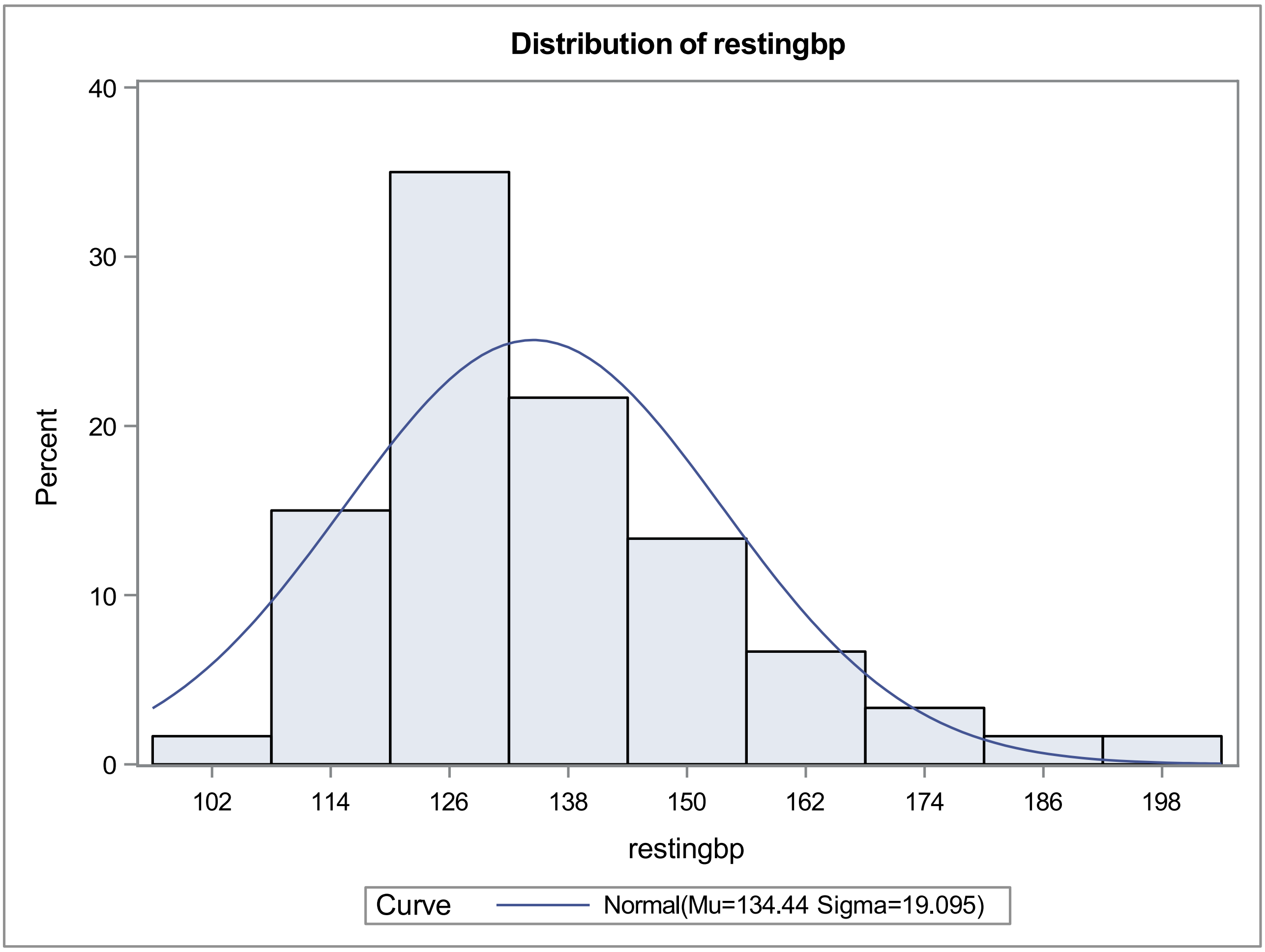

Distribution of restingbp
Percent
30
25
20
15
10
5
0
96
108
120
132
144
156
168
180
restingbp
Curve
Normal(Mu=128.87 Sigma=16.458)


Probability Plot for restingbp
restingbp
180
160
140
120
100
80
0.1
1
5
10
25
50
75
90
95
99
99.9
Normal Percentiles

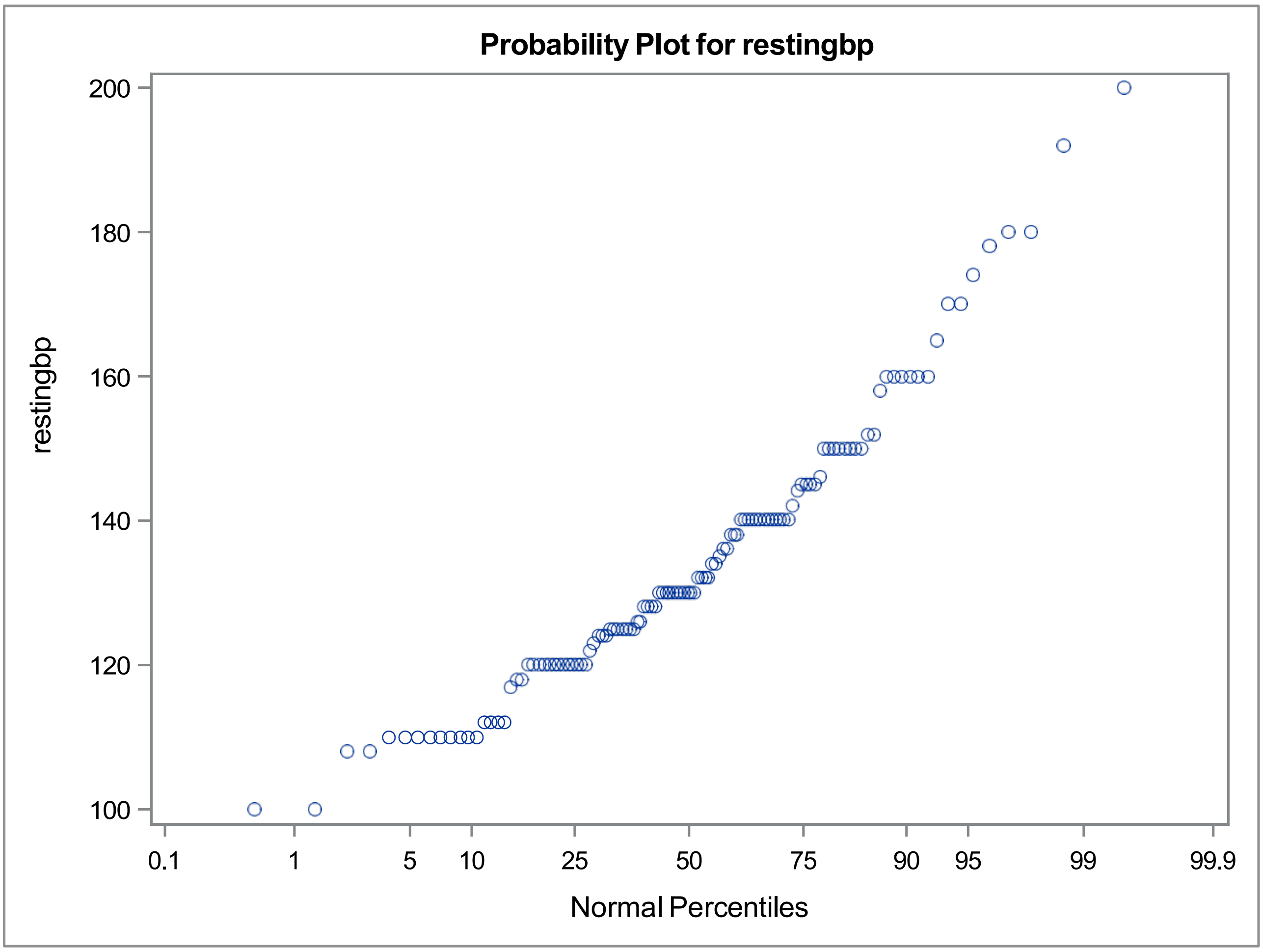
Variable: restingbp

heartdisease=presence

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.948046	**Pr < W**	0.0002
Kolmogorov-Smirnov	**D**	0.110494	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.253846	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	1.610518	**Pr > A-Sq**	<0.0050

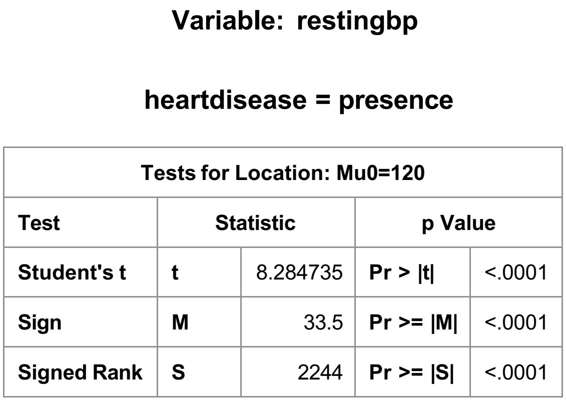


In the presence of heart disease, we see that resting blood pressure is also not normal. It is also can be observed from the histogram plot and probability plot. All the tests imply that the distribution is not normal as well. So, normality should not be assumed in this case.
In conclusion, we can say that group with heartdisease=presence is considered to be significantly far from the normality, based on the tests.


Probability Plot for restingbp
restingbp
200
180
160
140
120
100
0.1
1
5
10
25
50
75
90
95
99
99.9
Normal Percentiles

Exercise 3:

a) Based on the analysis from Exercise 2 we do not have normality assumption (cannot use t-test), so we can perform all the tests to check whether we can accept average resting blood pressure of 120 (129). So, we can consider sign test (since we have skewness and non-symmetricity). The null hypothesis is median of resting blood pressure = 120; alternative hypothesis is significantly higher than 120. The test suggest that we fail to accept null hypothesis for group with heartdisease = presence.

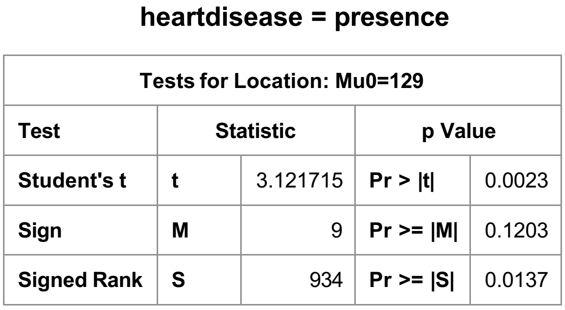


Variable: restingbp

heartdisease = presence

Tests for Location: Mu0=120				
Test	**Statistic**		**p Value**	
Student's t	**t**	8.284735	**Pr > |t|**	<.0001
Sign	**M**	33.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	2244	**Pr >= |S|**	<.0001

Similar to the first case, we can consider sign test (since we have skewness and non-symmetricity). The null hypothesis is median of resting blood pressure = 129; alternative hypothesis is significantly higher than 129. The test suggest that accept null hypothesis for group with heartdisease=presence. p-value = 0.1203 > significant value of 5%.



heartdisease = presence

Tests for Location: Mu0=129				
Test	**Statistic**		**p Value**	
Student's t	**t**	3.121715	**Pr > |t|**	0.0023
Sign	**M**	9	**Pr >= |M|**	0.1203
Signed Rank	**S**	934	**Pr >= |S|**	0.0137

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable restingbp Classified by Variable heartdisease

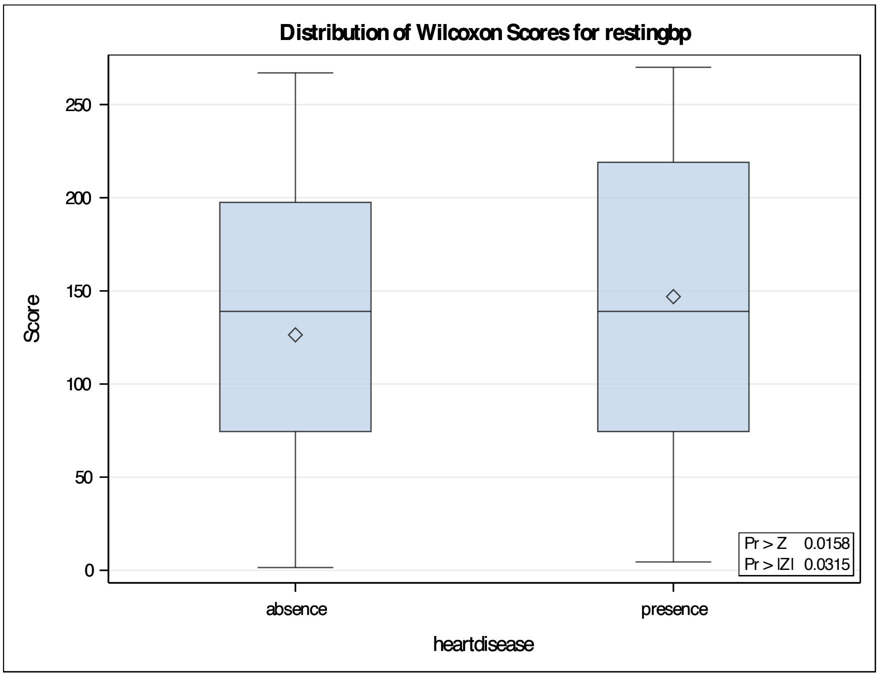
heartdisease	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
absence	150	18957.50	20325.0	635.759554	126.383333
presence	120	17627.50	16260.0	635.759554	146.895833

Average scores were used for ties.

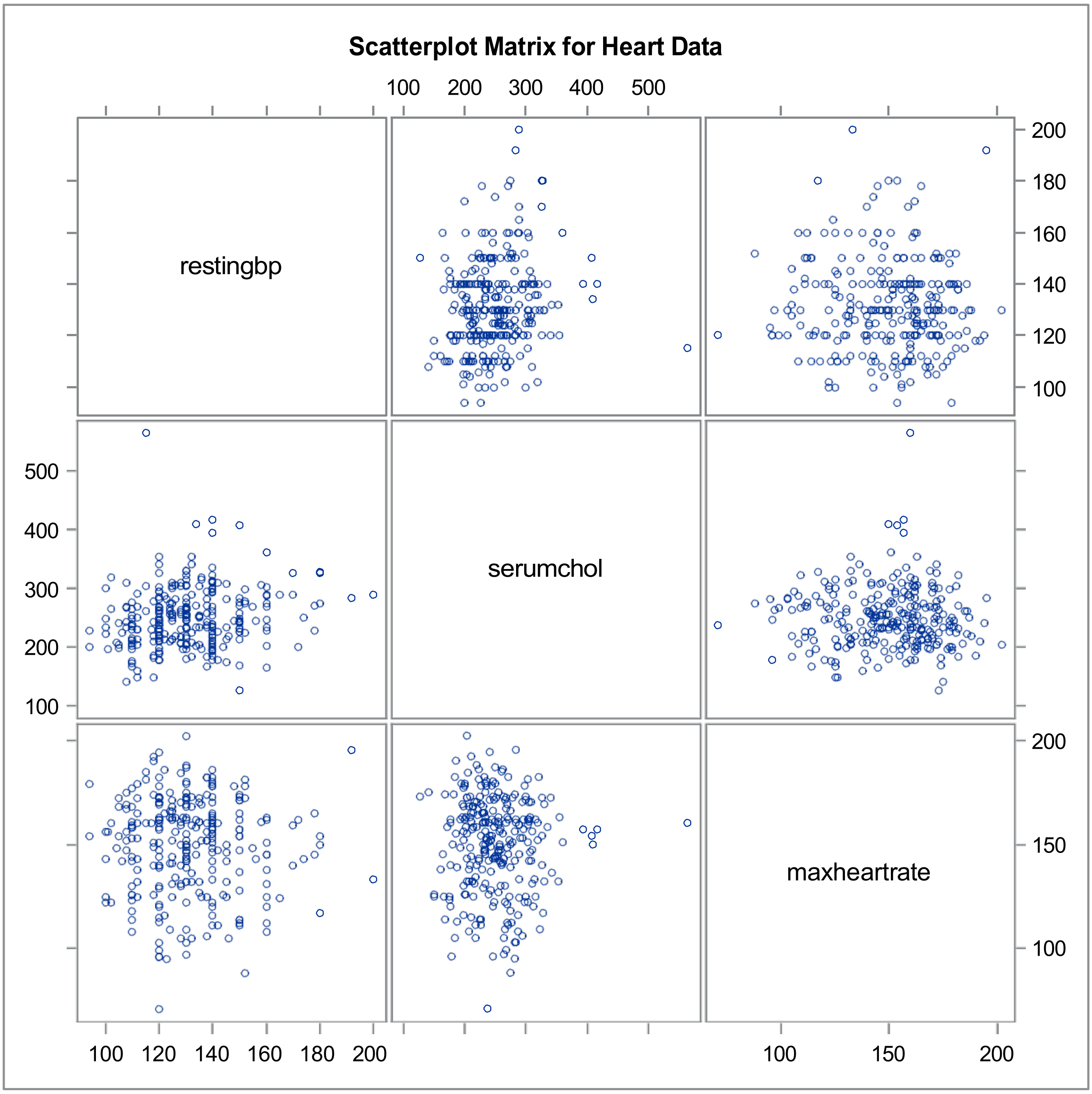
Wilcoxon Two-Sample Test

				t Approximation	
Statistic	**Z**	**Pr > Z**	**Pr > |Z|**	**Pr > Z**	**Pr > |Z|**
17627.50	2.1502	0.0158	0.0315	0.0162	0.0324

Z includes a continuity correction of 0.5.



b) As our resting blood pressure values were far from normal, we use rank sum test to compare restingbp of group with heart disease with group without heart disease. We almost have the same of number absences and presences. The null hypothesis is: “those with heart disease have higher resting blood pressure than those without heart disease”. The one-sided test is significant, so the conclusion is that group having heart disease had lower restingbp than the group without heart disease. Therefore, we fail to accept null hypothesis.



Exercise 4:

a) We first created pairwise scatter plot for **restingbp, serumchol and maxheartrate** variables. As can be seen from the plot above, there is not much linear trends in the data, and there are some extreme values. So, I think it would be better to consider Spearman correlation for the data.

   Below is the output of Spearman correlation test. Spearman correlation should tell us about the general tendency if the variable will go up/down as the other variable increases. Null hypothesis will be as the ranks of one variable increase, the ranks of the other variable do not increase (or decrease) (don't covary).

   - (Ranks) restingbp and serumchol has some positive relationship (slightly might increase if other increases). However, it is significant since we have p-value=0.0017<alpha=0.05. We fail to accept null hypothesis.
   - (Ranks) restingbp and maxheartrate are negatively correlated (almost not, small value) and p-value=0.4832>alpha, we accept null hypothesis.
   - (Ranks) serumchol and maxheartrate also are negatively correlated (almost not, small value) and p-value = 0.3553 >alpha, we accept null hypothesis.

The CORR Procedure

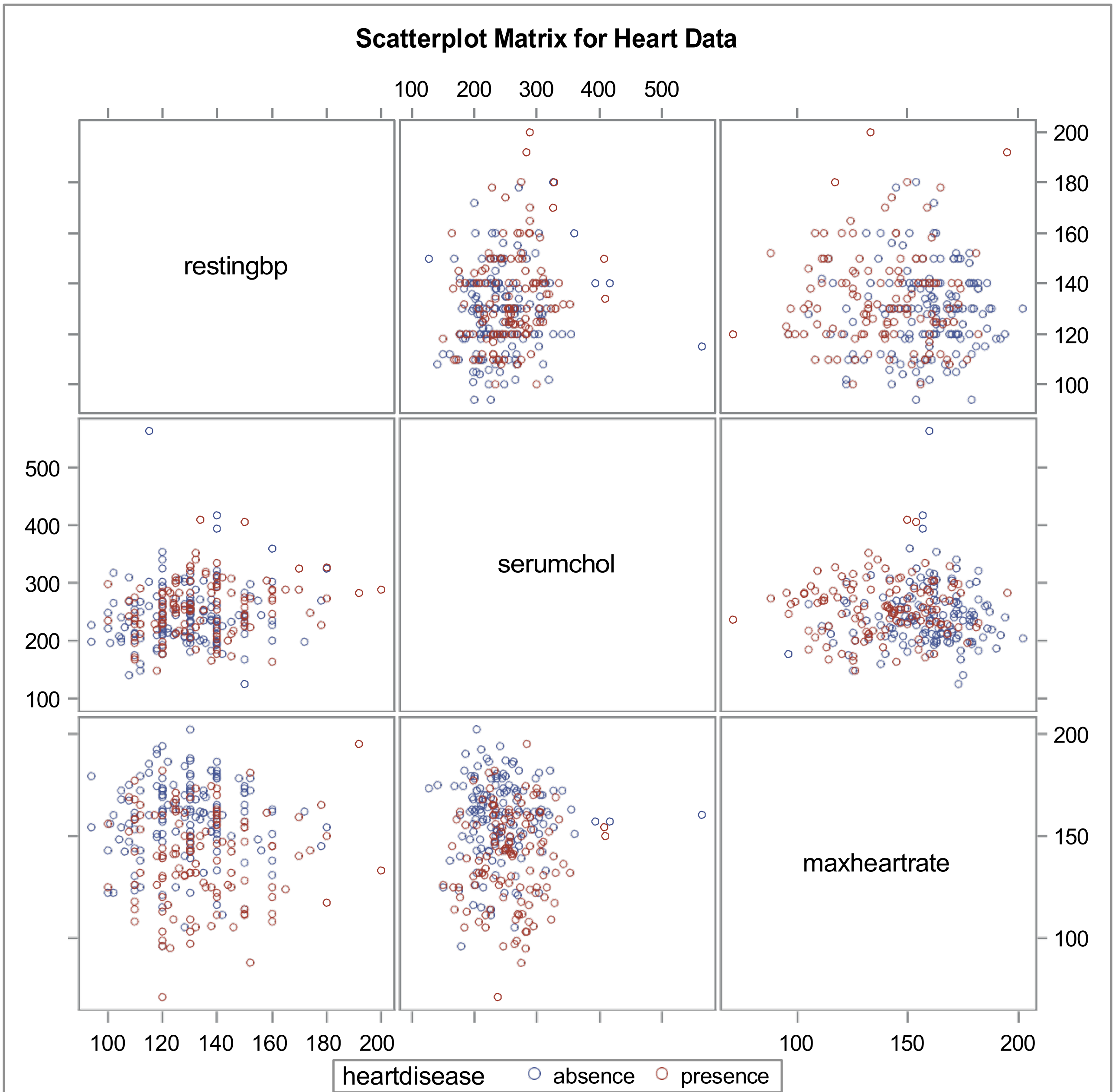
3 Variables:	restingbp serumchol maxheartrate

Simple Statistics						
Variable	**N**	**Mean**	**Std Dev**	**Median**	**Minimum**	**Maximum**
restingbp	270	131.34444	17.86161	130.00000	94.00000	200.00000
serumchol	270	249.65926	51.68624	245.00000	126.00000	564.00000
maxheartrate	270	149.67778	23.16572	153.50000	71.00000	202.00000

Spearman Correlation Coefficients, N = 270
Prob > |r| under H0: Rho=0

	restingbp	serumchol	maxheartrate
restingbp	1.00000	0.19048 0.0017	-0.04294 0.4823
serumchol	0.19048 0.0017	1.00000	-0.05647 0.3553
maxheartrate	-0.04294 0.4823	-0.05647 0.3553	1.00000

b) We repeated the same analysis as in part a): pairwise scatter plot and Spearman correlation test based on different groups.



heartdisease=absence

3 Variables:	restingbp serumchol maxheartrate

Simple Statistics						
Variable	**N**	**Mean**	**Std Dev**	**Median**	**Minimum**	**Maximum**
restingbp	150	128.86667	16.45766	130.00000	94.00000	180.00000
serumchol	150	244.21333	54.01909	236.00000	126.00000	564.00000
maxheartrate	150	158.33333	19.28336	161.00000	96.00000	202.00000

Spearman Correlation Coefficients, N = 150 Prob > |r| under H0: Rho=0			
	restingbp	**serumchol**	**maxheartrate**
restingbp	1.00000	0.11422 0.1640	0.04808 0.5590
serumchol	0.11422 0.1640	1.00000	-0.01894 0.8181
maxheartrate	0.04808 0.5590	-0.01894 0.8181	1.00000

heartdisease=presence

3 Variables:	restingbp serumchol maxheartrate

Simple Statistics						
Variable	**N**	**Mean**	**Std Dev**	**Median**	**Minimum**	**Maximum**
restingbp	120	134.44167	19.09542	130.00000	100.00000	200.00000
serumchol	120	256.46667	47.96917	255.50000	149.00000	409.00000
maxheartrate	120	138.85833	23.13072	141.50000	71.00000	195.00000

Spearman Correlation Coefficients, N = 120 Prob > |r| under H0: Rho=0			
	restingbp	**serumchol**	**maxheartrate**
restingbp	1.00000	0.25154 0.0056	-0.02543 0.7828
serumchol	0.25154 0.0056	1.00000	0.02042 0.8248
maxheartrate	-0.02543 0.7828	0.02042 0.8248	1.00000

In the absence of heart disease, we see that p-values are larger than alpha, hence we accept the null hypothesis that the ranks of variables do not convary.

In the presence of heart disease:

- (Ranks) Restingbp and serumchol are slightly positive correlated, and one variable might slighly increase if other increase. So, we fail to accept null hypothesis.
- (Ranks) restingbp and maxheartrate do not covary, so we accept null hypothesis, wih p-value = 0.7828.
- (Ranks) serumchol and maxheartrate do not covary, so we also accept null hypothesis, with p-value = 0.8248.

In conclusion, we see that in the absence of heart disease, all variables did not have any correlation, and in the presence of heart disease, we had slight correlation between cholesterol and resting blood pressure. The latter's output was similar to the output from the first part of the problem, where we considered the data without grouping.