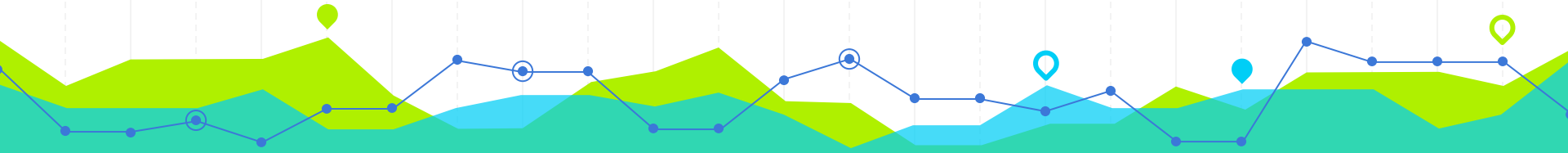
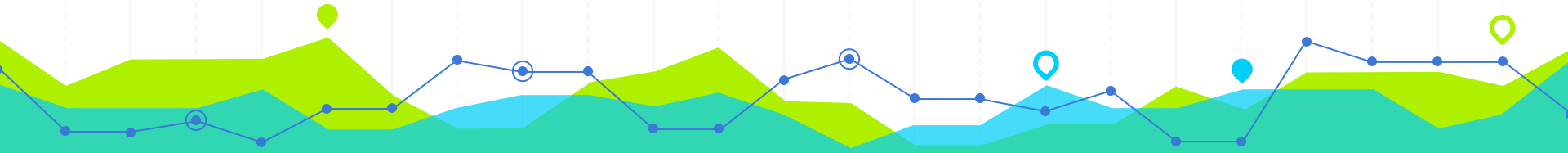


Testing the Normality of Data

Data Analysis Group A

Contents

- Dataset description
- Questions
- Methods for testing normality
  - Histograms & Box Plots
  - Q-Q and P-P Plots
  - Skewness & Kurtosis
  - Shapiro-Wilk, Kolmogorov-Smirnov, & Anderson-Darling
- Results/Question answers
  - Question 1
  - Question 2



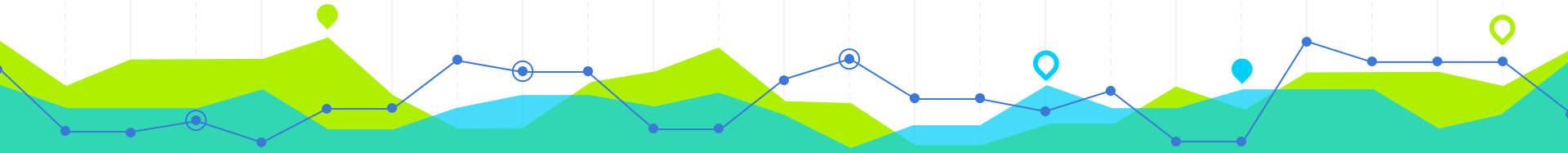
Data Set Description

1

Heart Disease Prediction

Heart Disease Prediction Dataset

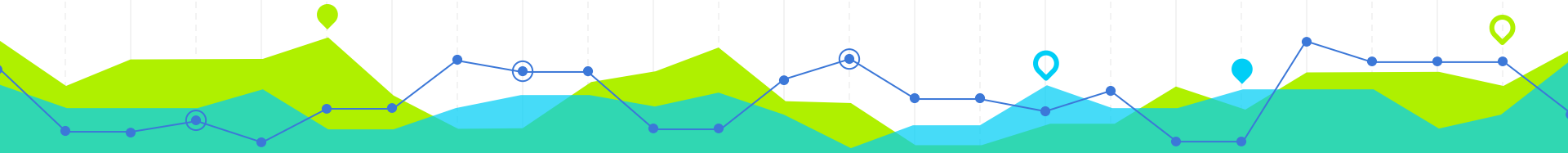
- 14 variables including Age, Sex, BP, etc…
- 270 rows of data
- The variables consist of a mix of numerical and categorical variables
- The variables that are relevant for this analysis are the continuous numeric variables: Age, BP, Cholesterol, Max HR, ST Depression



Questions 2

Questions

- Question 1: Which of the variables are approximately normally distributed?
- Question 2: Which of the variables are not normally distributed?
- Question 3: How might the normal distribution of these variables help us in possible future analysis of this data set?



Methods for Testing Normality

3

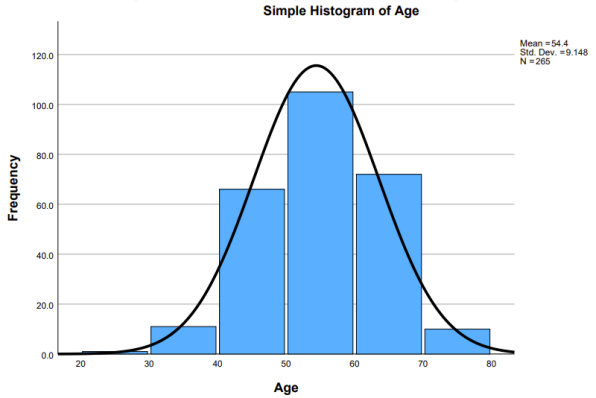
Histogram, Q-Q Plots, P-P Plots, Box Plots, Skewness & Kurtosis, Shapiro-Wilk & Kolmogorov Smirnov Tests

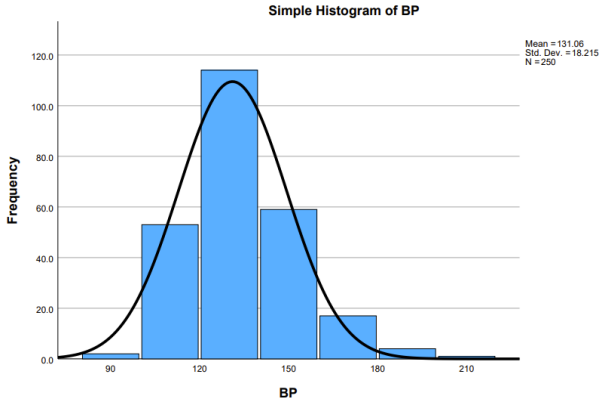
Methods of testing normality

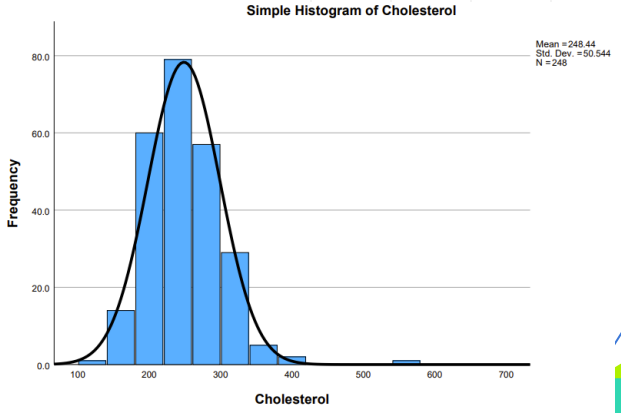
- Testing for normality is only for numerical data.
- An assessment of the normality of data is a prerequisite for many statistical tests because normal data is an underlying assumption in parametric testing.
- There are two main methods of assessing normality: **graphically and numerically.**

Histograms

- Age, BP, Cholesterol, and Max HR look approximately normal
- They are symmetric
- They have one mode
- The number of events less than the mean is approximately equal to the number of events above the mean

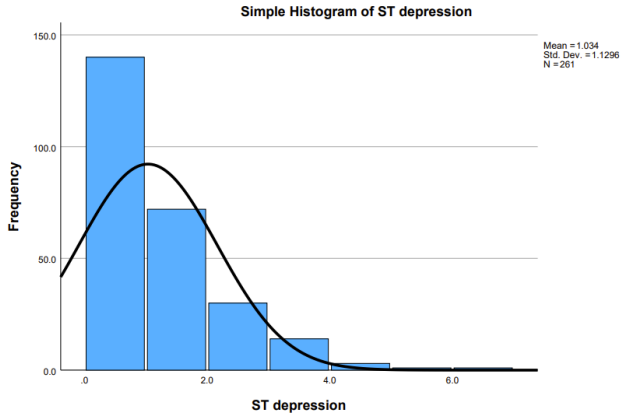


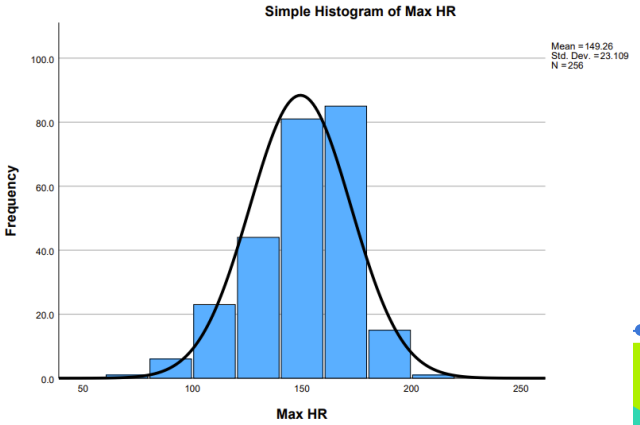




Histograms

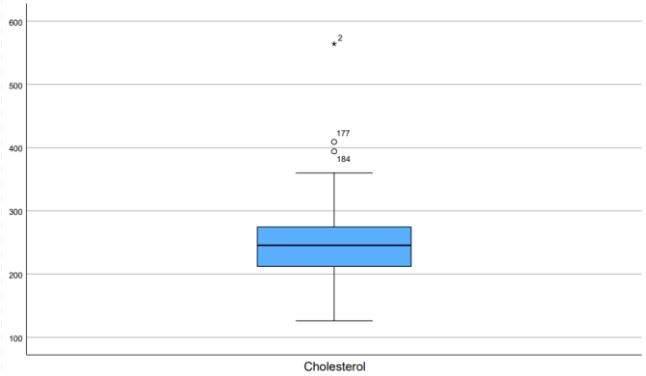
- ST depression does not look like it is normally distributed
- It is not symmetric
- The number of events less than the average is not equal to the number of events above the average

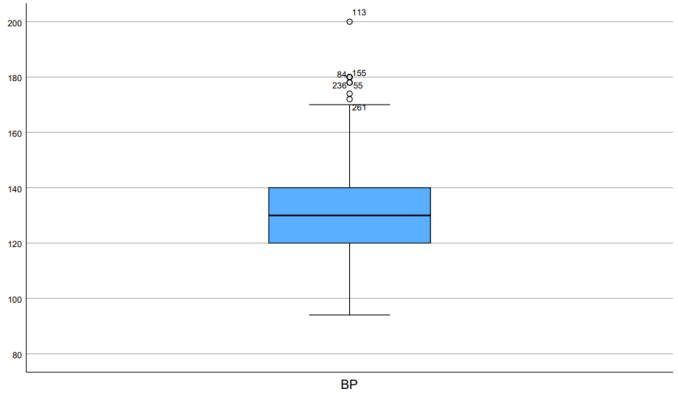


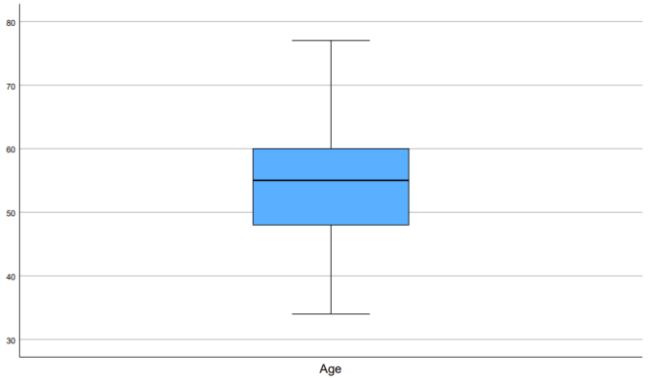


Box Plots

- Age, Cholesterol, andBP appear approximately normal
- Median is approximately at the center of the box
- Upper and lower quartile seem approximately equal
- Upper and lower whiskers seem approximately equal

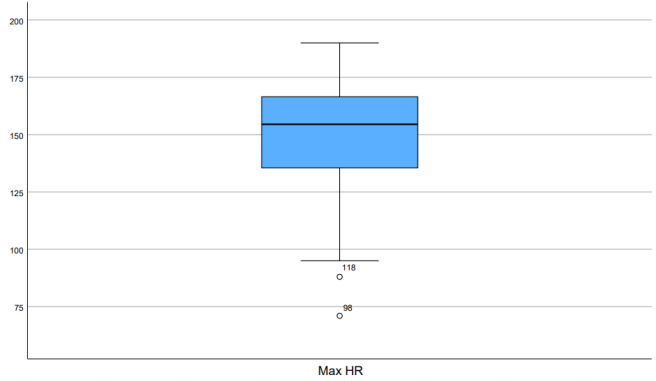


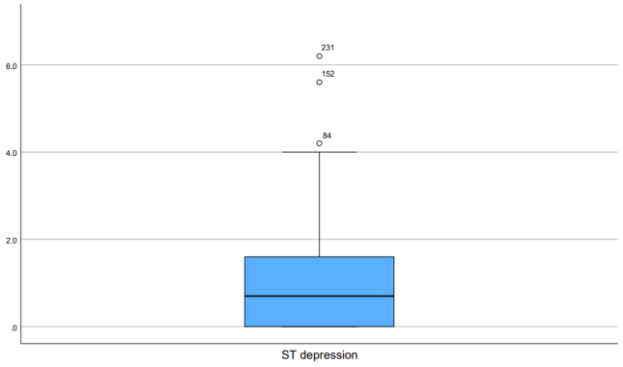




Box Plots

- Max HR and ST Depression seem to deviate from a normal distribution
- Median is not centered
- Upper and lower quartiles and whiskers are not equal



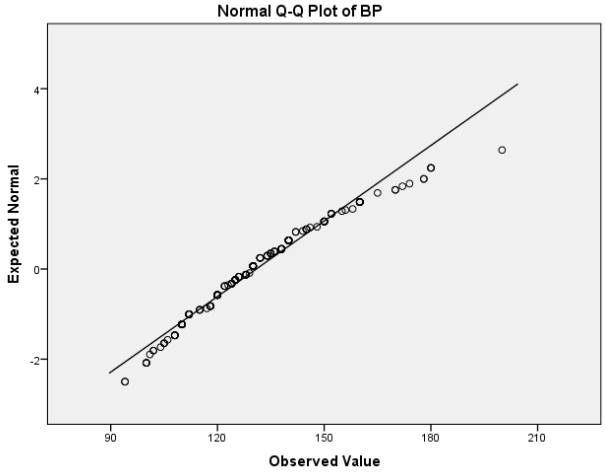


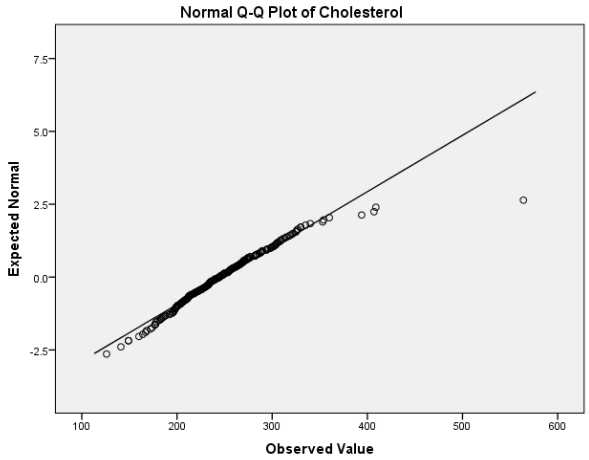
Q-Q and P-P Plots

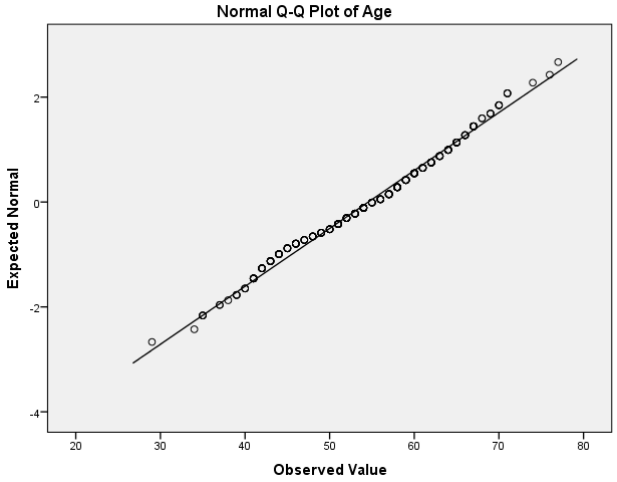
- Q-Q Plot: useful for checking whether a dataset follows a certain theoretical distribution, such as a normal distribution or a log-normal distribution. If the points on the Q-Q plot fall on a straight line, it indicates that the two datasets have the same distribution.

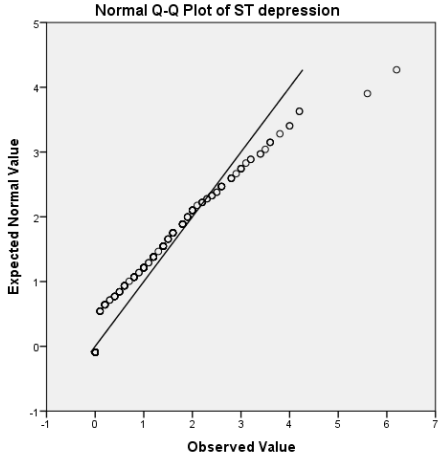
QQ plots

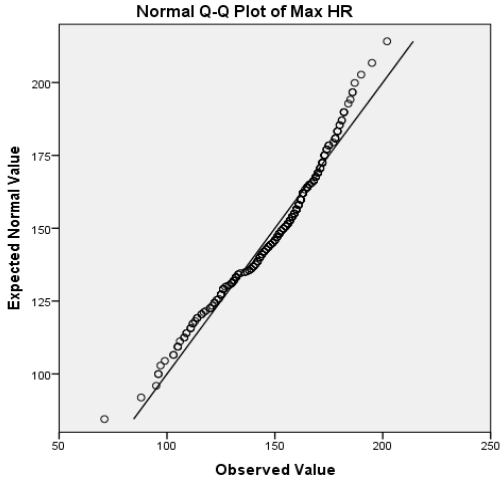
All variables show normality aside from ST depression.
BP and Cholesterol shows slight outliers, but overall show normal distribution







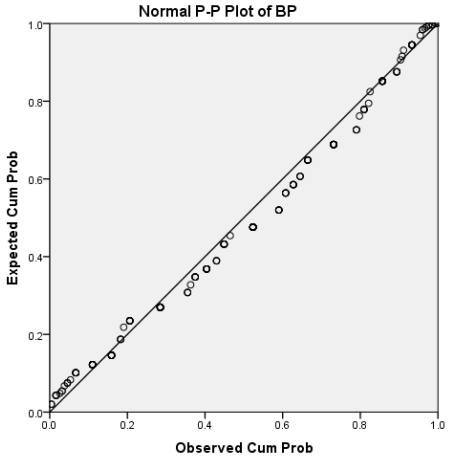


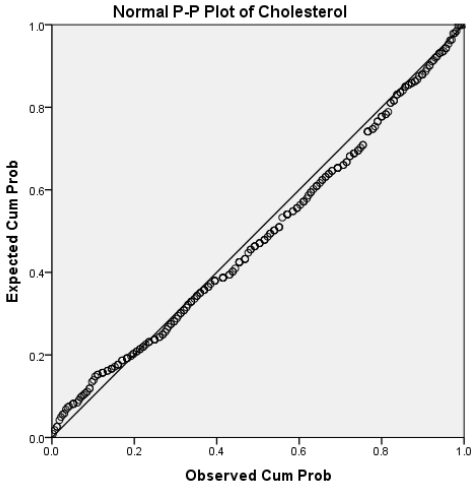


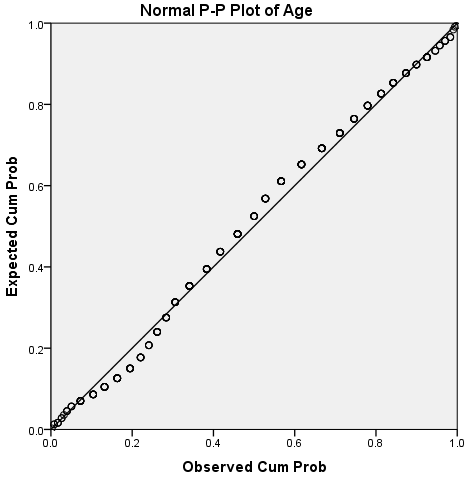
PP Plots

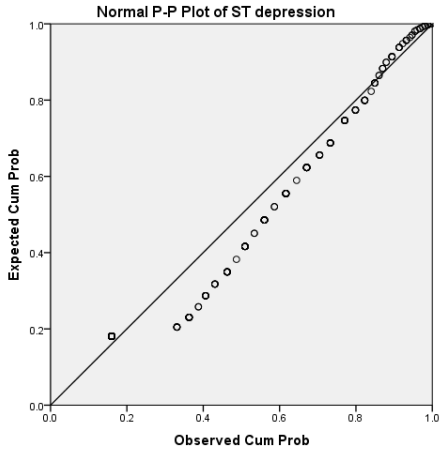
All tests aside from ST depression here shows normality as well

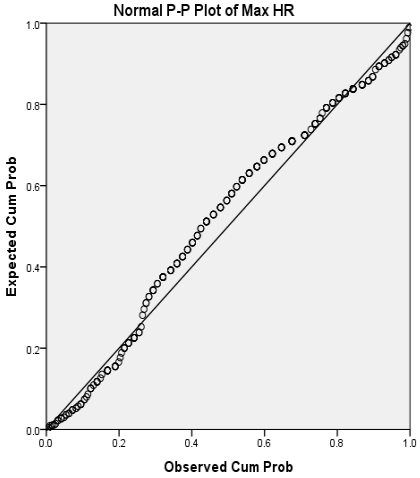
Max HR shows slight deviation from the line but sticks close enough to be considered normal, and BP is a bit messy but sticks to the line.











Skewness and Kurtosis Z-scores

As the sample is medium-size we establish the normality of the data if the Z-score is between -3.29 and +3.29.

Variable Age

- Skewness: 0.114/0.153=0.745
- Kurtosis : 0.580/0.304=1,98

Descriptives

			Statistic	Std. Error
Age	Mean		54,20	,578
	95% Confidence Interval for Mean	Lower Bound	53,06	
		Upper Bound	55,34	
	5% Trimmed Mean		54,25	
	Median		54,00	
	Variance		85,121	
	Std. Deviation		9,226	
	Minimum		29	
	Maximum		77	
	Range		48	
	Interquartile Range		14	
	Skewness		-,114	,153
	Kurtosis		-,580	,304

Skewness and Kurtosis Z-scores

Variable BP

- Skewness: 0.610/0.153=3.987
- Kurtosis : 0.811/0.304=2.668

We cannot consider the BP variable to be normally distributed due to the skewness z-score is greater than 3,29

Descriptives

			Statistic	Std. Error
SMEAN(BP)	Mean		129,896	1,0273
	95% Confidence Interval for Mean	Lower Bound	127,873	
		Upper Bound	131,919	
	5% Trimmed Mean		129,328	
	Median		129,896	
	Variance		269,088	
	Std. Deviation		16,4039	
	Minimum		94,0	
	Maximum		192,0	
	Range		98,0	
	Interquartile Range		20,0	
	Skewness		,610	,153
	Kurtosis		,811	,304

Skewness and Kurtosis Z-scores

Variable Cholesterol

- Skewness: 0.152/0.153=0.993
- Kurtosis : 0.172/0.304=0,566

The result of the skewness and kurtosis z-score of the cholesterol variable indicate that the data may be normally distributed

Descriptives

			Statistic	Std. Error
SMEAN(Cholesterol)	Mean		246,349	2,6950
	95% Confidence Interval for Mean	Lower Bound	241,041	
		Upper Bound	251,656	
	5% Trimmed Mean		245,955	
	Median		246,349	
	Variance		1852,055	
	Std. Deviation		43,0355	
	Minimum		126,0	
	Maximum		360,0	
	Range		234,0	
	Interquartile Range		60,0	
	Skewness		,152	,153
	Kurtosis		-,172	,304

Skewness and Kurtosis Z-scores

Variable Max HR

- Skewness: -0.577/0.153=-3.77
- Kurtosis : -0.046/0.304= -0.151

We cannot consider the MaxHR variable to be norm distributed due to the skewness z-score is greater than 3,29

Descriptives

			Statistic	Std. Error
SMEAN(MaxHR)	Mean		149,346	1,4372
	95% Confidence Interval for Mean	Lower Bound	146,515	
		Upper Bound	152,176	
	5% Trimmed Mean		150,161	
	Median		152,000	
	Variance		526,699	
	Std. Deviation		22,9499	
	Minimum		71,0	
	Maximum		202,0	
	Range		131,0	
	Interquartile Range		34,0	
	Skewness		-,577	,153
	Kurtosis		-,046	,304

Skewness and Kurtosis Z-scores

Z-scores

- Skewness: 0.902/0.153=5.89
- Kurtosis : -0.007/0.304=-0.023

Descriptives

			Statistic	Std. Error
SMEAN(STdepression)	Mean		,9713	,06359
	95% Confidence Interval for Mean	Lower Bound	,8461	
		Upper Bound	1,0966	
	5% Trimmed Mean		,8882	
	Median		,8000	
	Variance		1,031	
	Std. Deviation		1,01550	
	Minimum		,00	
	Maximum		4,20	
	Range		4,20	
	Interquartile Range		1,60	
	Skewness		,902	,153
	Kurtosis		-,007	,304

Shapiro-Wilk and Kolmogorov-Smirnov Tests

- These are statistical tests that examine the null hypothesis that the data came from a normally distributed population.
- **Null hypothesis:** The values are sampled from a population that is normally distributed.
- **Alternative hypothesis:** The values are sampled from a population that is not normally distributed.

Shapiro-Wilk and Kolmogorov-Smirnov Tests

- If P<0.05===> reject null hypothesis.(not normal)
- If P>0.05===> do not reject null hypothesis. (normal)

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Age	.066	270	.006	.988	270	.028
BP	.104	255	.000	.963	255	.000
Cholesterol	.052	253	.095	.943	253	.000
Max HR	.085	259	.000	.970	259	.000
ST depression	.181	266	.000	.850	266	.000

a. Lilliefors Significance Correction

Shapiro-Wilk and Kolmogorov-Smirnov Tests

- If we want to conduct a parametric test, normality testing should be done for each category of the independent variable separately.
- E.g: For the independent variable ‘Sex’ = 0 or 1 (Two categories)
- Since the sample size n>50, KS test will be more reliable.

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Sex	Statistic	df	Sig.	Statistic	df	Sig.
Age	0	.102	86	.028	.976	86	.117
	1	.070	184	.030	.989	184	.177
BP	0	.116	81	.009	.963	81	.018
	1	.122	174	.000	.963	174	.000
Cholesterol	0	.103	82	.032	.923	82	.000
	1	.045	171	.200*	.994	171	.738
Max HR	0	.157	82	.000	.909	82	.000
	1	.060	177	.200*	.982	177	.022
ST depression	0	.214	85	.000	.775	85	.000
	1	.166	181	.000	.874	181	.000

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Interpretation

- Based on the results of the KS and SW tests, the variable ‘Cholesterol’ is the only one that seems to be sampled from a population that follows a normal distribution.
- When the data is split based on the independent variable ‘Sex’, both Cholesterol and Max HR are normally distributed for the category ‘1’, as opposed to ‘0’.

Anderson Darling Test using R

- This test is not supported in SPSS, so it has been conducted using the following R code:

```
#install and load readxl package
install.packages('readxl')
library(readxl)

#import Excel file into R
data <- read_excel ("C:\\Users\\User\\Desktop\\Handson_data_analysis\\Group A. Heart_Disease_Prediction3.xlsx")

install.packages('nortest')
library(nortest)

ad.test(data$Age)
ad.test(data$BP)
ad.test(data$Cholesterol)
ad.test(data$`Max HR`)
ad.test(data$ST depression)
```

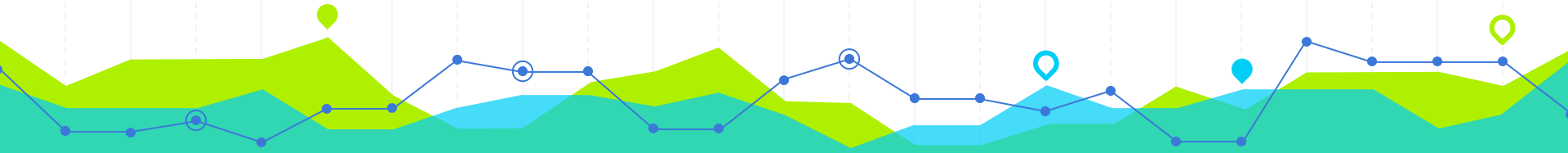
Anderson Darling Test using R

- The following is the output:

```
> ad.test(data$Age)

        Anderson-Darling normality test

data:  data$Age
A = 1.1234, p-value = 0.006011
> ad.test(data$BP)

        Anderson-Darling normality test

data:  data$BP
A = 2.3371, p-value = 6.23e-06
> ad.test(data$Cholesterol)

        Anderson-Darling normality test

data:  data$Cholesterol
A = 1.2954, p-value = 0.002259
```

```
> ad.test(data$`Max HR`)

        Anderson-Darling normality test

data:  data$`Max HR`
A = 2.6653, p-value = 9.817e-07
> ad.test(data$`ST depression`)

        Anderson-Darling normality test

data:  data$`ST depression`
A = 11.433, p-value < 2.2e-16
```

- The results of this test suggest that all of the variables are not sampled from a population that is normally distributed.



Results

Answers to Questions

4

Summary

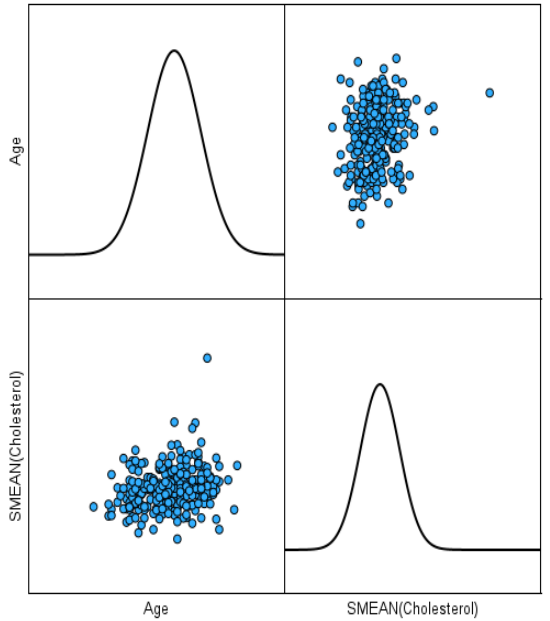
	Histograms	Box Plots	Q-Q Plots	P-P Plots	Skewness	Kurtosis	Shapio-Wilk	Kolmogorov-Smirnov	Anderson-Darling
Age	✓	✓	✓	✓	✓	✓	×	×	×
BP	✓	✓	✓	✓	×	×	×	×	×
Cholesterol	✓	✓	✓	✓	✓	✓	×	✓	×
Max HR	×	×	✓	✓	×	×	×	×	×
ST Depression	×	×	×	×	×	×	×	×	×

Conclusions

- Age and cholesterol are approximately normal

Question 1: Correlation between Age and Cholesterol?

- Data-set indicates, in-spite of how old people were, cholesterol levels were found to be not to be of a great deviation from one another
- Therefore age could not have been a deciding factor as to whether one would have high cholesterol, albeit the fact that the outlier (highest cholesterol level) was one case from old age.



Question 2: What are the chances of having a cholesterol level higher than 268?

Whit a normal distribution we can use Z-score to obtain the probability to happen of a score.

P(x=268)=P(z=0.5031)=70.19%

P(X>268)=1-0.7019=29,81%

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
SMEAN(Cholesterol)	255	126,0	360,0	246,349	43,0355	,152	,153	-,172	,304
Valid N (listwise)	255								

Thank you