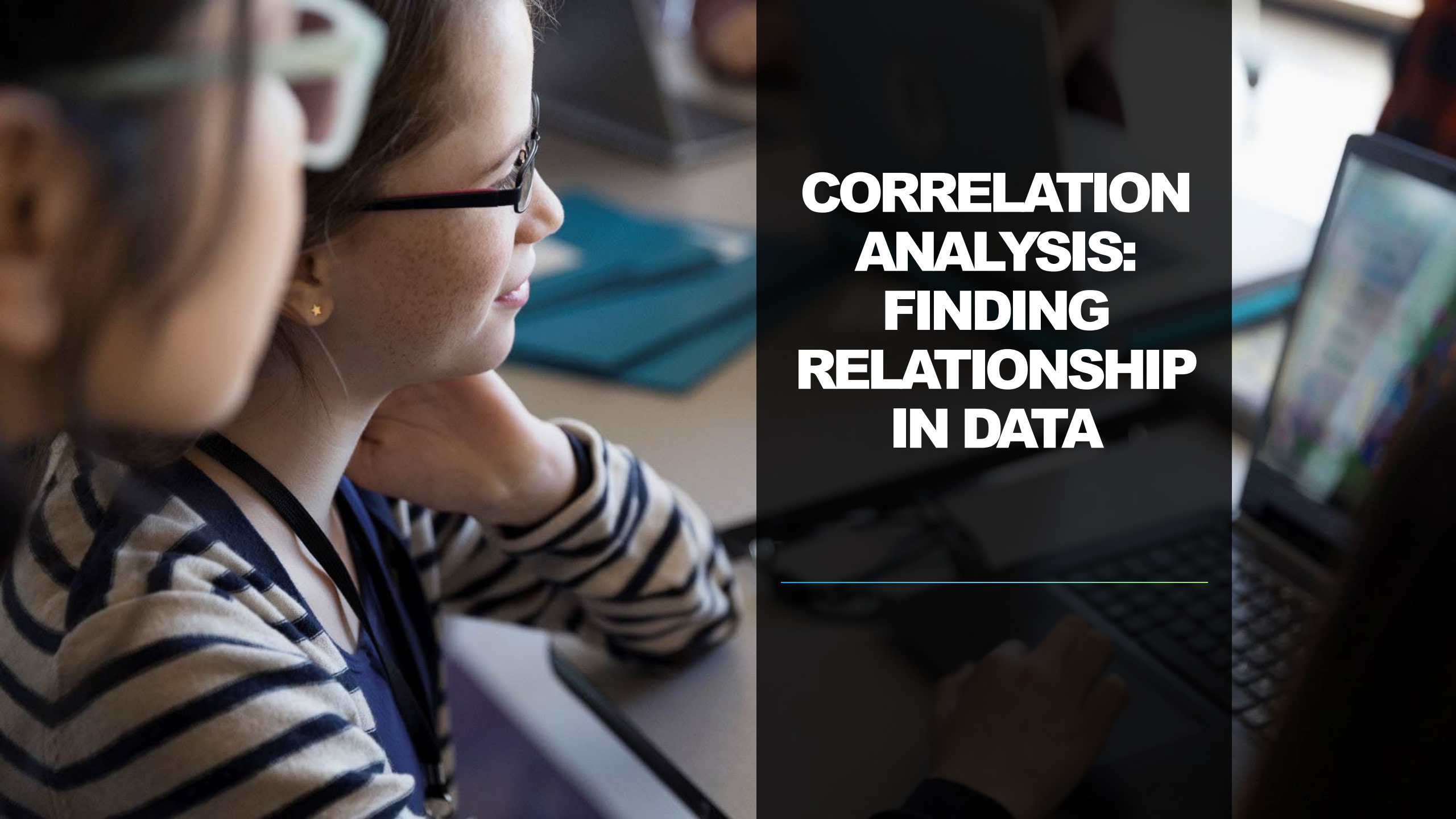


CORRELATION ANALYSIS: FINDING RELATIONSHIP IN DATA

CONTENT

Why Correlation?

Purpose to study correlation

Type of Relationship

Distinguish the differences of relationship

Scatter Plot

Visualization to see the relationship between variables

Assumption

Know the assumption for parametric and non-parametric way

Correlation Analysis

Correlation analysis for parametric and non-parametric data

WHY CORRELATION?

Correlation is a term in statistics that refers to the degree of association between two random variables.

Correlation analysis is used for spotting patterns within datasets. A positive correlation result means that both variables increase in relation to each other, while a negative correlation means that as one variable decreases, the other increases.

The variables are not designated as dependent or independent.

Correlation measure the strength of relationship. It is not for prediction.

Correlation and causation are not the same. Two variables may be correlated but it does not mean that one causes the other.

Causation One variable directly affects another.
It is implying that A and B have a cause-and-effect relationship with one another. You're saying A causes B. Causation is also known as causality.

Examples:

1. A study shows there exists a strong correlation between Math IQ and Feet size for Kindergarten children. Does this mean that Feet Size is the cause of Math IQ for kindergarten children?

2. A study finds a negative correlation between the number of hours people spend exercising and their body weight.
   Does this mean that exercising causes weight loss?

TYPE OF RELATIONSHIPS

Unrelated relationship:

Have no systematic relationship; changes in one variable simple are not related to the changes in the other variable.

Linear relationship

Represented and explained by a straight line on a scatter plot

Two type of linear relationship: Positive and negative relationship

No linear relationship

Takes at least one curve, or turn, to represent the data on a scatter plot.

Types of nonlinear relationship: curvilinear relationship (polynomial equation), U-shaped curvilinear relationship (two variables are related negatively until a certain point and then are related positively)

VISUALIZING VARIABILITY WITH SCATTER PLOT

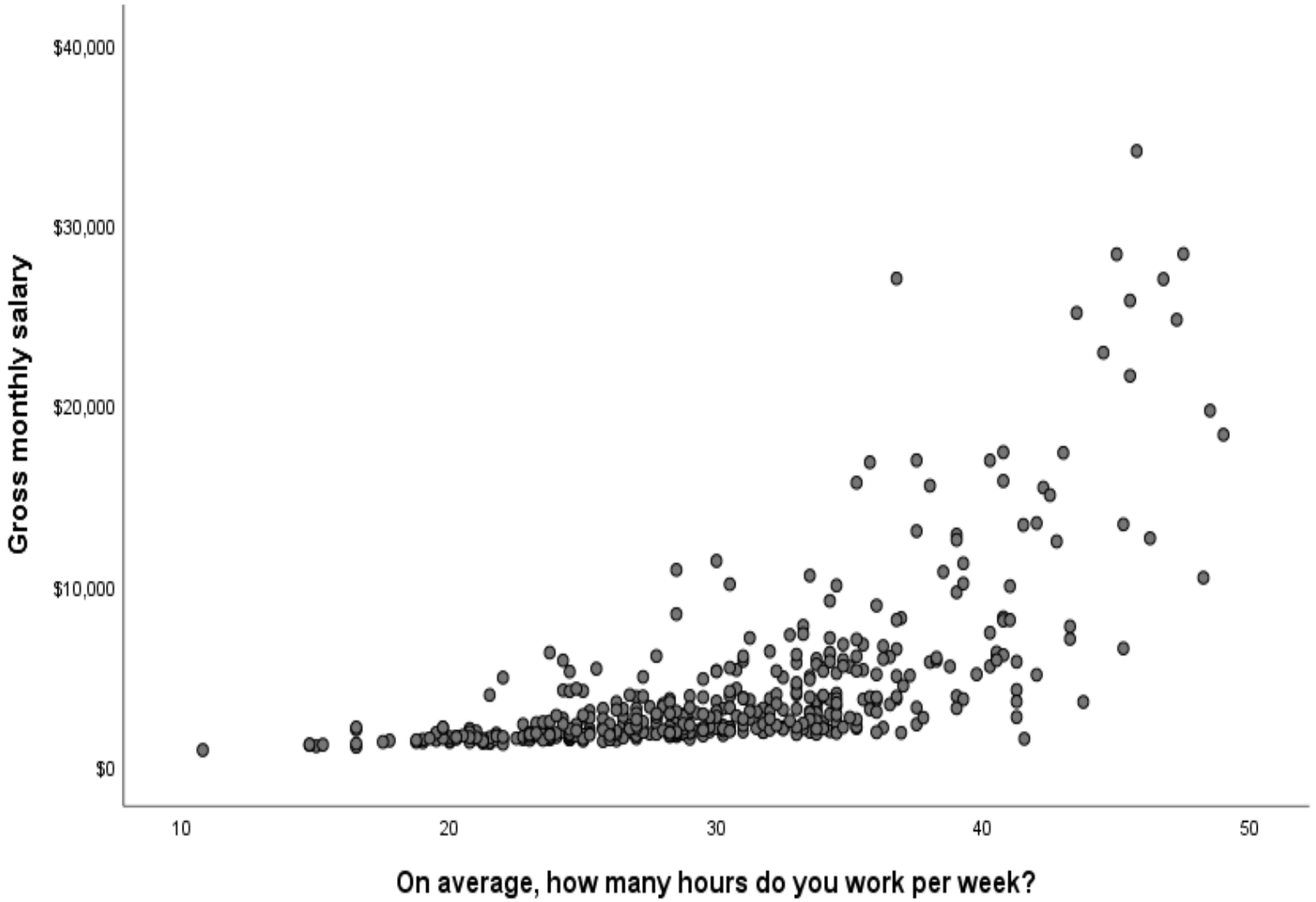
- Common visual tool for two-variable relationship: scatter plots (or scatter diagram)
- Scatter Plot/Scatter diagram: A two-dimensional plot showing the values for the joint occurrence of two quantitative variables measured on the same individual (observation→ row).
  - Each individual in the data set is represented by a point in the scatter diagram.

CAUTION !!!
Do not connect points when drawing scatter diagram

When examining Histogram (for single continuous variable), we look for center, spread and shape.
When examining Scatter plot (for two continuous variable) , we look for Trend (which is like center), strength (which is like spread) and shape.

WHAT YOU SEE FROM SCATTER PLOT?

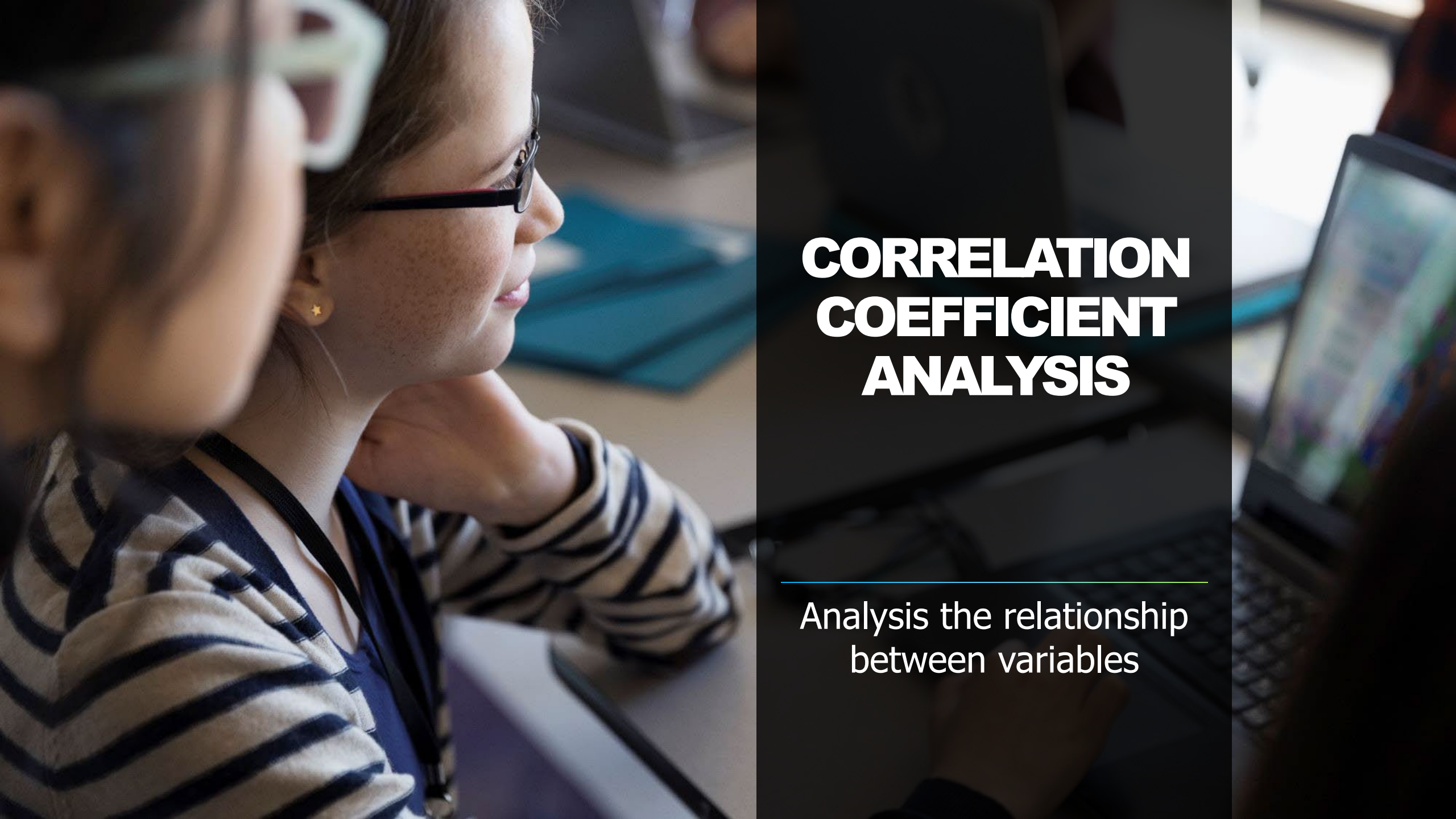
- Seeing the trend:
  - Increasing trend
  - Decreasing trend
  - Absence trend
- Seeing the strength of the trend:
  - Weak association/relation: have larger vertical variation
  - Strong association: have little vertical variation
- Seeing the shape of the trend:
  - Linear shape: Increase (or decrease) at the same rate
  - Non-Linear shape: it might fairly flat at first and then become steeper
- Seeing the Outliers/Extreme Values
  - Located far away



How can we summarize?

1. The dots become more dispersed as the respondents work more hours; the more hours people work, the greater the variability of monthly salary.

2. The pattern of dots "bend upwards" towards the right side of our chart. This is a clear indication of **nonlinearity**



CORRELATION COEFFICIENT ANALYSIS

Analysis the relationship between variables

CORRELATION COEFFICIENT

- What it is? - A quantitative measure of the strength of the relationship between variables.
- What can we measure?
  1. Strength of the relationship: weak, moderate, strong or no relationship
     - Is indicated through the value of correlation coefficient
     - This value is expressed in decimal value ranges from -1 to +1
     - The more correlation differs from 0.0, the stronger the linear relationship between the two variables.
     - A correlation coefficient of 0.00 means two variables are unrelated, at least in a linear manner.
  2. Direction of the relationship: positive or negative direction
     - Is indicated through the sign (+ or -) of correlation coefficient

Below are some of the measurements to calculate correlation coefficient according to variable type:

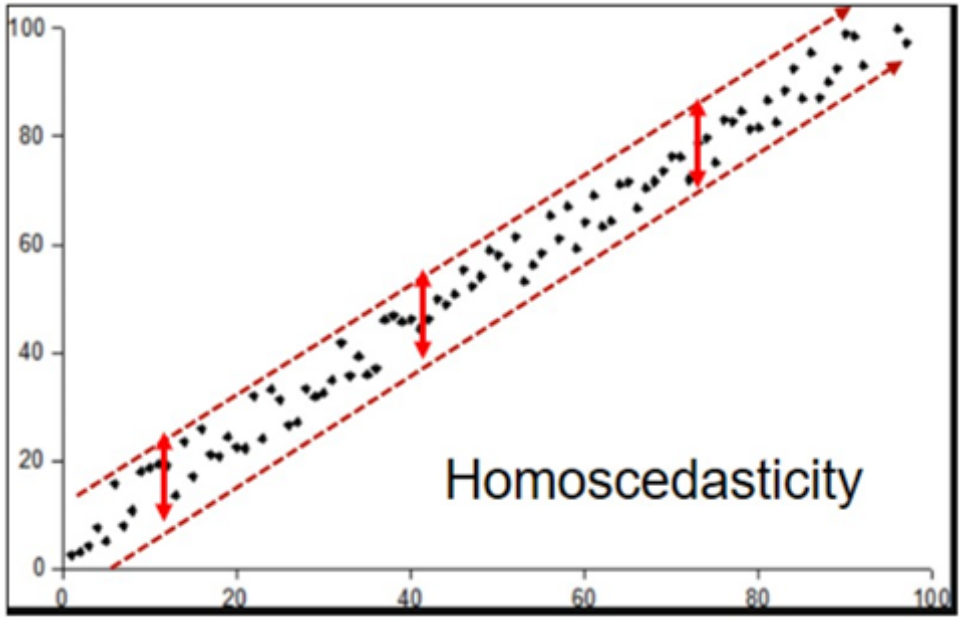
1. Relationship between ratio/interval variables:
   - Pearson Product Moment Correlation (r-value): measure of the strength and direction of association in correlation that met the assumption of parametric data.
2. Relationship between ordinal variables:
   - Spearmen rho correlation: nonparametric version of Pearson product used to compare two setts of ranked scores for the same group or the ranked scores of various item by two different groups
   - Kendall's correlation (tau):nonparametric measure of the strength and direction of association that exists between two variables measured on at least an ordinal scale.
3. Relationship between nominal variables:
   - Cramer's V: calculating correlation in tables which have more than 2x2 rows and columns. It is used as post-test to determine strengths of association after chi-square has determined significance.

SCENARIO 1: BOTH VARIABLES ARE QUANTITATIVE

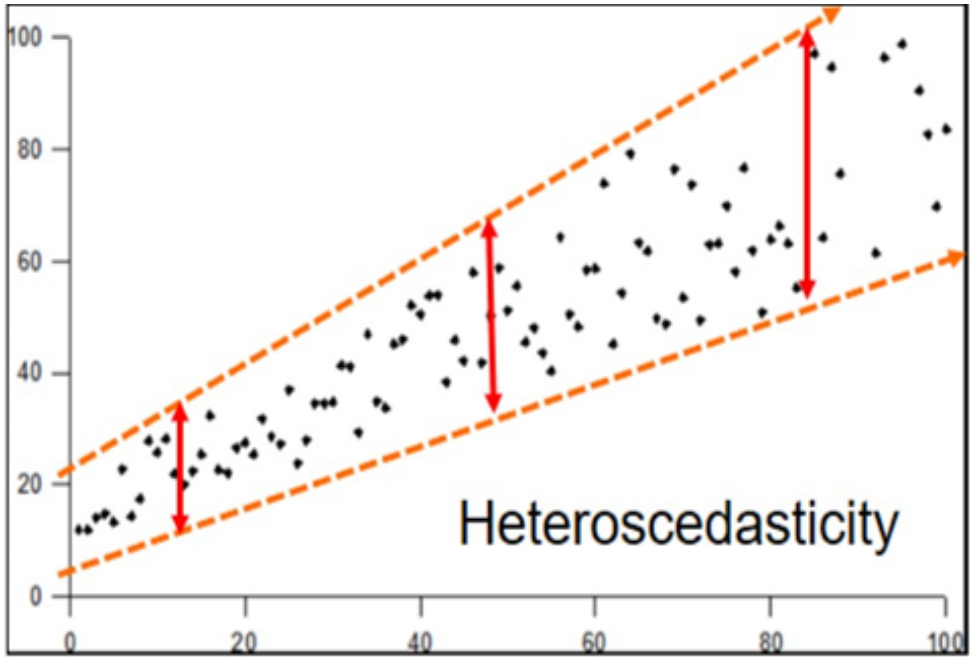
PEARSON CORRELATION

SPEARMAN CORRELATION

DETERMINING PARAMETRIC OR NON-PARAMETRIC CORRELATION ANALYSIS



VS



PARAMETRIC CORRELATION COEFFICIENT

Refers to the distance between the points to that straight line. "having the same scatter." The shape of the scatter plot should be tube-like in shape (ecliptically shape)

NON-PARAMETRIC CORRELATION COEFFICIENT

Refers to the circumstance in which the variability of a variable is unequal across the range of values of a second variable. Data with a different (*hetero*) dispersion (*skedasis*). Heteroscedastic data tends to follow a cone shape on a scatter plot.

SPSS FOR PARAMETRIC CORRELATION ANALYSIS

PEARSON MOMENT PRODUCT CORRELATION (R-VALUE)

IBM SPSS 26

PEARSON CORRELATION COEFFICIENT

- Parametric measure for correlation between variables
- Measure the strength and direction of association in correlation between two continuous variable
- Denoted by the symbol: r
- R value is unitless

INTERPRETING PEARSON CORRELATION

R VALUE	Strength of Relationship
< 0.2	Very Weak
0.2 – 0.3	Weak
0.4 – 0.6	Moderate
0.7 – 0.8	Strong
> 0.8	Very Strong

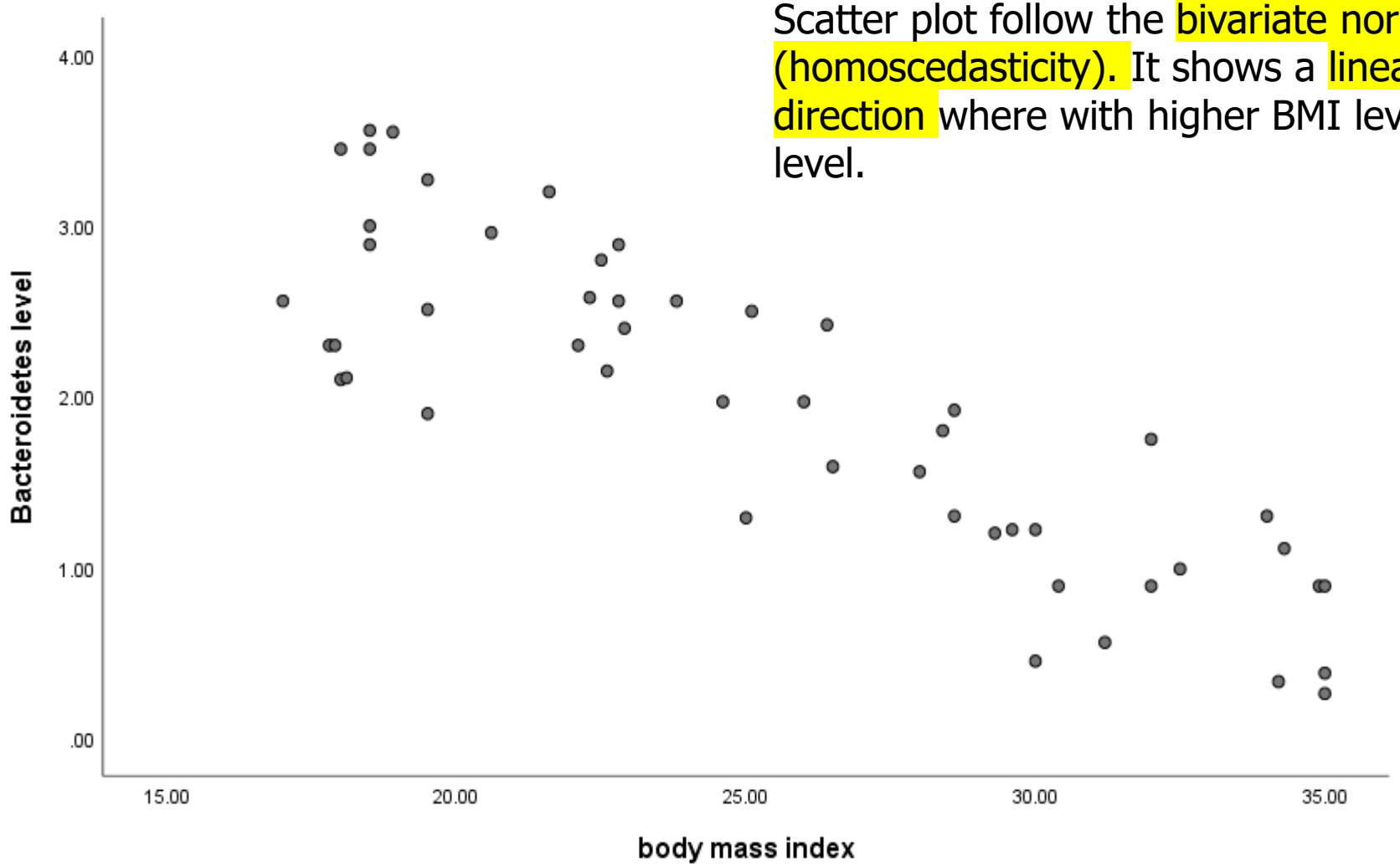
Research Question:

Changes in intestinal microbial composition are believed to be an important causal factor in development of obesity. The most common organisms in human gut microbiota are members of the gram-positive Firmicutes and the gram-negative Bacteroidetes. Previous studies indicate that obese persons have a significantly higher level of Firmicutes and lower level of Bacteroidetes compared to normal-weight and lean adults. A researcher wish to determine whether there is a relationship between the gut Bacteroidetes level ($x10^7$ copies/μL) and BMI (kg/m^2) among a group of adult.

Data Set: gut microbiota_BMI.sav

OUTPUT: Scatter Plot

Data Set: gut microbiota_BMI.sav

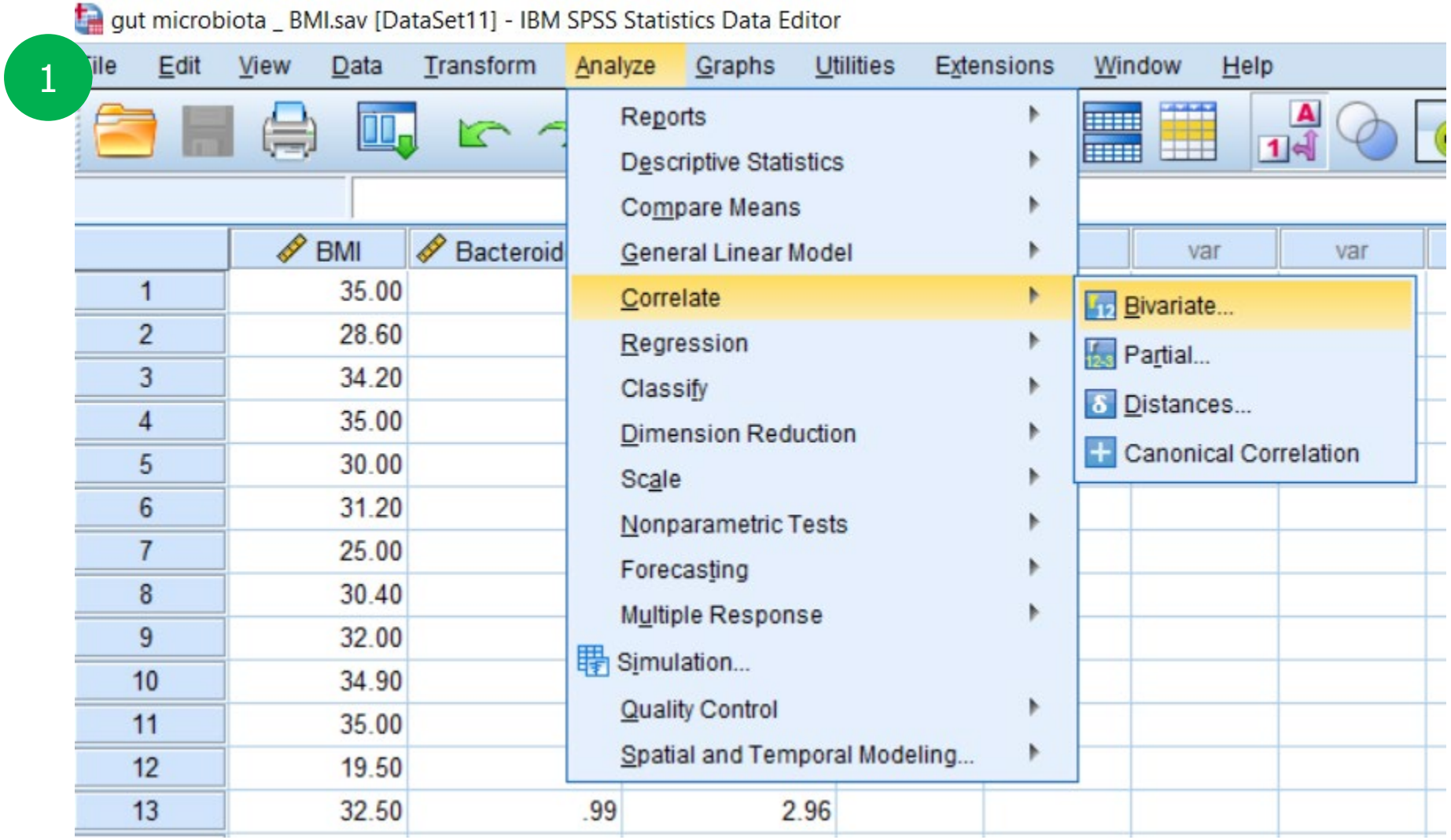


Scatter plot follow the bivariate normal distribution (homoscedasticity). It shows a linear relationship with negative direction where with higher BMI level, the lower Bacteroidetes level.

Pearson Correlation

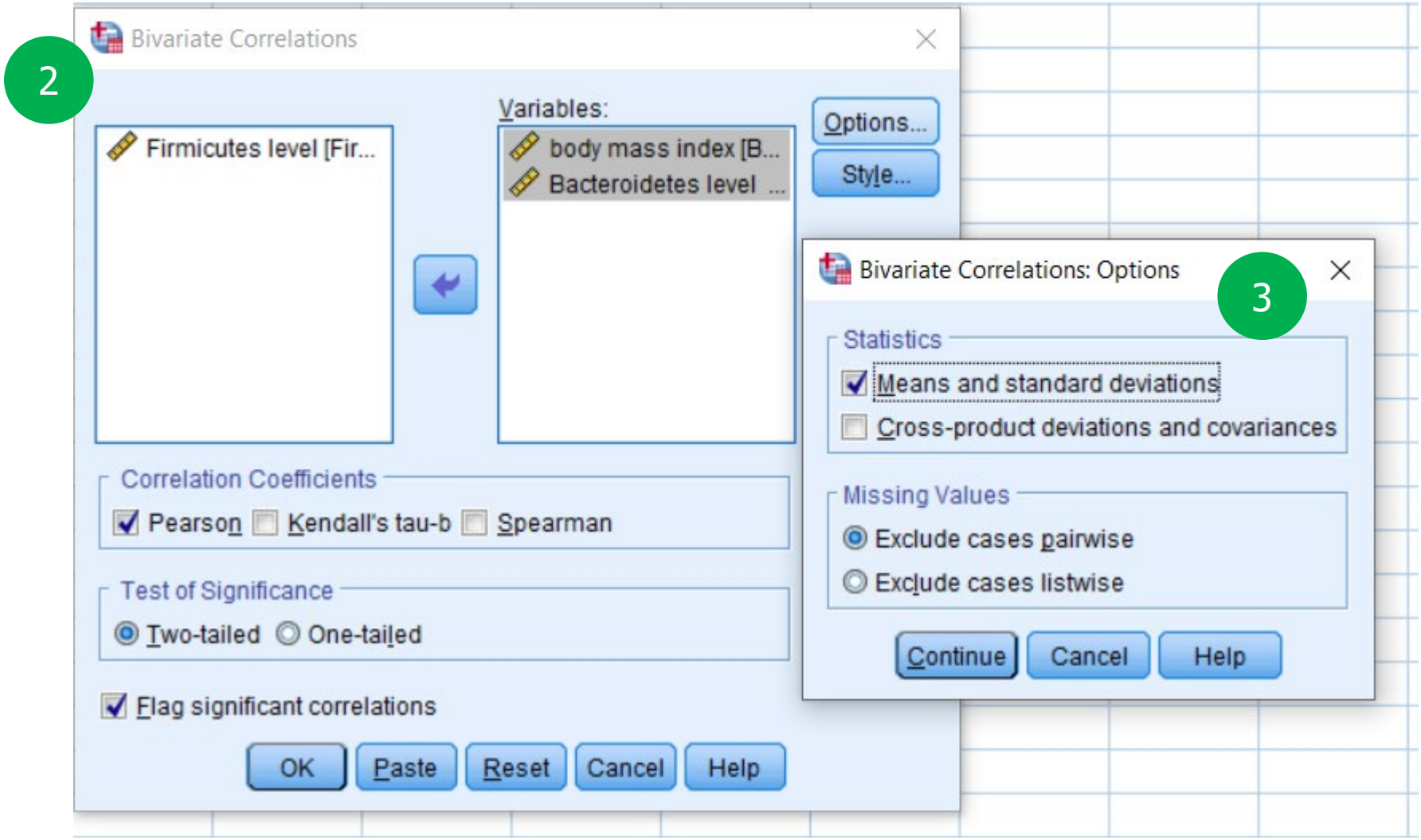
Data Set: gut microbiota_BMI.sav

Analyze > Correlate > Bivariate..



Pearson Correlation

Data Set: gut microbiota_BMI.sav



OUTPUT: Pearson Correlation

Data Set: gut microbiota_BMI.sav

Hypothesis Test for testing the correlation significance:

$H_0 : \rho = 0$ (There are no correlation between variables)

$H_A : \rho \neq 0$ (There exists correlation between variables)

Descriptive Statistics

	Mean	Std. Deviation	N
body mass index	25.3760	5.87178	50
Bacteroidetes level	1.9600	.92669	50

P-value 0.000 < α at 0.05 that results in rejecting the hypothesis null that indicate there is a significant (linear– shows from plot) correlation between BMI and Bacteroidetes level.

Correlations

		body mass index	Bacteroidetes level
body mass index	Pearson Correlation	1	-.859**
	Sig. (2-tailed)		.000
	N	50	50
Bacteroidetes level	Pearson Correlation	-.859**	1
	Sig. (2-tailed)	.000	
	N	50	50

**. Correlation is significant at the 0.01 level (2-tailed).

The correlation coefficient (r value) is -0.859 which suggest a strong negative correlation.

How about the gram-positive Firmicutes? Is there a relationship between the gut Firmicutes level and BMI (kg/m^2)?

Data Set: gut microbiota_BMI.sav

SPSS FOR NON-PARAMETRIC CORRELATION ANALYSIS

SPEARMAN 'S RANK ORDER CORRELATION

(SPEARMAN CORRELATION)

IBM SPSS 26

SPEARMAN CORRELATION

- Non- Parametric measure for correlation between variables
- Measure the strength and direction of association in correlation between two ranked variables (ordinal or continuous)
- Relatively robust to outliers
- Denoted by the symbol: r_s
- It determines the strength and direction of the monotonic but non-linear relationship

INTERPRETING SPEARMAN CORRELATION

R_s VALUE	Strength of Relationship
< 0.2	Very Weak
0.2 – 0.3	Weak
0.4 – 0.6	Moderate
0.7 – 0.8	Strong
> 0.8	Very Strong

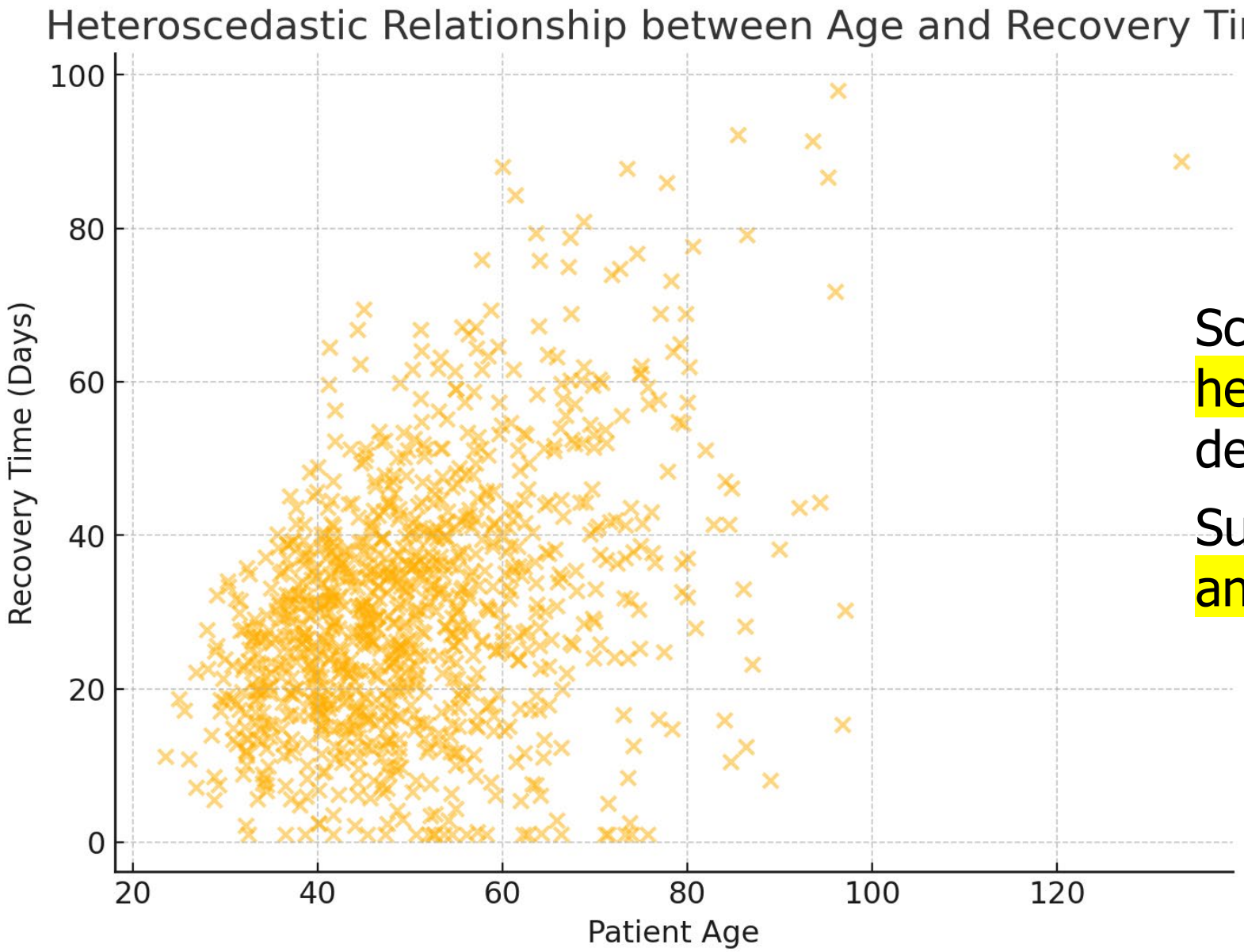
Research Question: Spearman correlation with both continuous data

This study focuses on analyzing the relationship between **patient age and recovery time** in a healthcare setting. The dataset contains medical records of **1,000 patients** from a multi-specialty hospital, including information on patient demographics, disease severity, and recovery duration. The goal is to assess whether an **association exists between patient age and the duration of hospital stay**.

Data Set: Hospital_Recovery.xlsx

OUTPUT: Scatter Plot

Data Set: Hospital_Recovery.xlsx



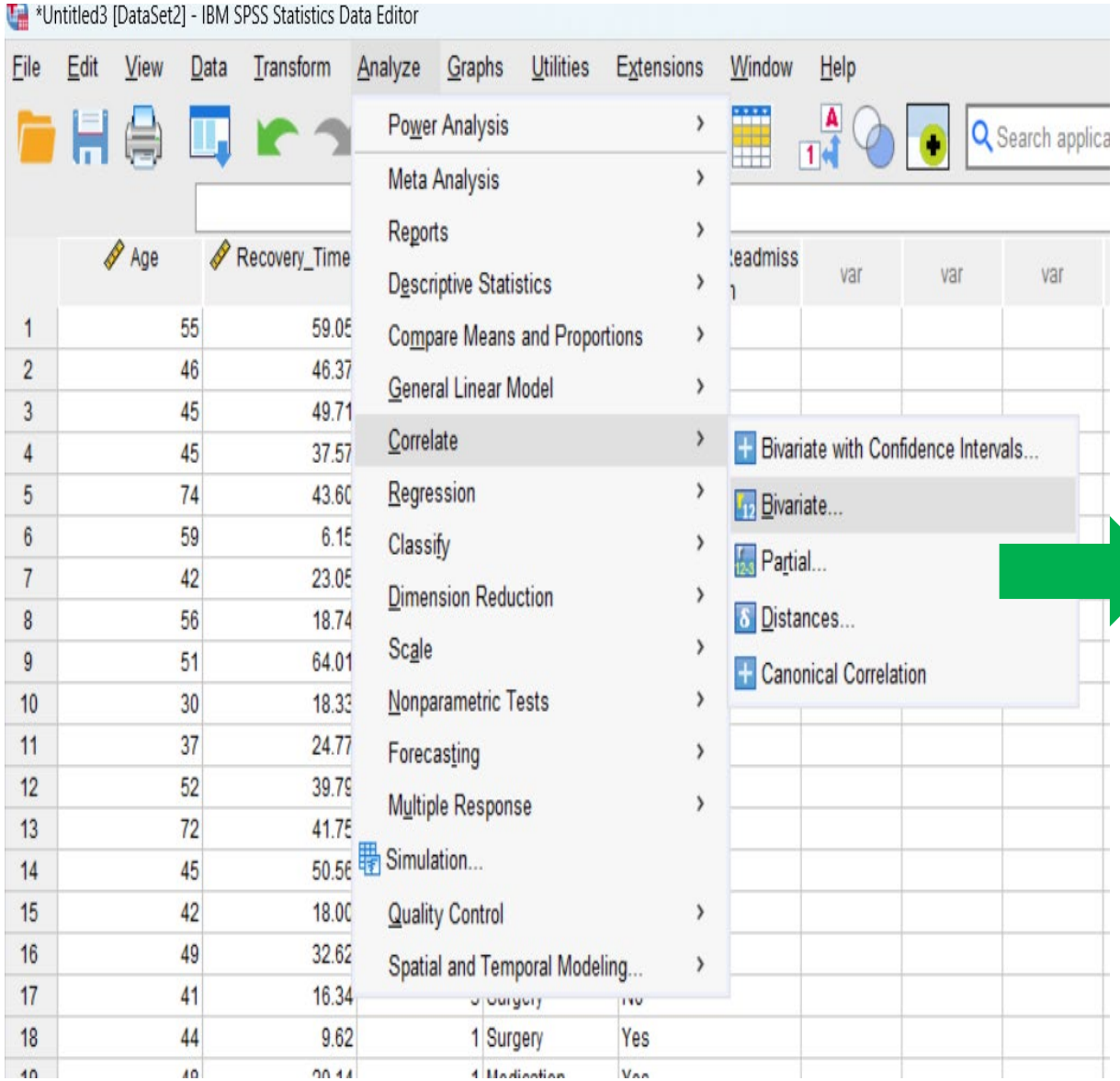
Scatter plot shows heteroscedasticity with some value detect as outliers.

Suggest: Spearman correlation analysis

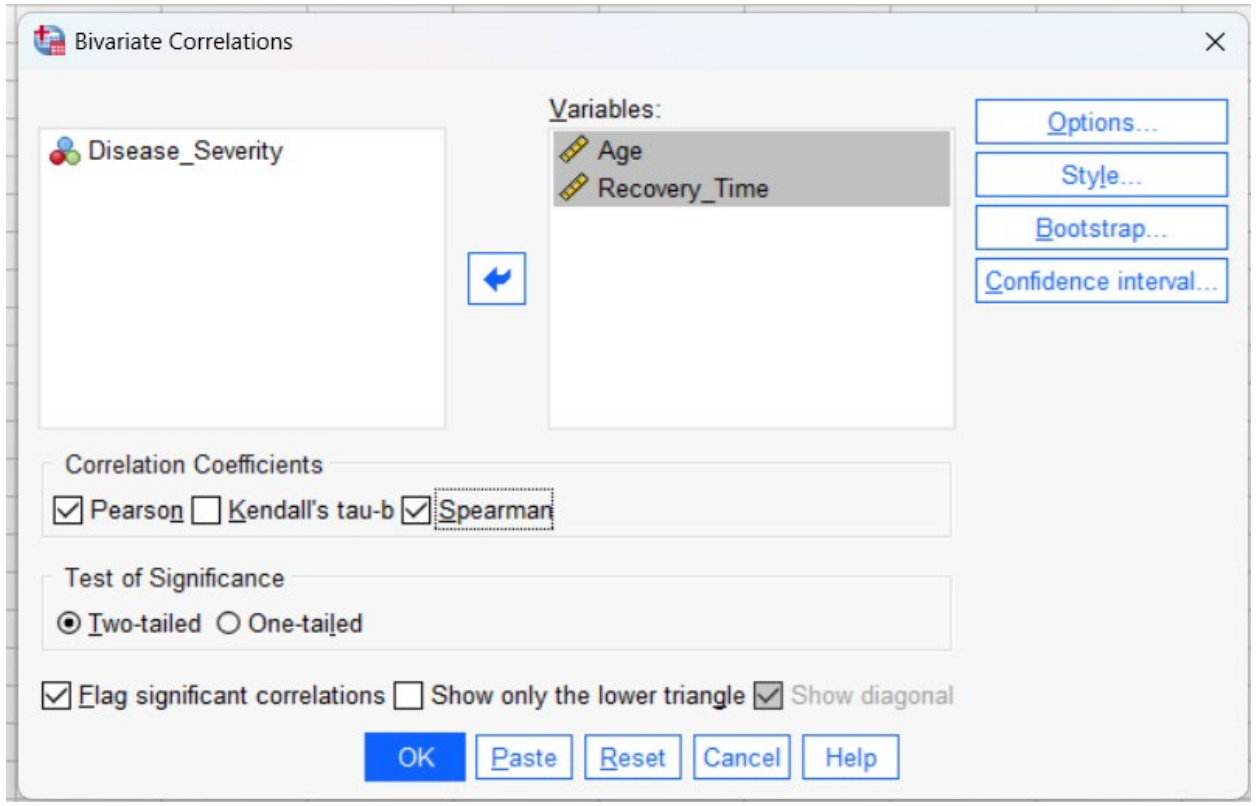
Spearman Correlation

Data Set: Hospital_Recovery.xlsx

Analyze > Correlate > Bivariate..



Select the variables to analyze



OUTPUT: Spearman Correlation

Data Set: Hospital_Recovery.xlsx

Hypothesis Test for testing the correlation significance:
$H_0 : \rho = 0$ (There are no correlation between variables)
$H_A : \rho \neq 0$ (There exists correlation between variables)

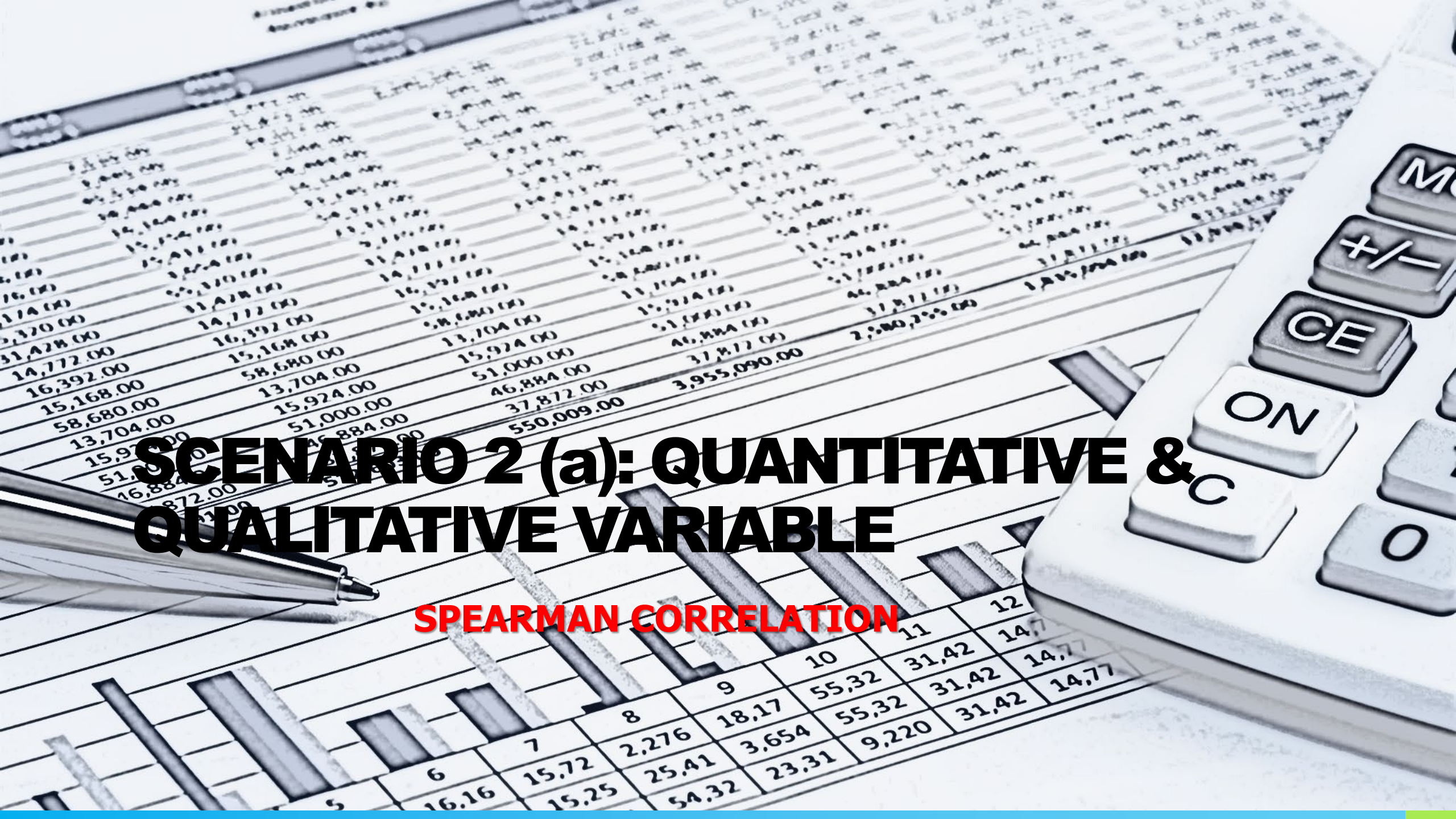
Correlations

			Age	Recovery_Time
Spearman's rho	Age	Correlation Coefficient	1.000	.352**
		Sig. (2-tailed)	.	<.001
		N	1000	1000
	Recovery_Time	Correlation Coefficient	.352**	1.000
		Sig. (2-tailed)	<.001	.
		N	1000	1000

**. Correlation is significant at the 0.01 level (2-tailed).

P-value 0.000 < α at 0.05 that results in rejecting the hypothesis null that indicate there is a significant correlation between age (in years) and recovery time.

The Spearman rank correlation (Rs) = 0.352 which suggest weak positive relationship between age (in years) and recovery_time



SCENARIO 2 (a): QUANTITATIVE & QUALITATIVE VARIABLE

SPEARMAN CORRELATION

Problem Statement:
This study investigates the relationship between **pain severity levels** (Mild, Moderate, Severe, Extreme Pain) and **patient age** (continuous) among post-surgical patients. The dataset includes **1,000 patient records** from a major hospital, covering demographics, medical conditions, and post-operative recovery details.

Research Question:
Is there a significant correlation between post-surgical pain severity and the length of hospital stay?

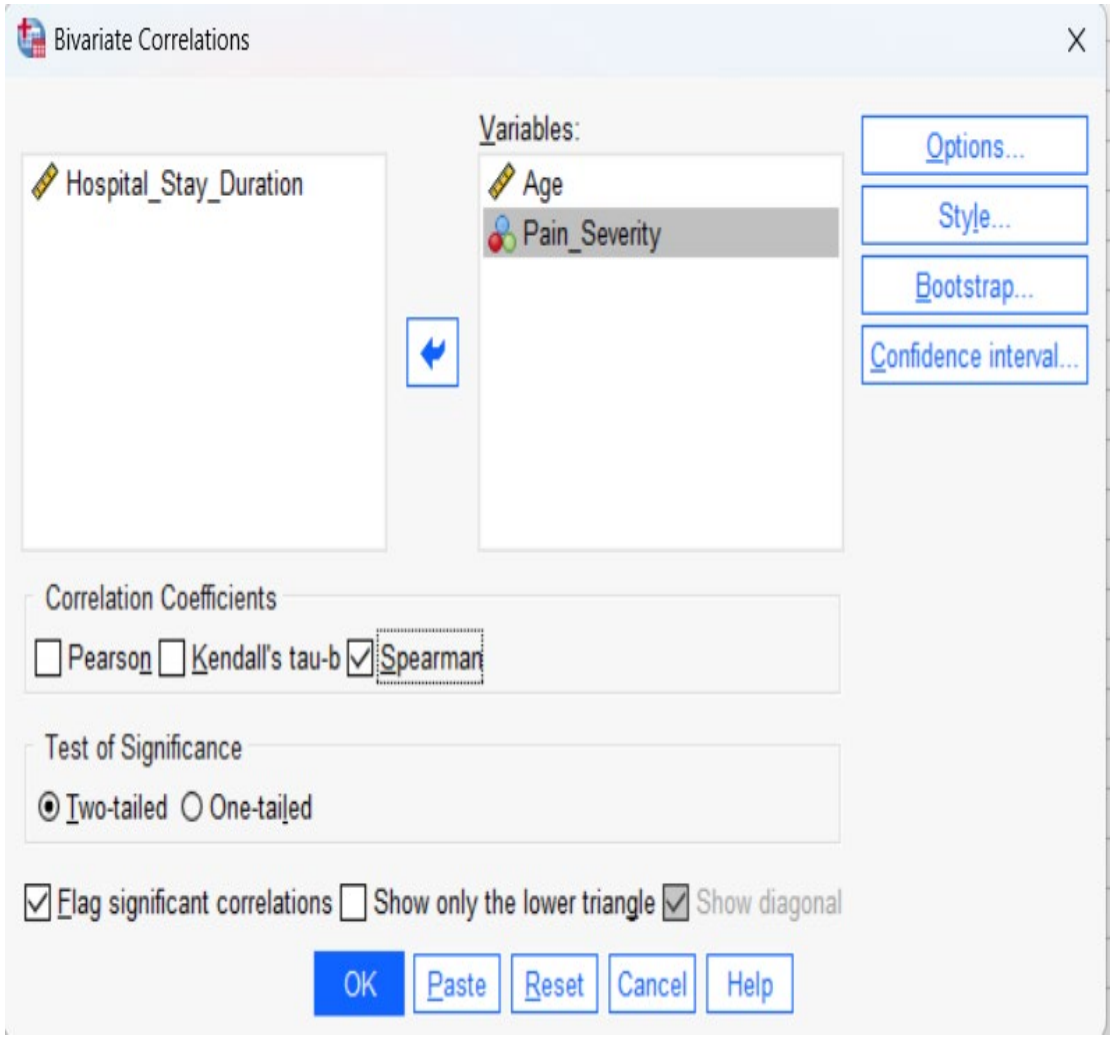
Data Set: Hospital_Pain_Recovery.xlsx

Note: Since **pain severity is an ordinal variable** and **patient age is a continuous variable**, **Spearman and Kendal's Tau correlation** is an appropriate method to analyze the **relationship** between these two variables.

Spearman Correlation

Data Set: Hospital_Pain_Recovery.xlsx

Analyze > Correlate > Bivariate..



RESULT

Correlations

			Age	Pain_Severity
Spearman's rho	Age	Correlation Coefficient	1.000	.018
		Sig. (2-tailed)	.	.575
		N	1000	1000
	Pain_Severity	Correlation Coefficient	.018	1.000
		Sig. (2-tailed)	.575	.
		N	1000	1000

Hypothesis Test for testing the correlation significance:

$H_0 : \rho = 0$ (There are no correlation between variables)

$H_A : \rho \neq 0$ (There exists correlation between variables)

P-value 0.575 > α at 0.05 that results in **ACCEPTING the hypothesis null** that indicate there is NO significant correlation between age (in years) with Pain Severity.

The Spearman rank correlation (Rs) = 0.018 which suggest **a NO relationship between** age (in years) with Pain Severity.

SCENARIO 2 (b): QUALITATIVE VARIABLE – ORDINAL HIERARCHY

KENDAL'S TAU CORRELATION

KENDAL'S TAU CORRELATION

- **Kendall's Tau** is a **correlation coefficient** used to measure the **strength and direction** of the relationship between **two sets of rankings** (ordinal data).
- Kendall's Tau compares the **number of concordant pairs** and **discordant pairs**.
  - **Concordant pair**: If one item is ranked higher than another in **both** lists.
  - **Discordant pair**: If one item is ranked higher in one list but **lower** in the other.
- 💡 Positive τ means the variables tend to **increase together**.
  💡 Negative τ means one variable **increases as the other decreases**.

INTERPRETING KENDAL'S TAU CORRELATION

τ VALUE	Strength of Relationship
< 0.2	Very Weak
0.2 – 0.3	Weak
0.4 – 0.6	Moderate
0.7 – 0.8	Strong
> 0.8	Very Strong

Problem Statement: *Relationship Between Patient Pain Level and Nurse's Response Time in a Triage Setting*

Scenario:
In a hospital triage unit, patients report their **self-assessed pain level** on a scale from 1 (no pain), 2 (mild pain), 3 (moderate pain), 4 (severe pain) to 5 (extreme pain). Nurses also record the **response time priority** on a scale from 1 (immediately attention needed), 2 (very urgent), 3 (urgent), 4 (less urgent) to 5 (not urgent). You want to determine if patients who report higher pain levels are more likely to be seen sooner. Is there a correlation between **patient-reported pain level** and **nurse-assigned response time**?

Data Set: kendall_tau_healthcare.xlsx

Hypothesis:

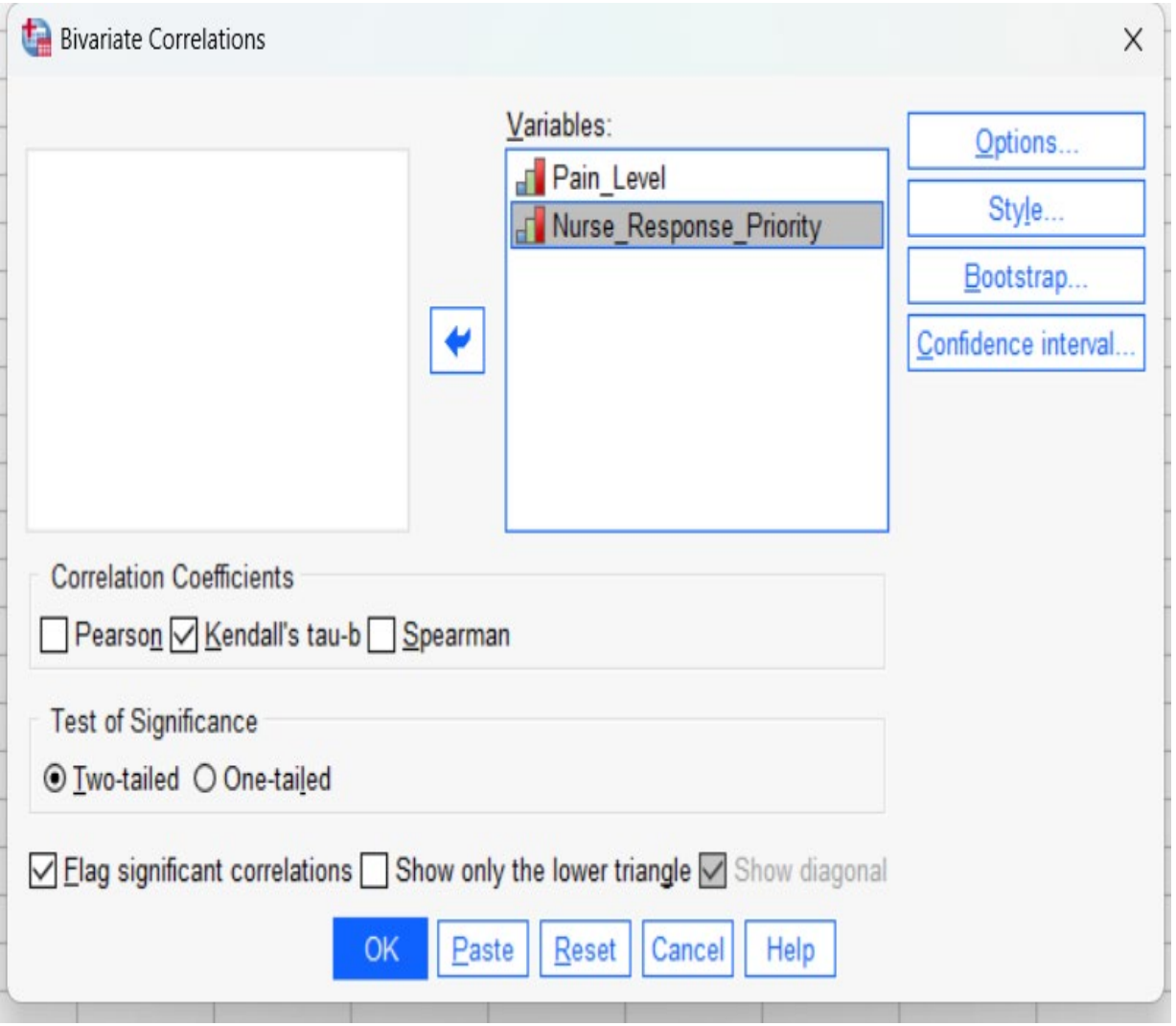
H_0: There is **no correlation** between pain level and nurse response time.

H_1: There **is** a correlation between pain level and nurse response time.

Kendall Tau Correlation

Data Set: kendall_tau_healthcare.xlsx

Analyze > Correlate > Bivariate..



RESULT

Correlations

			Pain_Level	Nurse_Response_Priority
Kendall's tau_b	Pain_Level	Correlation Coefficient	1.000	-.798**
		Sig. (2-tailed)	.	<.001
		N	30	30
	Nurse_Response_Priority	Correlation Coefficient	-.798**	1.000
		Sig. (2-tailed)	<.001	.
		N	30	30

**. Correlation is significant at the 0.01 level (2-tailed).

Hypothesis Test for testing the correlation significance:

H_0: There is **no correlation** between pain level and nurse response time.

H_1: There **is** a correlation between pain level and nurse response time.

P-value 0.000 < α at 0.05 that results in REJECTING the hypothesis null that indicate there EXISTS a correlation between pain level and nurse response time

The Kendall tau correlation = - 0.798 which suggest a NEGATIVE STRONG correlation between pain level and nurse response time (higher pain = faster response).

SCENARIO 3: BOTH ARE QUALITATIVE VARIABLE

CRAMER'S V AND PHI CORRELATION

PHI COEFFICIENT AND CRAMER'S V CORRELATION

- Using a Chi-Square Test of Independence allows researchers to assess whether the relationship observed between the nominal variables in a particular sample is also likely to be found in the population.
- Phi is a measure for the strength of an association between two categorical variables in a 2×2 contingency table.
- Cramer's V is an alternative to phi in tables bigger than 2×2 tabulation.
- Both measures varies between 0 and 1 without any negative value

INTERPRETING PHI COEFFICIENT AND CRAMER'S V CORRELATION

Phi & Cramer's VALUE	Strength of Relationship
> 0.2	Very Strong
> 0.15	Strong
> 0.10	Moderate
> 0.05	Weak
0 and < 0.05	No and very weak

Problem Statement:

To determine whether statistically significant associations exist between the **type of surgery**, **post-operative complications**, and **pain severity** experienced by 200 patients during recovery. This study aims to quantify the strength of association among these categorical variables.

Research Question: Phi Coefficient and Cramer's V Correlation

1. Does Post Operative Complication have a relation with pain severity?
2. Does Post Operative Complication have a relation with Type of Surgery?
3. Does Type of Surgery have a relation with pain severity?

Data Set: Hospital_Pain_Recovery_Cramers.xlsx

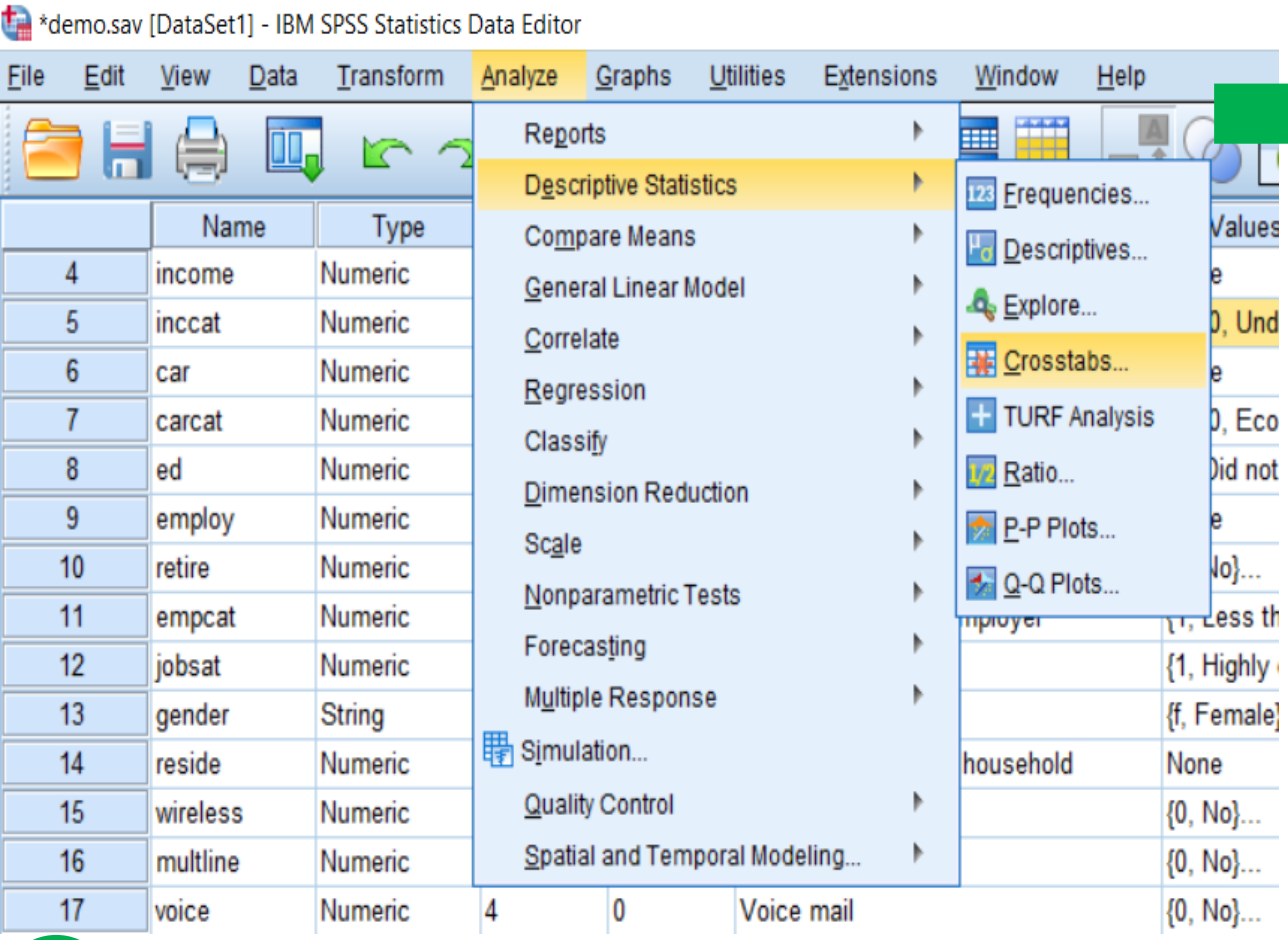
Variables:

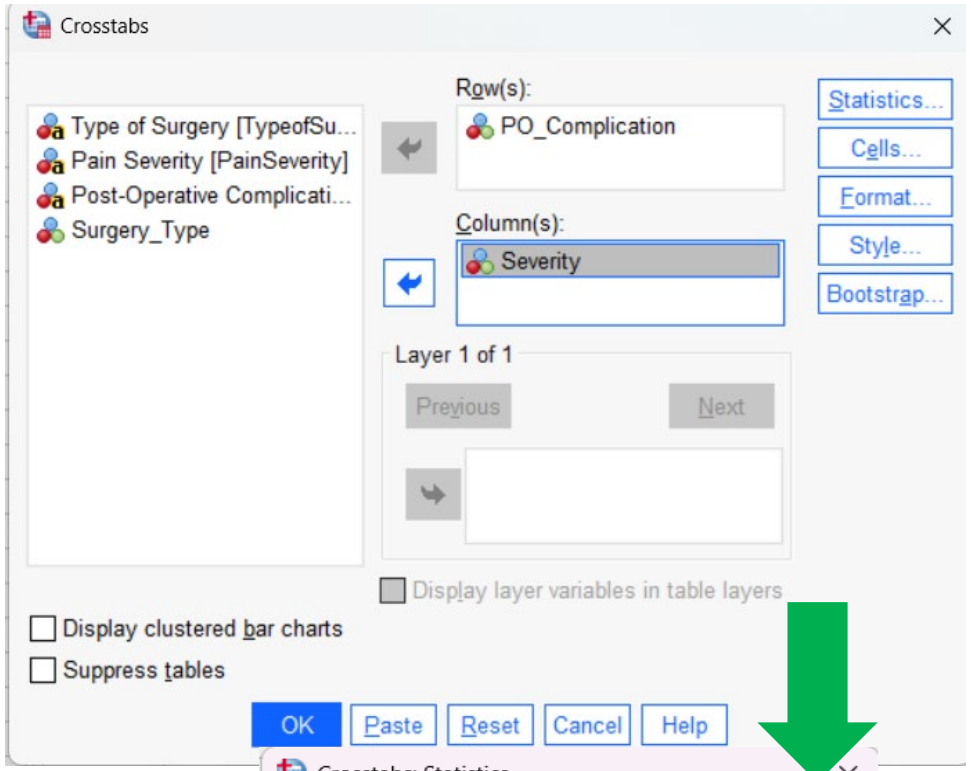
1. Pain Severity (Ordinal: Mild, Moderate, Severe, Extreme Pain)
2. Type of Surgery (Categorical: General, Orthopedic, Cardiovascular, Neurological, Other)
3. Post-Operative Complications (Binary: Yes/No)

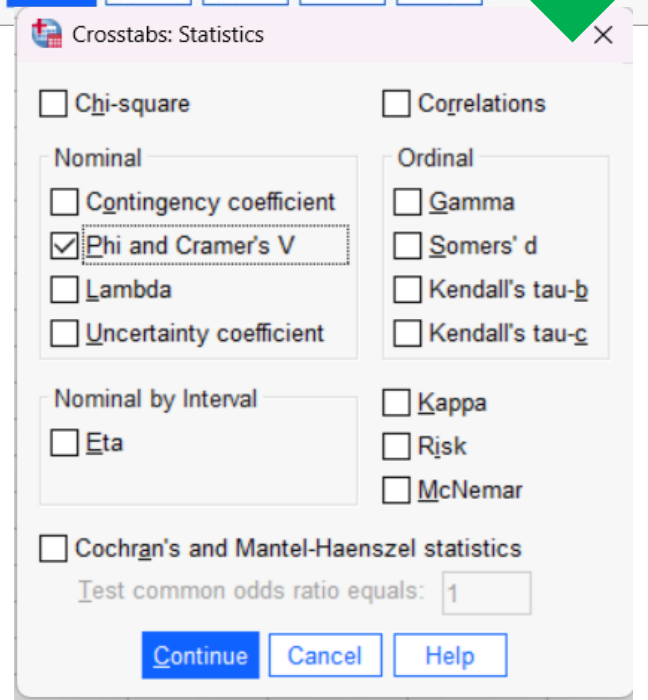
Phi Coefficient and Cramer's V correlation

Data Set: Hospital_Pain_Recovery_Cramers.xlsx

Analyze > Descriptive Statistics > Crosstabs..







1

2

3

OUTPUT → Phi Coefficient and Cramer's V correlation: Does Post Operative Complication have a relation with pain severity?

Data Set: Hospital_Pain_Recovery_Cramers.xlsx

1

PO_Complication * Severity Crosstabulation

Count

		Severity				
		Mild	Moderate	Severe	Extreme Pain	Total
PO_Complication	No	42	51	19	16	128
	Yes	7	10	34	21	72
Total		49	61	53	37	200

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.476	<.001
	Cramer's V	.476	<.001
N of Valid Cases		200	

3

Hypothesis Test for testing the correlation significance:

H_0 : There are no correlation between variables

H_A : There exists correlation between variables

2

The cramer's V = 0. 476 which suggest VERY STRONG relationship between **post operative complication and pain severity.**

P-value 0.000 < α at 0.05 that results in REJECTING the hypothesis null that indicate there EXISTS significant correlation between post operative complication and pain severity

OUTPUT → Phi Coefficient and Cramer's V correlation: Does Post Operative Complication have a relation with Type of Surgery?

Data Set: Hospital_Pain_Recovery_Cramers.xlsx

1

PO_Complication * Surgery_Type Crosstabulation

Count

		Surgery_Type					
		General	Orthopedic	Cardiovascular	Neurological	Other	Total
PO_Complication	No	38	29	23	19	19	128
	Yes	8	21	17	10	16	72
Total		46	50	40	29	35	200

3

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.222	.042
	Cramer's V	.222	.042
N of Valid Cases		200	

The cramer's V = 0.222 which suggest VERY STRONG relationship between **post operative complication and Type of Surgery**.

Hypothesis Test for testing the correlation significance: 2

H_0 : There are no correlation between variables

H_A : There exists correlation between variables

P-value 0.042 < α at 0.05 that results in REJECTING the hypothesis null that indicate there EXISTS significant correlation between post operative complication and Type of Surgery

OUTPUT → Phi Coefficient and Cramer's V correlation: Does Type of Surgery have a relation with pain severity?

Data Set: Hospital_Pain_Recovery_Cramers.xlsx

1

Surgery_Type * Severity Crosstabulation

Count

		Severity				
		Mild	Moderate	Severe	Extreme Pain	Total
Surgery_Type	General	19	15	6	6	46
	Orthopedic	9	10	13	18	50
	Cardiovascular	9	14	12	5	40
	Neurological	5	10	10	4	29
	Other	7	12	12	4	35
Total		49	61	53	37	200

Hypothesis Test for testing the correlation significance:

H_0 : There are no correlation between variables

H_A : There exists correlation between variables

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.357	.013
	Cramer's V	.206	.013
N of Valid Cases		200	

2

The cramer's V = 0.206 which suggest VERY STRONG relationship between **surgery type and pain severity.**

3

P-value 0.013 < α at 0.05 that results in REJECTING the hypothesis null that indicate there EXISTS significant correlation between post operative complication and pain severity

SUMMARY

- Correlation analysis is used to quantify the degree to which two variables are related.
- Through the correlation analysis, you evaluate correlation coefficient that tells you how much one variable changes when the other one does.
- Pearson correlation is used for correlation analysis for BOTH continuous variables with bivariate normal distribution.
- Spearman correlation is used for correlation analysis for BOTH continuous nonparametric data and can be mixed of continuous variable and ordinal variables.
- Kendall Tau correlation is used for correlation analysis for BOTH ordinal variables.
- Cramer's V correlation is used for correlation analysis for BOTH qualitative variables.

TEST YOUR UNDERSTANDING

Exploring Relationships Between Patient Characteristics and Recovery Outcomes

Objective:
In this activity, you will analyze relationships between different patient factors and health outcomes using three correlation techniques: **Pearson**, **Spearman**, **Kendall's Tau** and **Cramer's V**.

Dataset:
📂 **Healthcare_Correlation_Study_Final.xlsx**

Variables in Dataset:

1. **Age** (Numerical)
2. **Recovery Time (Days)** (Numerical)
3. **Pain Score** (Numerical score from 1–10)
4. **Hospital Type** (Categorical: Public, Private, Specialty)
5. **Smoking Status** (Categorical: Smoker, Non-Smoker)
6. **Insurance Type** (Categorical: Private Plan, Public Plan)
7. **Diet Category** (Categorical: Unhealthy, Moderate, Healthy)
8. **Physical Activity Level** (Categorical: Low, Moderate, High)

Research Question:

1. What is the pattern between **Age and Recovery Time (Days)**? Suggest suitable correlation coefficient measurement.
2. Does exists linear positive relationship between Age and Recovery Time (days)?
3. What do you think a better correlation coefficient measurement for **Age and Pain Score**? Calculate its value and what can be conclude?
4. What is the relationship between **Recovery Time (days) with Diet Category**. What can be conclude from the analysis?
5. Identify the relationship between:

   a) Smoking status with Diet Category

   b) Physical Activity with Diet Category


THANK YOU