# SUMMARY: DATA ANALYSIS-IV

## SESSION OVERVIEW:

By the end of this session, the students will be able to:

- Understand bivariate analysis.
- Understand different types of bivariate analysis
- Perform bivariate analysis in datasets.

## KEY TOPICS AND EXAMPLES:

***In this session there might be certain topics which we have already covered in the previous session and it might be a revision to some of you but this session will make you aware of those topics in terms of analysis point of view.***

### Understanding bivariate analysis:

Bivariate analysis is a statistical method used to explore the relationship between two variables. It aims to understand how changes in one variable are associated with changes in another variable. Bivariate analysis is essential for patterns, trends, and associations in data.

**Why is it important to perform bivariate analysis in dataset:**

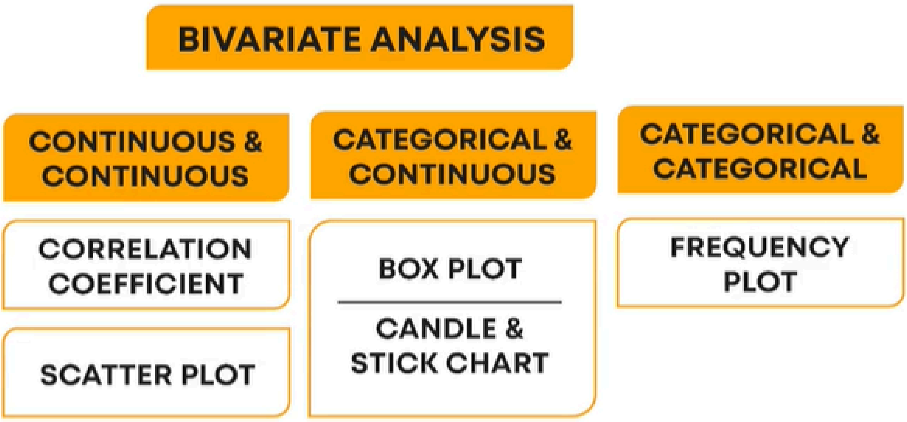
- **Identifying Relationships:** It helps to identify and understand the relationships between two variables.
- **Detecting Patterns and Trends**: Bivariate analysis allows analysts to detect patterns and trends that may exist between two variables. This information is valuable for making informed decisions in various fields such as business, healthcare, social sciences, and more.
- **Diagnostic Purposes:** In some cases, bivariate analysis can serve diagnostic purposes by uncovering unexpected relationships or outliers in the data.

**Example:**
Let's consider an example of bivariate analysis involving two variables: "Study Time" and "Exam Score."
**Variables:**
**Independent Variable (X)**: Study Time
This variable represents the number of hours a student spends studying for an exam.
**Dependent Variable (Y):** Exam Score
This variable represents the score achieved by the student on the exam.

**Conclusion that we will acquire from the given information:**
Allows us to explore and understand the relationship between study time and exam score. By analyzing the data and interpreting the results, we gain insights into how study habits may influence academic performance.

**Understanding different types of bivariate analysis:**

Here we will be only working on the contingency table using pivot tables in Excel for different types of bivariate analysis.



BIVARIATE ANALYSIS

CONTINUOUS & CONTINUOUS
- CORRELATION COEFFICIENT
- SCATTER PLOT

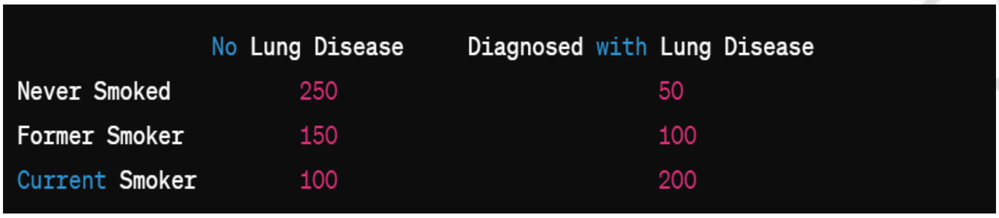
CATEGORICAL & CONTINUOUS
- BOX PLOT
- CANDLE & STICK CHART

CATEGORICAL & CATEGORICAL
- FREQUENCY PLOT

1. **Categorical to categorical:**

   Bivariate analysis of categorical variables involves exploring the relationship between two categorical variables. In this analysis, both variables consist of categories or groups rather than numerical values. The goal is to understand whether there is an association or dependency between the two variables and to describe the nature of this relationship.

   **Example:**
   Let's consider an example of bivariate analysis involving two categorical variables: "Smoking Status" and "Lung Disease Diagnosis."
   - Define Variables:
     Smoking Status: Categories could include "Never Smoked," "Former Smoker," and "Current Smoker."
     Lung Disease Diagnosis: Categories could include "No Lung Disease" and "Diagnosed with Lung Disease."
   - Gather data on individuals, recording their smoking status and whether they have been diagnosed with lung disease.
   - Construct a contingency table to display the frequencies of each combination of categories.



```
                No Lung Disease     Diagnosed with Lung Disease
Never Smoked          250                       50
Former Smoker         150                       100
Current Smoker        100                       200
```

This is the dataset in which we have already performed and created some pivot tables and charts. But now we will be understanding the dataset using bivariate analysis.

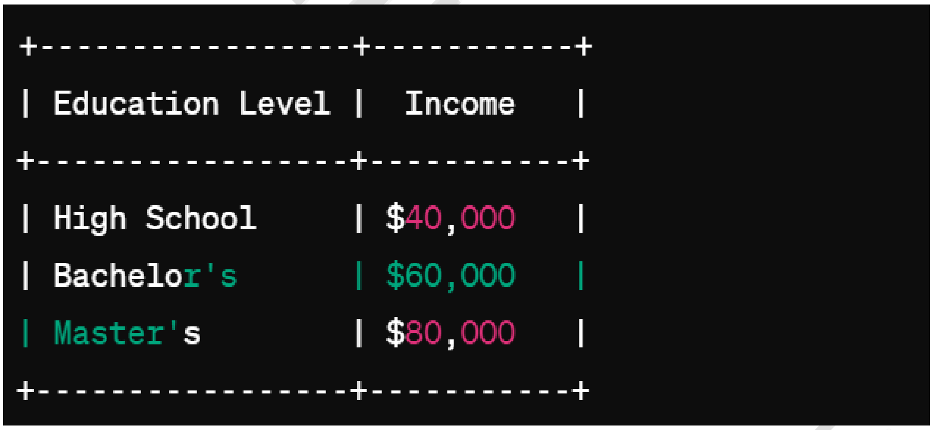
2. **Categorical to continuous:**

   The categorical variable divides the data into distinct groups or categories, while the continuous variable represents quantitative data that can take any value within a range.

   The purpose of this analysis is to examine how the continuous variable varies across different categories of the categorical variable. It aims to identify whether there are significant differences in the distribution or means of the continuous variable among the categories of the categorical variable.

   **Example:**
   Let's consider an example of bivariate analysis involving one categorical variable: "Education level" and another continuous variable "Income". Here we are trying to understand how the level of education affects the income.

   - **Categorical Variable**: "Education Level"
     - Categories: High School Diploma, Bachelor's Degree, Master's Degree
   - **Continuous Variable:** "Income"
   - **Purpose:** We want to explore how income varies across different education levels.
   - Construct a contingency table to display the frequencies of each combination of categories and the continuous variable that we are considering.



```
+----------------+----------+
| Education Level |  Income   |
+----------------+----------+
| High School     | $40,000   |
| Bachelor's      | $60,000   |
| Master's        | $80,000   |
+----------------+----------+
```

   This is the dataset in which we have already performed and created some pivot tables and charts. But now we will be understanding the dataset using bivariate analysis.

3. **Continuous to continuous:**

   Continuous to continuous bivariate analysis involves exploring the relationship between two continuous variables. This type of analysis is used to understand how changes in one variable correspond to changes in another variable.

   The following are the different types of methods to determine the relationships between two continuous variable:

   1. **Correlation Coefficient:**

The correlation coefficient, such as Pearson's correlation coefficient, is a statistical measure used to quantify the strength and direction of the linear relationship between two continuous variables.
The strength of the correlation is determined by correlation coefficient, which varies between -1 to +1. Correlation coefficient is used to measure the strength and direction of a linear relationship between two variables. Here is a table that shows the strength of the correlation based on its magnitude (absolute value):

| Magnitude of Correlation | Strength of Correlation |
|---|---|
| 0.00-0.19 | Very Weak |
| 0.20-0.39 | Weak |
| 0.40-0.59 | Moderate |
| 0.60-0.79 | Strong |
| 0.80-1.00 | Very strong |

**Note:** This is just a reference. The strength of the correlation varies from industry to industry and thus industry knowledge is important.

**Example of 0.8 being a weak correlation -**
consider the relationship between temperature and ice cream sales. We might expect a very high correlation because, intuitively, as the temperature increases, more people buy ice cream. However, if we find a correlation of 0.8, it might be considered weaker than expected. This is because, although 0.8 indicates a strong relationship, the direct impact of temperature on ice cream sales is so intuitive and expected to be so strong that any deviation from a near-perfect correlation (closer to 1) might suggest other factors are also at play (like rain on a hot day reducing sales), or there might be inaccuracies in the data.

**Example for 0.8 as strong correlation -** same example of study time vs exam score

**Note:** The sign of the correlation (positive or negative) indicates the direction of the relationship, while the magnitude (absolute value) indicates the strength of the relationship. A positive correlation indicates that both variables tend to increase or decrease together, while a negative correlation indicates that as one variable increases, the other tends to decrease.

**Direction of the correlation**

- **Positive Correlation:** When the value of one variable increases, the value of the other variable also tends to increase. For example, there is a positive correlation between the amount of exercise a person gets and their overall health. As a person exercises more, their health tends to improve.

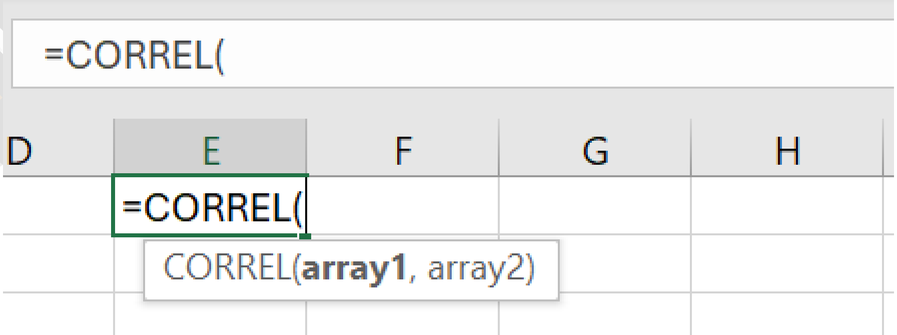
- **Negative Correlation:** When the value of one variable increases, the value of the other variable tends to decrease. For example, there is a negative correlation between the number of hours a person spends watching TV and their academic performance.

As a person spends more time watching TV, their academic performance tends to decrease.

- **No Correlation:** When there is no relationship between two variables. For example, there is no correlation between a person's shoe size and their IQ. The size of a person's feet has no impact on their intelligence.

**Correlation Coefficient in Excel:**

To calculate the correlation coefficient in Excel, you can use the CORREL function.



```
=CORREL(
```

```
=CORREL(
CORREL(array1, array2)
```

Here, array1 and array2 are the two arrays of data for which you want to calculate the correlation coefficient.
This is the dataset in which we have already performed and created some pivot tables and charts. But now we will be understanding the dataset using bivariate analysis.

2. **Scatter Plot:**

**Uses:**

- **Exploring Relationships:** Scatter plots are used to visually explore the relationship between two continuous variables. They help in identifying patterns, trends, clusters, outliers, and the overall nature of the relationship between the variables.

- **Assessing Correlation:** Scatter plots are particularly useful for assessing correlation between two continuous variables. The pattern of points on the plot can indicate whether there is a positive correlation, negative correlation, or no correlation between the variables.

- **Identifying Trends**: Scatter plots can reveal trends or patterns in the data, such as linear trends, nonlinear trends, exponential growth, or decay. This information is valuable for understanding the behavior of the variables over a range of values.

- **Detecting Outliers:** Outliers, or data points that deviate significantly from the overall pattern, can be easily identified on a scatter plot. Outliers may represent data errors, measurement inaccuracies, or genuine anomalies that require further investigation.

3. **Box plot:**

Creating a box plot for a categorical to continuous bivariate analysis in Excel is straightforward since box plots are ideally suited for visually displaying the distribution of a continuous variable across different categories of a categorical variable.

Excel provides an in-bulit feature of creating box plots. Categorical to continuous bivariate analysis is used when you want to understand the relationship or compare the characteristics of a continuous variable across different groups defined by a categorical variable.

(**NOTE**: How to create box plots has been taught in the previous sessions and this is where we learn about the implementation of the box plot through bivariate analysis.)

**Understanding multivariate analysis:**

Multivariate analysis is used when we have a dataset which is a little more complex and the analysis should have multiple variables involved.

**Example:** Suppose we have a dataset containing information about students, including their scores on three different subjects: Math, Science, and English. We want to explore the relationships among these subjects and identify any underlying patterns in the students' performance.

## ***FUN TIME:***

***This is the part where we will be using the ABC Company dataset to perform the complete bivariate analysis and understand the types of customers who will be willing to buy the product and what kind of customers are more inclined towards this company and its products.***

***Using bivariate analysis, we will be acquiring the answers of the following questions:***

- ***What is the correlation between the amount of fruits and sweets?***
- ***What is the correlation between income and amount spent on fruits/sweets?***
- ***What is the correlation between income and the total purchase?***
- ***What is the correlation between Age and total purchases?***
- ***What is the correlation between TotalKidsHome and amount spent on sweets/fruits?***
- ***How the levels of marital status influences the amount spent on fruits/sweets?***
- ***By using the box plot concept, we can find the correlation between education and the amount spent on sweets/fruits.***

***So, these questions will help us understand different aspects.***
***For e.g. as the number of kids at home increases, how does the amount spent on sweets/fruits differ?***
***As the Age increases, do they get more attracted to spending on sweets/fruits of the company or not?***