SUMMARY: DATA ANALYSIS-III

SESSION OVERVIEW:

By the end of this lecture, you will be able to:

- Understand the different aspects involved in continuous univariate analysis.
- Understand the different aspects involved in categorical univariate analysis.
- Perform univariate analysis in datasets.
- Understand the introductory topics of bivariate analysis.

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.

Understand the different aspects involved in continuous univariate analysis:

1. **Types of continuous univariate feature:**
    a. **Descriptive statistics:**

Descriptive analysis is a fundamental step in understanding and summarizing the characteristics of a dataset, particularly for continuous univariate features. It provides a concise overview of the central tendency, dispersion, and distribution of the data.

Measures of Central Tendency:

(This part has been covered in detail in session 8 when we were discussing data cleaning. Here we will be discussing mean, median in terms of data analysis perspective.)

Mean: The arithmetic average of the values, calculated by summing all the values and dividing by the total number of observations.
Median: The middle value in the ordered sequence of values, with an equal number of observations above and below it.
Mode: The value(s) that occur(s) most frequently in the dataset.

Measures of Dispersion:

(This part has been covered in detail when we were discussing data cleaning. Here we will be discussing mean, median in terms of data analysis perspective.)

Variance: The average squared deviation from the mean, quantifying how spread out the values are from the mean.
Standard Deviation: The square root of the variance, providing a more interpretable measure of spread in the same units as the original data.
Interquartile Range (IQR): The range between the 25th and 75th percentiles, representing the spread of the middle 50% of the data.

b. Trimmed mean:

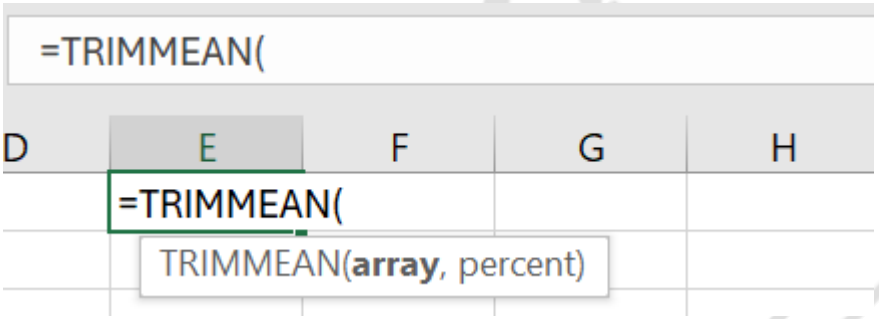
The trimmed mean is a robust measure of central tendency that excludes a certain percentage of the highest and lowest values from the calculation. It is particularly useful when there are outliers or extreme values in the data that could heavily influence the regular mean.

The importance of the trimmed mean in univariate analysis lies in its ability to provide a more reliable estimate of the central tendency when dealing with skewed distributions or datasets containing outliers. By trimming a small percentage of the extreme values (e.g., 5% from each tail), the trimmed mean reduces the impact of these extreme observations on the overall measure of central tendency.

In situations where outliers or heavy-tailed distributions are present, the regular mean can be heavily influenced and may not accurately represent the central tendency of the majority of the data. The trimmed mean, on the other hand, provides a more robust estimate by focusing on the central values and minimizing the effect of extreme observations.

Important:
The amount it trims is determined by the percentage value you provide as the second argument to the function. For example, if you set this percentage to 0.1 (or 10%), Excel will remove the lowest 5% and the highest 5% of the values from the data set before calculating the average of the remaining values.

In excel, there is a function called TRIMMEAN () which helps us get the trim mean value.



Comparing the TRIMMEAN() and AVERAGE() values, we will get the idea of the presence of outliers in the dataset.
The difference between the TRIMMEAN value and the AVERAGE value helps us determine the extent of the presence of outliers in the dataset.

- **If the difference between the TRIMMEAN value and the AVERAGE value is large then it indicates that there is a significant amount of outliers in the dataset.**
- **If the difference between the TRIMMEAN value and the AVERAGE value is small then it indicates that there is a lesser amount of outliers in the dataset.**

Example: [(Reference Dataset)]
Here we are going to use this dataset to understand the distribution of the data using several methods. We will use the TRIM MEAN and AVERAGE function for now to check if there is any outlier present and if yes, then what is the difference between the average value and the trimmean value.

c. Creating histograms in Excel:

- Histograms provide a graphical representation of the frequency or density of data values within specified bins or intervals.
- They help identify the shape of the distribution, such as normal, skewed (positively or negatively), bimodal, or multimodal.
- This visual information can guide further analyses and transformations to meet assumptions of statistical tests or modeling techniques.
- Histograms can reveal the presence of outliers or anomalies in the data, which may appear as isolated bars or spikes at the extremes of the distribution.

Determining bin sizes and frequency:
Binning in Excel is a powerful tool used to organize and analyze data. It is an excellent way to quickly categorize data into useful groups, such as age groups, income levels, or product categories.

A bin is a type of data structure used to categorize data into groups. It is also referred to as a “binning” or “bucket” system. This data structure is used to make information more manageable and easier to analyze. In an MS Excel, bins are created by setting up ranges of values within a column. For example, a single column could contain age groups such as 18-25, 26-35, and 36-45.

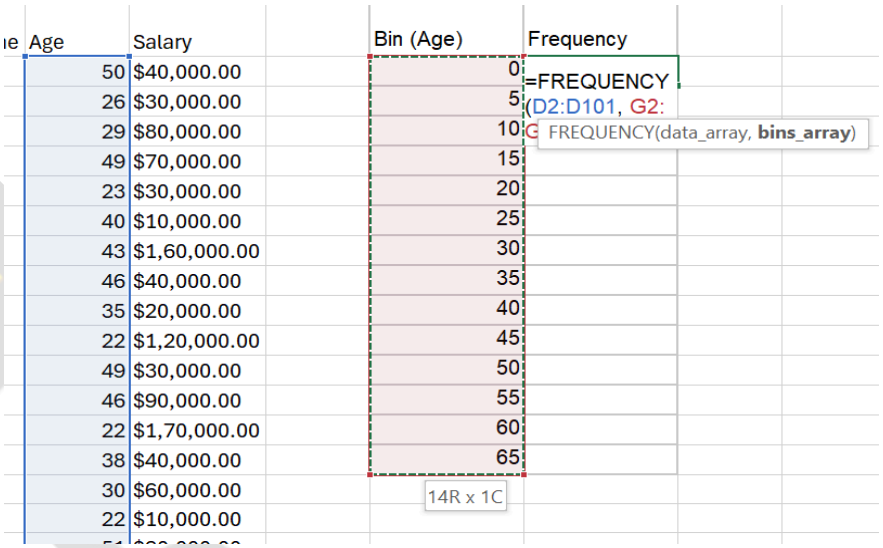
How to create bins in Excel:
[(Reference Dataset)]

1. **Create manually:**
   We can simply create the bins according to our needs. For example, suppose you have a dataset on employee age and you want to bin the age column by a difference of 5. So, in this scenario you can manually create bins in a separate column and apply the frequency function in Excel which helps us to create the frequency table.

   A frequency table is a useful tool that helps you understand how data is distributed in your dataset. You can use it to create a histogram, which gives you have an even better idea of the data distribution.

   For example, you have a list of ages for different people. By creating a frequency table with a bin size of 5, you can see how many people fall into each age group. This helps you understand how Age is distributed across

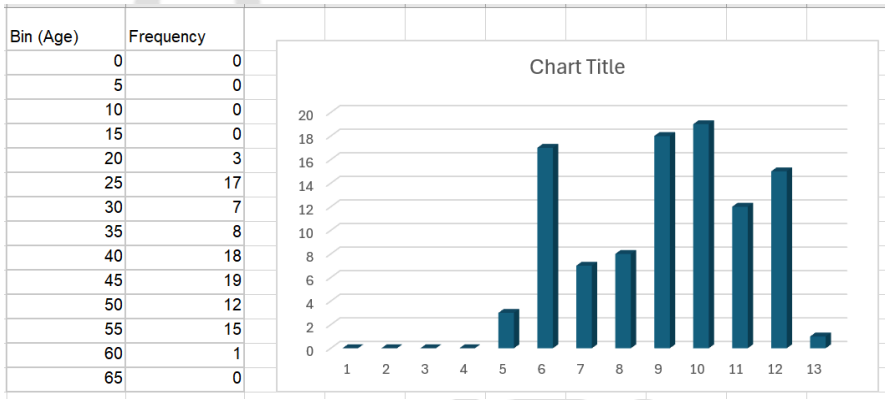
different categories or classes. If you want to analyze data like this, you can use a frequency table and histogram to make things easier to understand.



ıe	Age	Salary	Bin (Age)	Frequency
	50	$40,000.00	0	=FREQUENCY
	26	$30,000.00	5	(D2:D101, G2:
	29	$80,000.00	10	G
	49	$70,000.00	15	
	23	$30,000.00	20	
	40	$10,000.00	25	
	43	$1,60,000.00	30	
	46	$40,000.00	35	
	35	$20,000.00	40	
	22	$1,20,000.00	45	
	49	$30,000.00	50	
	46	$90,000.00	55	
	22	$1,70,000.00	60	
	38	$40,000.00	65	
	30	$60,000.00		
	22	$10,000.00		

FREQUENCY(data_array, **bins_array**)

14R x 1C

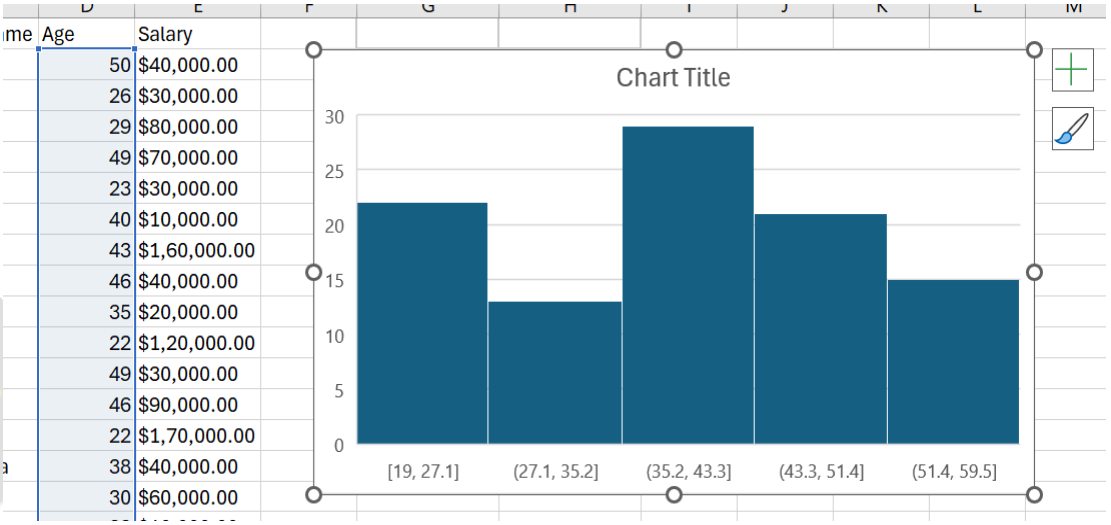


Bin (Age)	Frequency
0	0
5	0
10	0
15	0
20	3
25	17
30	7
35	8
40	18
45	19
50	12
55	15
60	1
65	0

Using the column chart here we can easily understand how the data is organized for the age column.

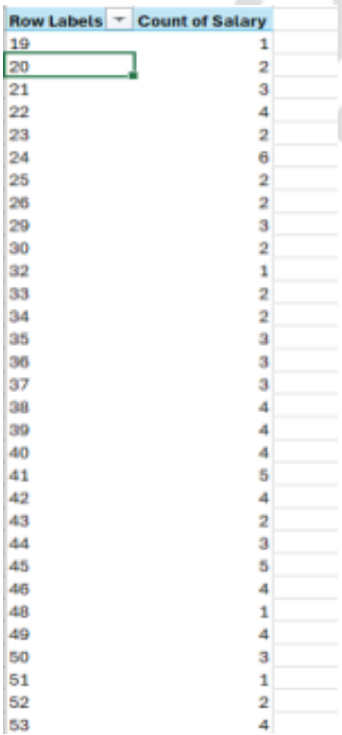
Another problem that we might encounter when we are setting the bins manually is that when the dataset is huge we might not be aware of the range of data that is available in the dataset. In this scenario, we can quickly check for the maximum and minimum value that is available in the dataset using MAX() and MIN function in Excel.
Once we are aware of the maximum and minimum values of the dataset, now we can easily divide the range equally. Further we will use the frequency function followed by inserting column charts to observe the distribution of data.

2. **Using histogram charts:**
   The excel provides an inbuilt property of histogram which helps us to create bins automatically. The histogram chart sets the range of bins by considering the dispersibility of data.



Here in the above image we can see that after applying histogram chart, Excel has automatically divided the “Age” data into ranges.

3. **Using pivot tables:**
   Binning in pivot tables involves grouping continuous data into discrete intervals, or "bins," to summarize and analyze the data more effectively.



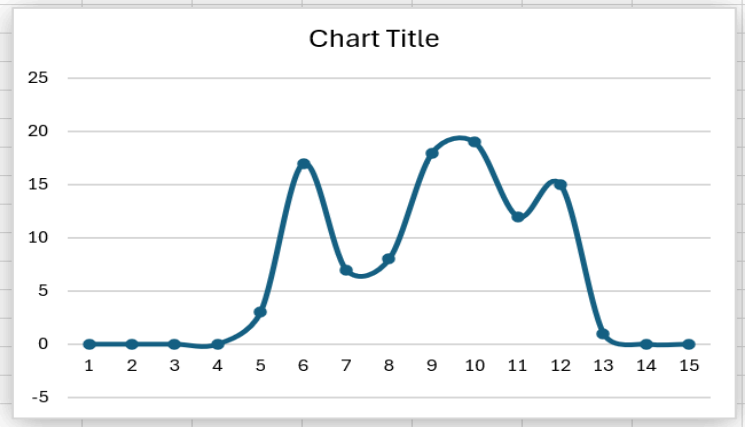
Row Labels	Count of Salary
19-28	22
29-38	23
39-48	32
49-58	23
Grand Total	**100**

d. Creating density plot:
[(Reference Dataset)]

A density plot, also known as a density curve or density trace, is a graphical representation that displays the estimated probability density function (PDF) of a continuous variable. It provides a smooth, continuous visualization of the distribution

of the data, allowing for a detailed examination of its shape, modality, and other characteristics.

Bin (Age)	Frequency
0	0
5	0
10	0
15	0
20	3
25	17
30	7
35	8
40	18
45	19
50	12
55	15
60	1
65	0
	0



When to use histogram vs density plot:

Example -
Imagine you're a teacher looking at the grades of your students on a test. If you want to see how many students scored within specific grade ranges (e.g., 60-69, 70-79), a histogram is perfect. It shows you the frequency of scores in each range, giving you a clear picture of how many students fall into each grade bracket.

Now, if you're interested in understanding the overall performance pattern, like whether grades are mostly clustered around a high, low, or middle point, a density plot is more insightful. It smooths out the grade distribution, showing you peaks where grades are concentrated, helping to identify if most students performed similarly or if there are variations in their performance.

Understand the different aspects involved in categorical univariate analysis:

(Reference Dataset)

Frequency Distribution: Calculate the frequency of each category or group within the categorical variable. This involves counting the number of observations in each category. Using pivot tables, we can easily get an idea of the frequency of the distribution.

Bar Charts: Visualize the frequency distribution of categories using a bar chart. Bar charts represent each category on the x-axis and their corresponding frequencies or proportions on the y-axis.

Pie Charts: Alternatively, you can use pie charts to visualize the proportion of each category relative to the whole. Each category is represented as a slice of the pie, with the size of the slice proportional to its frequency or proportion.

FUN TIME:

This is the part where we will be using the ABC Company dataset to perform the complete univariate 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 univariate-Continuous analysis, we will be acquiring the answers of the following questions:

- ***Which age group of customers are willing to buy the product?***
- ***Which income range of customers will be willing to buy the product?***
- ***What is the maximum amount of sweets and amount of fruits sold?(Range can be identified)***
- ***What is the pattern of the customers buying the products of ABC company with kids?***
- ***What is the pattern of total purchases of the customers?***

Using Univariate-Categorical analysis, we will be acquiring the answers of the following questions:

- ***What is the category of customers who are more likely to buy the product?***
    - ***Marital status***
    - ***Education***
    - ***Age group***

Furthermore, to get a clearer picture of the analysis and deep dive into it we have to understand BIVARIATE ANALYSIS in detail.

Understand the introductory topics of bivariate analysis:

a. What is bivariate analysis?

Bivariate analysis is a statistical technique used to examine the relationship between two variables in a dataset. It involves analyzing the association, correlation, or dependence between these two variables.

There are several methods and techniques used in bivariate analysis, depending on the types of variables (continuous, categorical, or a combination) and the nature of the relationship being investigated.

Different types of bivariate analysis:

i. Continuous-Continuous Bivariate Analysis:

Scatter Plots: Visualize the relationship between two continuous variables.
Correlation Analysis: Measures the strength and direction of the linear relationship between two continuous variables.

ii. Continuous-Categorical Bivariate Analysis:

Box Plots: Create a Box and Whisker Plot in Excel to compare the distribution of a continuous variable across different categories of a categorical variable. You can use the Box and Whisker Plot option in the

Insert Statistic Chart menu. (Only for MS Excel, will not be useful in Spreadsheets.)

iii. **Categorical-Categorical Bivariate Analysis:**

Cross-Tabulation (Contingency Table): Use Excel's PivotTable feature to create a cross-tabulation of two categorical variables. This will show the frequency distribution of each category combination.