

SUMMAR: DATA ANALYSIS-1

SESSION OVERVIEW:

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

- Understand briefly about feature engineering.
- Understand the different aspects of pivot tables.
- Understand advanced features in pivot tables.
- Understand the different aspects of charts in Excel.
- Perform data analysis using pivot tables and charts in Excel.

KEY TOPICS AND EXAMPLES:

1. Understanding on feature engineering: (15 mins)

a. What is feature engineering:

In data analysis, feature engineering refers to the process of selecting, creating, or transforming features (variables or attributes) in a dataset to facilitate analysis, gain insights, or improve model performance. Feature engineering aims to enhance the quality of the data by making it more informative, relevant, and suitable for the analysis tasks at hand. Here's a breakdown of what feature engineering entails in data analysis:

Feature Selection: Identifying and selecting the most relevant features from the dataset based on their importance or contribution to the analysis objectives. This involves evaluating the correlation between features, considering domain knowledge, and using techniques such as statistical tests or feature importance rankings.

Example: Let's assume we are performing a credit risk analysis problem, selecting features such as income, employment status, credit history, and debt-to-income ratio, as they are highly relevant for predicting loan default risk.

Feature Creation: Generating new features from existing ones to capture additional information or improve model performance. This may involve mathematical transformations, aggregations, or deriving features from domain-specific knowledge.

Example: Let's assume we are performing a marketing campaign analysis. Creating a new feature called "customer lifetime value" by aggregating historical purchase data and calculating the total revenue generated by each customer would mean that it's a part of the feature creation process.

Feature Transformation: Modifying the characteristics of features to meet the requirements of the analysis or modeling techniques. Common transformations include scaling features to a common range, normalizing distributions, or applying logarithmic transformations.

Domain-Specific Feature Engineering: Incorporating domain knowledge to create features that capture relevant information specific to the problem domain. This may

involve understanding the underlying processes or relationships in the data and engineering features accordingly.

b. **Example: ABC dataset**

This dataset can be used in terms of understanding feature engineering. We can add columns for accurate analysis or to get a greater exposure of the data which will ease the process of analysis.

Operations in the dataset can include, addition of age column to figure out which age group of customer might be interested in the new product or would churn out.

Secondly, if we add both the columns of “teenhome” and “kidshome” and call it as “childrenhome”, then it would be easier to analyze as we are analyzing aggregately.

2. **Understanding different aspects of pivot tables:**

a. **What are pivot tables?**

Pivot tables are a powerful feature in spreadsheet software, such as Microsoft Excel or Google Sheets, used for summarizing and analyzing large datasets. They allow users to quickly and easily transform raw data into meaningful insights by organizing, summarizing, and aggregating data in a customizable format.

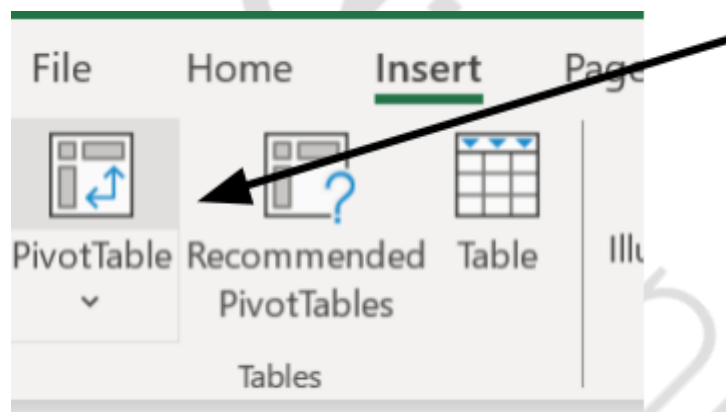


Figure: Represents pivot table in Excel

b. **Why are pivot tables useful?**

Pivot tables offer several benefits that make them invaluable tools for data analysis:

- **Data Summarization:** Pivot tables can efficiently summarize large datasets into manageable and understandable summaries, enabling users to grasp trends, patterns, and relationships within the data.
- **Data Exploration:** Pivot tables provide flexibility in analyzing data from different perspectives. Users can easily rearrange and customize the layout of pivot tables to explore data from various angles and dimensions.

- **Interactivity:** Pivot tables are interactive, allowing users to dynamically filter, sort, and drill down into the data to uncover insights and explore specific details.
- **Ease of Use:** Despite their powerful capabilities, pivot tables are user-friendly and require minimal technical expertise to create and manipulate. They automate many complex data analysis tasks, saving time and effort for users.
- **Visual Representation:** Pivot tables can be accompanied by pivot charts, providing visual representations of the summarized data for enhanced understanding and presentation.

c. **Overview of pivot tables components:**

([Dataset](#) used in explanation of the topic)

- **Rows:** Think of rows as the vertical arrangement of your data. They organize your information based on categories or groups. For example, if you're analyzing sales data, the rows could represent different product categories or customer segments. Each row shows the data related to that specific category.
- **Columns:** Columns are like the horizontal organization of your data. They help you group your information across different criteria. For instance, in the sales data example, columns could represent different time periods like months or quarters. Each column displays data for a specific time period.
- **Values:** Values are the actual numbers you're interested in analyzing. These could be sales revenue, quantities sold, or any other numerical data. Values are summarized or aggregated based on the rows and columns they intersect with. For example, the total sales revenue for a particular product category in a specific time period.
- **Filters:** Filters allow you to narrow down the data shown in your pivot table based on specific conditions. You can use filters to focus on particular categories, time periods, or any other criteria of interest. They help you analyze subsets of your data without changing the overall structure of the pivot table.

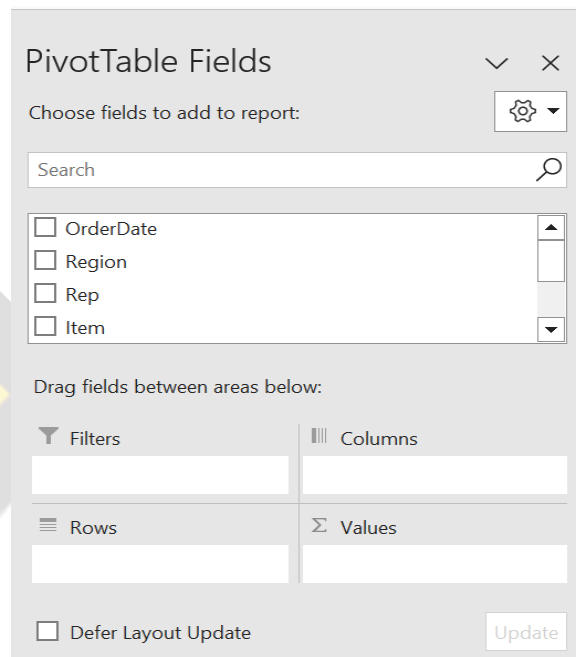


Figure: Represents the dialogue box associated to pivot table in Excel

d. Basic terminology:

- **Fields:** Fields are like the different types of information you have in your dataset. In simpler terms, we can say fields are the columns of your table. For example, in sales data, fields could include things like product names, sales dates, or customer names. Each field represents a specific piece of data you can use to analyze your information.
- **Items:** Items are the individual options or categories within each field. In simpler terms, we can say unique values present in a column. For instance, if you have a field for product names, the items would be the actual product names themselves. They're the specific values or categories that you can choose from when creating your pivot table.
- **Report Layouts:** Report layouts are the different ways you can arrange and display your data in the pivot table. Think of them as the formats or styles you can choose from. Each layout option, like compact, outline, or tabular, changes how your data is organized and presented in the pivot table. It's like choosing the design of your table to best fit your needs.

3. Understanding different features in pivot tables:

a. Applying Filters in Pivot Tables:

(Reference [Dataset](#))

a. Uses of filter in pivot table:

- Narrow down the data displayed to focus on specific subsets.

- Exclude certain categories or values from analysis.
- Filter by value criteria, such as greater than or less than.
- Quickly switch between different views of your data for deeper analysis.
- Customize your analysis based on specific requirements or questions.

b. Steps that should be followed:

Create the Pivot Table:

- Select your data range.
- Go to the "Insert" tab on the Excel ribbon.
- Click on "PivotTable" and select where you want the pivot table to be placed (e.g., a new worksheet).
- Click "OK."

Add Fields to the Pivot Table:

- In the PivotTable Field List pane, drag the fields you want to analyze into the Rows, Columns, Values, or Filters areas.
- For example, you might drag "Product" to Rows and "Sales" to Values to analyze sales by product.

Apply Filters:

- To apply a filter, click the dropdown arrow next to the field you want to filter in the Rows, Columns, or Filters area of the pivot table.
- Uncheck the box next to any items you want to exclude from the analysis, or use the search box to find specific items.
- Click "OK" or "Apply" to apply the filter.

Filtering by Values:

- In addition to filtering by field items, you can also filter by values.
- Click the dropdown arrow next to the field you want to filter in the Values area.
- Choose "Filter by Value" and specify the criteria for filtering, such as greater than, less than, equal to, etc.

Clear Filters:

- To remove filters, click the dropdown arrow next to the field you want to clear the filter for.
- Select "Clear Filter" or "Clear Filter from [Field Name]" to remove the filter.

c. Example:

This [dataset](#) will be used to explain the filters in the pivot table. Here we have a dataset of office supplies which contains sales data of office supplies in different regions.

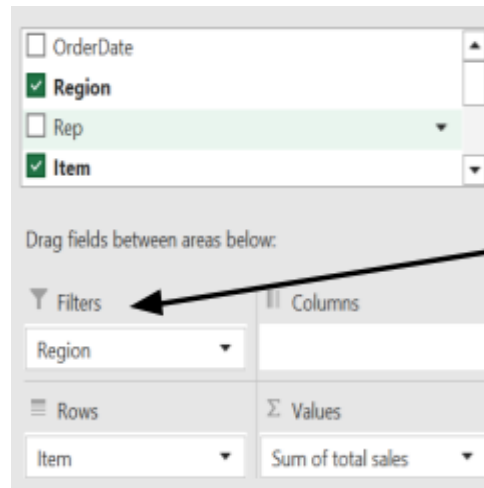
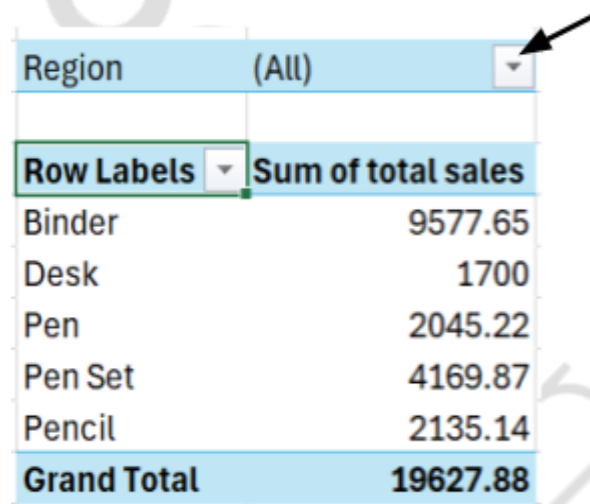


Figure: Represents the filter in pivot table in Excel

- This is the dialogue box which opens up when we apply a pivot table in the data. The arrow indicates the filter which helps us to filter out the required data. Also, we can notice that we have put “region” as one of the filters.



Region	(All)
Row Labels	Sum of total sales
Binder	9577.65
Desk	1700
Pen	2045.22
Pen Set	4169.87
Pencil	2135.14
Grand Total	19627.88

Figure: Represents the application of filter in pivot table

- Once we have added “region” as one of the filters in the PivotTable Field List pane a dialogue box appears as the above image.
- The arrow in the right corner indicates that a filter has been created in the following pivot table. The presence image shows the total sales of all regions in each item category.

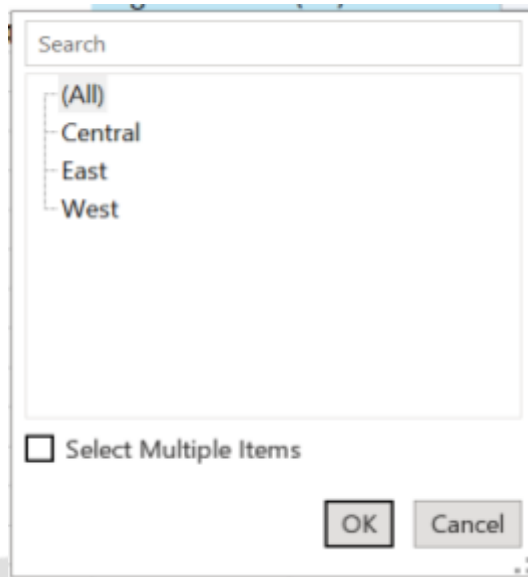


Figure: Represents the filter dropdown box in pivot table

- Once we click the button in the right most corner, this dropdown opens which helps us to select the criteria for which we want the values.

b. Drilldown in pivot table:

[\(Reference Dataset\)](#)

i. Uses of drilldown:

1. Explore detailed data underlying summary values.
2. Investigate specific data points by expanding or collapsing hierarchical levels.
3. Understand the composition of summary values by viewing underlying details.
4. Analyze trends or patterns within specific categories or groupings.
5. Gain insights into the factors driving summary values by examining individual data entries.

ii. Steps used to apply drilldown:

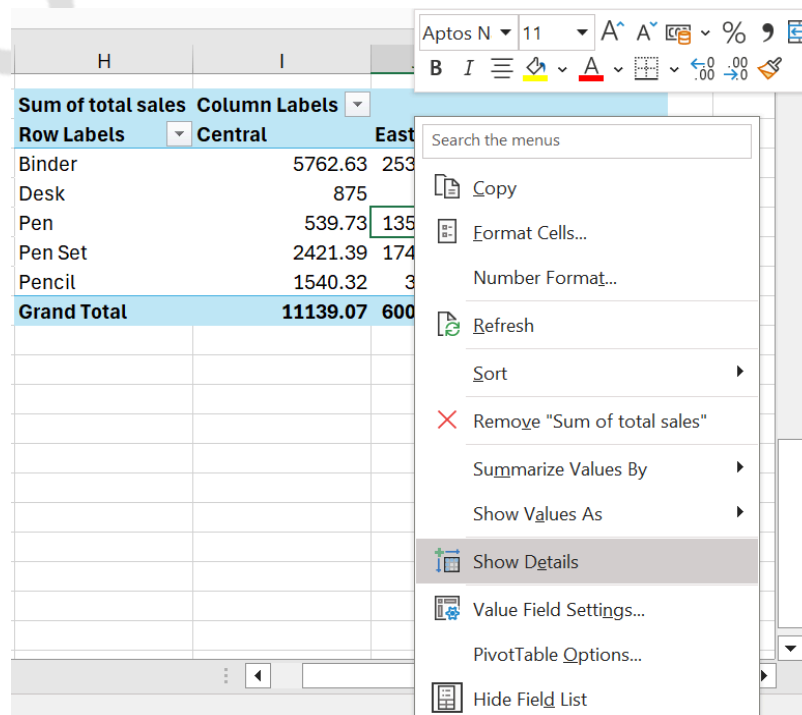
- Click on any cell within the pivot table that contains summarized data you want to drill down into.
- Right-click on the cell.
- In the context menu that appears, select "Show Details" or "Drill Down", depending on your Excel version.
- Excel will automatically generate a new worksheet containing the detailed data corresponding to the selected cell.
- Review the detailed data to explore the underlying information.
- To return to the pivot table view, simply navigate back to the original worksheet containing the pivot table.

iii. Example:

The dataset that will be used to explain the drill down feature in pivot table.

Sum of total sales		Column Labels			
Row Labels	Central	East	West	Grand Total	
Binder	5762.63	2535.66	1279.36	9577.65	
Desk	875		825	1700	
Pen	539.73	1354.25	151.24	2045.22	
Pen Set	2421.39	1748.48		4169.87	
Pencil	1540.32	363.7	231.12	2135.14	
Grand Total	11139.07	6002.09	2486.72	19627.88	

- Suppose we want to apply the drilldown feature in the following cell of the following pivot table.



The screenshot shows an Excel interface with a pivot table. The pivot table is titled 'Sum of total sales' and has 'Row Labels' and 'Column Labels'. The 'Row Labels' are 'Binder', 'Desk', 'Pen', 'Pen Set', and 'Pencil'. The 'Column Labels' are 'Central', 'East', and 'West'. The 'Grand Total' row is highlighted. The cell containing '1354.25' (Pen, East) is selected. A right-click context menu is open over this cell, showing options like 'Copy', 'Format Cells...', 'Number Format...', 'Refresh', 'Sort', 'Remove "Sum of total sales"', 'Summarize Values By', 'Show Values As', 'Show Details' (which is highlighted), 'Value Field Settings...', 'PivotTable Options...', and 'Hide Field List'.

- We select the cell and right click on the cell which leads us to a dialogue box similar to the above mentioned image. Then we click on the “show details” mentioned in the dialogue box.

OrderDate	Region	Rep	Item	Units	Unit Price	total sales
27-04-2015	East	Nick	Pen	96	4.99	479.04
08-11-2014	East	Susan	Pen	15	19.99	299.85
22-10-2014	East	Richard	Pen	64	8.99	575.36

- After we click on the “show details” a separate sheet opens up which contains the elaborated version of the data in the pivot table.

OrderDate	Region	Rep	Item	Units	Unit Price	total sales
27-04-2015	East	Nick	Pen	96	4.99	479.04
08-11-2014	East	Susan	Pen	15	19.99	299.85
22-10-2014	East	Richard	Pen	64	8.99	575.36

Formatting | Charts | Totals | Tables | Sparklines

Data Bars
 Color Scale
 Icon Set
 Greater Than
 Top 10%
 Clear Format

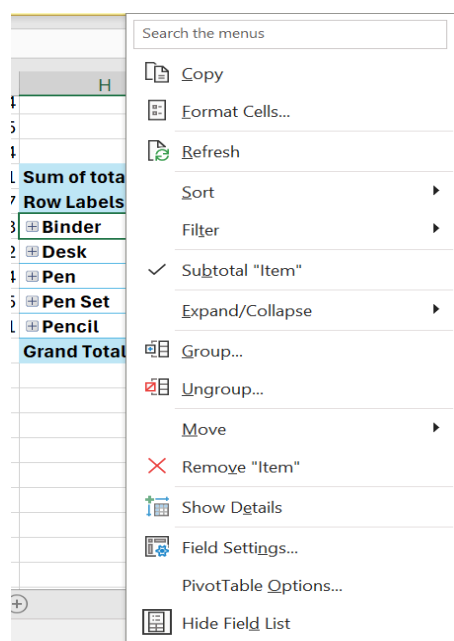
Conditional Formatting uses rules to highlight interesting data.

- This also helps us apply different formatting, charts, to immediately analyze the data and get insights from the data.

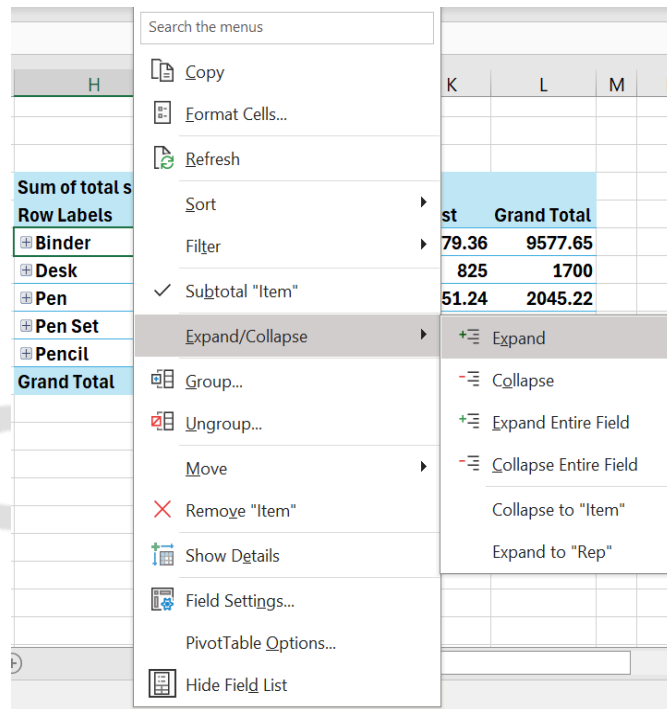
Alternative method:

Sum of total sales	Column Labels			
Row Labels	Central	East	West	Grand Total
Binder	5762.63	2535.66	1279.36	9577.65
Desk	875		825	1700
Pen	539.73	1354.25	151.24	2045.22
Pen Set	2421.39	1748.48		4169.87
Pencil	1540.32	363.7	231.12	2135.14
Grand Total	11139.07	6002.09	2486.72	19627.88

- Suppose we want to apply the drilldown feature in the following selected cell in the following pivot table.



- We select the cell and right click on the cell which leads us to a dialogue box similar to the above mentioned image. Then we click on the “Expand/Collapse” mentioned in the dialogue box.



- After clicking on the Expand/Collapse option, a separate dialogue box opens where if we click on the expand then it will work as drill down and collapse as drill up.

Sum of total sales	Column Labels			
Row Labels	Central	East	West	Grand Total
Binder	5762.63	2535.66	1279.36	9577.65
Alex	1933.95			1933.95
Bill	1132.74			1132.74
James			139.93	139.93
Matthew	999.5			999.5
Morgan	251.72			251.72
Nick		57.71		57.71
Rachel	139.72			139.72
Richard		858.76		858.76
Smith	1305			1305
Susan		1619.19		1619.19
Thomas			1139.43	1139.43
Desk	875		825	1700
Pen	539.73	1354.25	151.24	2045.22
Pen Set	2421.39	1748.48		4169.87
Pencil	1540.32	363.7	231.12	2135.14
Grand Total	11139.07	6002.09	2486.72	19627.88

- This is how it looks when drilldown is applied to the pivot table.

Alternative method:

Sum of total sales		Column Labels			
Row Labels		Central	East	West	Grand Total
Binder		5762.63	2535.66	1279.36	9577.65
Alex		1933.95			1933.95
Bill		1132.74			1132.74
James				139.93	139.93
Matthew		999.5			999.5
Morgan		251.72			251.72
Nick			57.71		57.71
Rachel		139.72			139.72
Richard			858.76		858.76
Smith		1305			1305
Susan			1619.19		1619.19
Thomas				1139.43	1139.43
Desk		875		825	1700
Pen		539.73	1354.25	151.24	2045.22
Pen Set		2421.39	1748.48		4169.87
Pencil		1540.32	363.7	231.12	2135.14
Grand Total		11139.07	6002.09	2486.72	19627.88

- In the left corner of the pivot table there is also a “+” and “-” sign which will help us get the detailed summary of the data.

c. Aggregating data in pivot table:

[\(Reference Dataset\)](#)

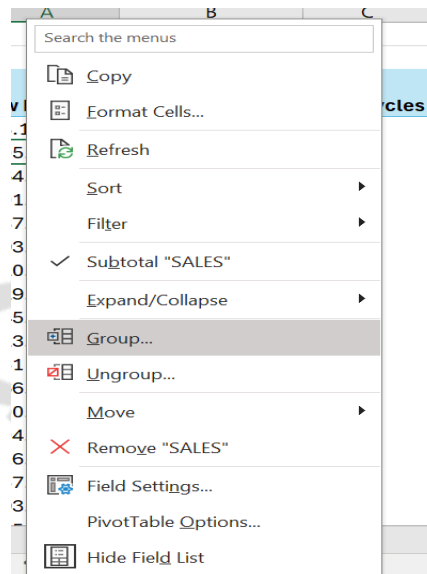
- Sum: Adds up the values in the selected field.
- Count: Counts the number of items in the selected field.
- Average: Calculates the average of the values in the selected field.
- Min: Finds the smallest value in the selected field.
- Max: Finds the largest value in the selected field.
- Product: Multiplies all the values in the selected field.
- Count Numbers: Counts only the numeric values in the selected field.
- Standard Deviation: Calculates the standard deviation of the values in the selected field.
- Variance: Calculates the variance of the values in the selected field.
- % of Column Total: Calculates the percentage of each value relative to the total in the column.
- % of Row Total: Calculates the percentage of each value relative to the total in the row.
- % of Grand Total: Calculates the percentage of each value relative to the grand total.

d. Grouping of numbers and dates in pivot table:

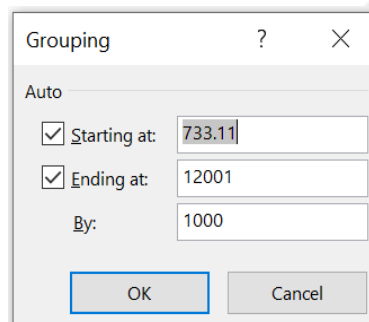
Grouping in pivot tables is a feature that allows you to organize and categorize data into logical groups based on specific criteria. It is particularly useful for summarizing and analyzing large datasets.

Steps: ([Dataset](#))

- i. First, create a pivot table with sales in the rows and “productline” in the columns.
- ii. Go to the pivot table and select a cell and right click on the cell. Once you right click, a dialogue box opens up then selects “group”.



- iii. Once you click on a group another dialogue box opens up which specifies about the maximum value, minimum value and the range according to which you want to group the data. You can make changes in the dialogue box accordingly.



- iv. Accordingly, the grouping will be completed and then you can add the desired values.

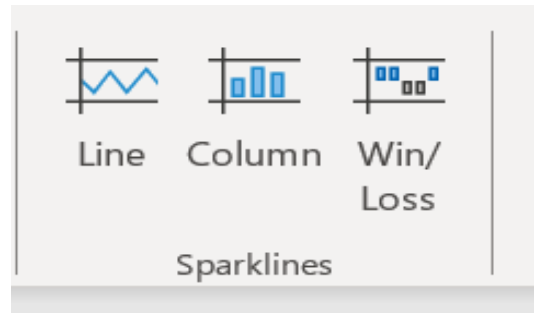
e. Sparklines in pivot table:

([Reference dataset](#))

Sparklines in pivot tables are miniature charts that provide visual representations of data trends within individual cells. They are a useful tool for quickly analyzing patterns and trends in large datasets without the need for creating separate charts.

Steps:

- Select the cell where you want to insert the sparkline within the pivot table.



- Go to the "Insert" tab on the Excel ribbon.
- Click on the "Sparklines" dropdown menu in the "Sparklines" group.
- Choose the type of sparkline you want to insert, such as Line, Column, or Win/Loss.
- In the "Data Range" dialog box, specify the range of data you want to use for the sparkline. This can be a range of cells or a data range reference within the pivot table.

Count of QUANTITYORDERED	Column Labels												
Row Labels	1	2	3	4	5	6	7	8	9	10	11	12	Grand Total
Classic Cars	14	13	12	6	17	7	10	10	11	19	39	9	167
Motorcycles	6	12	4	10	10	4	7	9	4	11	21	6	104
Trucks and Buses	2	2	3		4	1	2	1	2	3	6	2	28
Vintage Cars	6	2	6	2	4	2	2	2	4	6	12	4	52
Grand Total	28	29	25	18	35	14	21	22	21	39	78	21	351

If the above steps are followed then we can create the trendlines as visible in the image above.

f. Activation of DISTINCT COUNT in Excel:

[\(Reference dataset\)](#)

To activate and get a distinct count in a PivotTable in Excel, follow these steps:

Create a PivotTable:

- Select any cell in your data range.
- Go to the "Insert" tab, and click on "PivotTable".
- In the "Create PivotTable" dialog box, choose the data range and click "OK".

Place the field(s) you want to count distinct values for in the "Rows" or "Columns" area:

- iv. In the PivotTable Fields pane, locate the field(s) you want to count distinct values for.
- v. Drag and drop the field(s) into the "Rows" or "Columns" area.

Add the field(s) to the "Values" area:

- vi. In the PivotTable Fields pane, locate the same field(s) you placed in the "Rows" or "Columns" area.
- vii. Drag and drop the field(s) into the "Values" area.

Change the summarization function to "Distinct Count":

- viii. In the "Values" area of the PivotTable, right-click on the field you just added.
- ix. Select "Value Field Settings" from the context menu.
- x. In the "Value Field Settings" dialog box, change the "Summarize Values By" option to "Distinct Count".
- xi. Click "OK" to apply the changes.

The PivotTable will now display the distinct count of values for the selected field(s) in the "Values" area, grouped by the fields in the "Rows" or "Columns" area.

For example, if you have a dataset with columns for "Product", "Region", and "Sales", you can place "Product" in the "Rows" area, drag "Region" to the "Values" area, and change the summarization function to "Distinct Count". This will give you the count of distinct regions for each product.

Note: The "Distinct Count" option is available in Excel 2007 and later versions. If you're using an earlier version of Excel, you may need to create a separate PivotTable with unique values and then use the COUNT function to count them.

g. **Slicers in pivot table:**

([Reference dataset](#))

i. Uses of slicer in pivot table:

1. Interactive filtering for quick data exploration.
2. Visual representation of filtering options.
3. Multi-selection capability for complex analysis.
4. Cross-filtering across multiple visualizations.
5. Easy reset functionality for seamless user experience.

ii. Steps involved in slicers in pivot table:

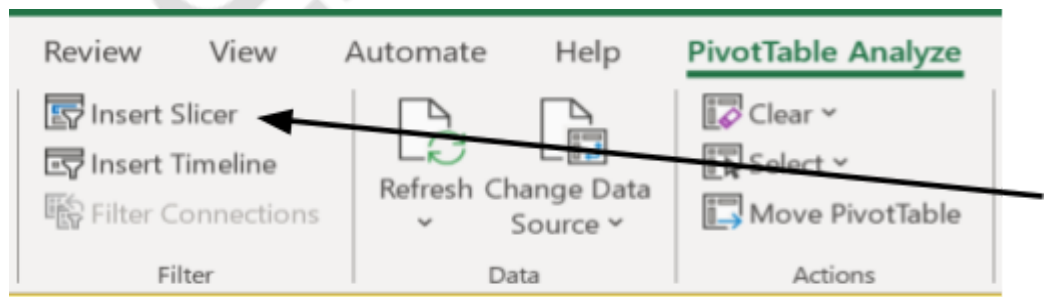
- Click anywhere inside the pivot table that you want to add a slicer to.
- Go to the "PivotTable Analyze" or "Analyze" tab in the Excel ribbon.

- In the "Filter" group, click on "Insert Slicer."
- In the "Insert Slicers" dialog box that appears, check the box next to each field you want to create a slicer for.
- Click "OK."
- Excel will insert slicers for each selected field on the worksheet.
- Arrange the slicers as desired on the worksheet.
- To filter the pivot table data, simply click on the desired slicer button(s) corresponding to the filter criteria you want to apply.

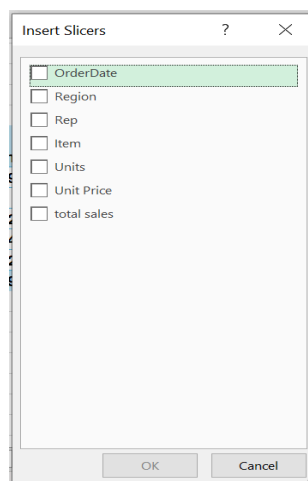
iii. Example:

Sum of total sales		Column Labels			
Row Labels		Central	East	West	Grand Total
Binder		5762.63	2535.66	1279.36	9577.65
Desk		875		825	1700
Pen		539.73	1354.25	151.24	2045.22
Pen Set		2421.39	1748.48		4169.87
Pencil		1540.32	363.7	231.12	2135.14
Grand Total		11139.07	6002.09	2486.72	19627.88

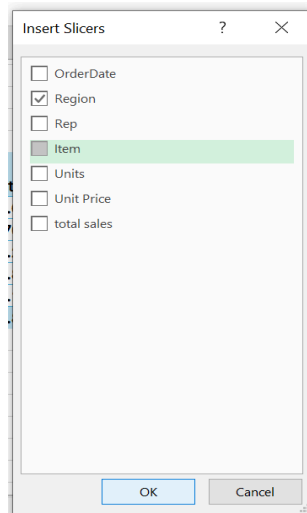
- The above image shows the pivot table on which we will be applying slicer.



- In the PivotTable Analyze in the Excel ribbon, we can find an insert slicer.



- Once we click on "Insert slicer" a dialogue box will appear from which we can choose on which column we want to create a slicer.



- Once we select the desired check box from the insert slicers dialogue box, click OK.

Sum of total sales		Column Labels			
Row Labels		Central	East	West	Grand Total
Binder		5762.63	2535.66	1279.36	9577.65
Desk		875		825	1700
Pen		539.73	1354.25	151.24	2045.22
Pen Set		2421.39	1748.48		4169.87
Pencil		1540.32	363.7	231.12	2135.14
Grand Total		11139.07	6002.09	2486.72	19627.88

- So, here we have created a slicer for the “Region column” which provides us the three different regions that are available in the dataset.

Item	Region
Binder	Central
Desk	East
Pen	West
Pen Set	
Pencil	

- Similarly we can create multiple slicers for multiple columns or variables.

Sum of total sales		
Row Labels	Column Labels	
	Central	Grand Total
Desk	875	875
Grand Total	875	875

Region

Central

West

East

Item

Binder

Desk

Pen

Pen Set

Pencil

- This is an example where we wanted the total sales of desk in the central region which is visible in the pivot table.

h. Automatic adjustments when source data changes:

[\(Reference dataset\)](#)

1. When the changes are within the pivot table range:
 - a. We can simply refresh the pivot table once the source data has been updated.
2. When the changes are beyond the pivot table range:
 - a. We can create the original data into table format in Excel.
 - b. You can add new data in the source table.
 - c. Then, you can go to the pivot table that has been created and go to PivotTable Analyze.
 - d. Once you click PivotTable Analyze, we can see an option called change data source.
 - e. A dialogue box appears where we have to select the table with the table name.
 - f. Once the whole table is selected, click Ok.
 - g. The desired changes will be made in the pivot table as well.

i. GETPIVOTDATA in pivot tables:

[Reference dataset](#)

- **Dynamic Reporting:** GETPIVOTDATA allows you to create dynamic reports that automatically update when the underlying pivot table data changes. You can reference

specific cells within the pivot table and extract the data you need for your reports.

- **Customized Analysis:** With GETPIVOTDATA, you can retrieve specific data points from your pivot table and perform customized analysis. For example, you can extract sales figures for a particular product or region, or calculate the average order size for a specific customer segment.
- **Linked Formulas:** GETPIVOTDATA enables you to link formulas to pivot table data, making it easier to perform calculations and analysis. You can reference GETPIVOTDATA formulas in other cells or worksheets to create dynamic reports and dashboards.
- **Error Handling:** GETPIVOTDATA automatically adjusts to changes in the structure of the pivot table, ensuring accurate data retrieval even if rows or columns are added or removed. This helps prevent errors in your reports and ensures data integrity.
- **Advanced Analysis:** GETPIVOTDATA supports complex criteria and conditions, allowing you to extract data based on multiple criteria and logical operators. This flexibility enables you to perform advanced analysis and gain deeper insights from your pivot table data.

Drawbacks of pivot tables in Excel:

- **Limited adjustment of data:** Once a pivot table is created, it does not automatically update when the source data changes. The user must refresh it manually, which can be a limitation in dynamic analysis environments. If we have added/deleted columns or have made changes to the data, it does not get reflected in pivot outputs, we should ensure changes in the data are properly reflected.
- **Limited Formatting Options:** Pivot tables have limited formatting options compared to other visualization tools, which may restrict customization and presentation options.
- **Inflexibility:** Pivot tables require data to be in a structured format. If the data is unclean or poorly organized, significant preprocessing may be necessary to use pivot tables effectively. Ex: if your table has a blank column header, pivot is unable to analyze that.
- **Calculation Limitations:** Pivot tables may not support all types of calculations or advanced analytical functions, limiting their capability for certain types of analysis.
- **Compatibility Issues:** Pivot tables may not be compatible with older versions of Excel or other spreadsheet software, making it challenging to share or collaborate on data analysis projects.

Using this [dataset](#) you can play around with the different features of pivot tables.



codingninjas