Intro to Data Science
Descriptive Analytics for Business

1   Key Metrics Analysis

Key Metrics or Key Performance Indicators (KPIs) are the main method that we use to measure the performance of a company, division, department or team.   Key Metrics provide an overall understanding of how the business is performing. Key Metrics usually include the following types of questions:

a)   What is the Total Sales?

b)   What is the Total Profit?

c)   What is the Profit Ratio?

d)   How many transactions were there processed?

e)   What was the average Sales amount?

f)   What was the highest Sales amount?

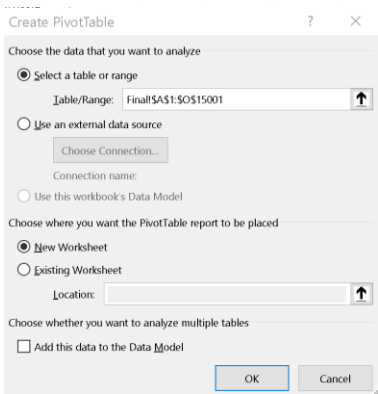g)   What was the lowest Sales amount?

File to use: `SalesData-v1.xlsx`

Create a Pivot table from the current range of data.
Make sure you have selected a cell in the range.
Insert -> Pivot table
The Create Pivot table should automatically detect the complete range of cells with valid values.
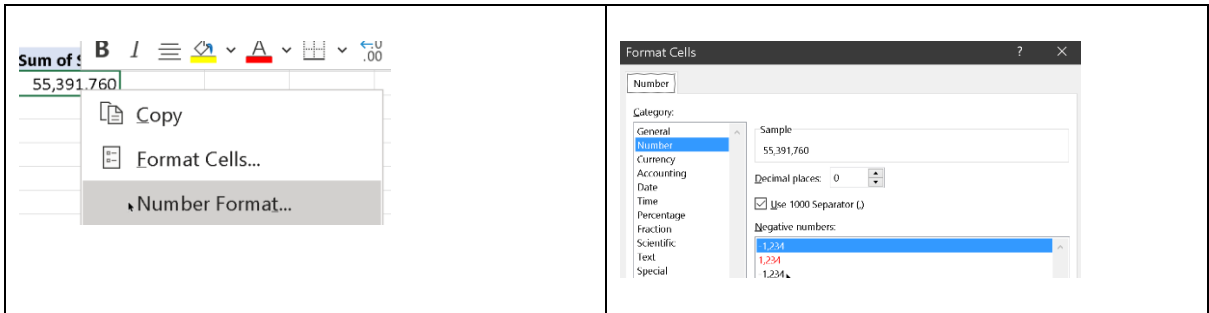Make sure you create the Pivot table in a new worksheet.

You can change the order of the Pivot Table fields as they appear in the main list: its sometimes useful to sort them in alphabetical order if you have many fields.
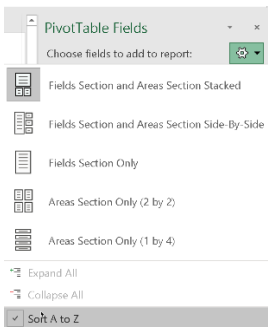
Select the Sales field to get the Sum of Sales, then perform a Number Format to make it easier to view.

Repeat this with the Profit field to get the Sum of Profit, with a similar Number Format as well.

We can now create a Calculated Field to obtain the Profit Ratio.
With at least once cell selected in the pivot table, select Pivot Table Analyze from the top main menu, select Fields, Items and Sets
Next insert a Calculated Field to obtain the Profit Ratio, and format the result accordingly

Sum of Sales	Sum of Profit	Sum of Profit Ratio
55,391,760	35,589,375	64.3%

To count the number of transactions, we would typically use a unique Sales ID, which we don't have here in this dataset. We can initially use the Sales Field to start with (drag and drop into the Values area to obtain a new Sum of Sales 2).

Then select this cell and from the context menu, change it to Summarize Values by Count to get the actual number of transactions (which we can also double confirm from the source dataset)
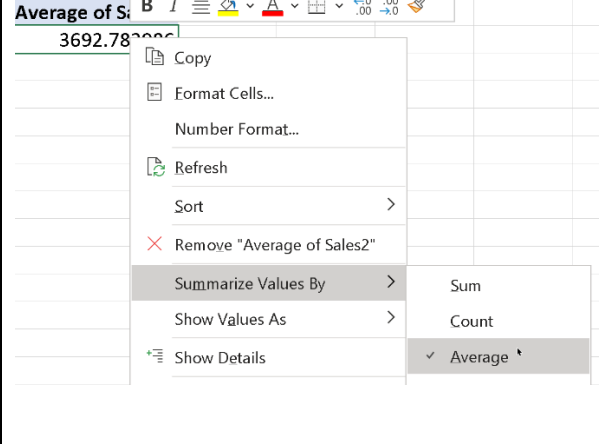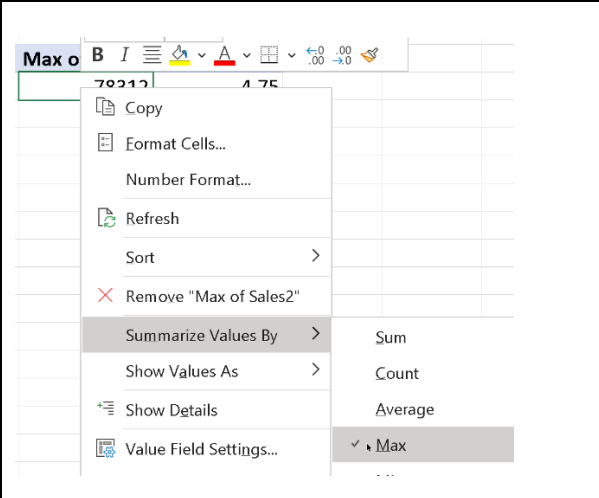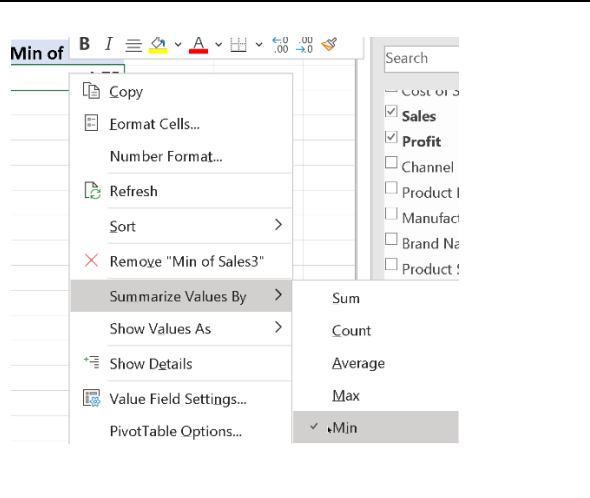

To get the average Sales, use the Sales field again in the same manner (drag and drop into the Values area to obtain a new Sum of Sales 2).

Then select this cell and from the context menu, change it to Summarize Values by Average to get the average Sales value, then format it accordingly.

Count of Sales2	Average of Sales2
15000	3,693

Add two more Sales Fields into the Values area in a similar manner as previously, and then perform a Summarize Values by Min and Max for these two new fields and format them appropriately.

Now that we have all our numbers, you can double click on any (or all) of the field headers to change the field names to something more appropriate if you wish.

We now have our key metrics listed out in a row:

Sum of Sales	Sum of Profit	Sum of Profit Ratio	Num Transactions	Average of Sales	Highest Sales	Lowest Sales
55,391,760	35,589,375	64.3%	15000	3,693	78312	5

You can transpose the rows and columns for a better layout:

To zoom in and get more details on the metrics based on a certain category (for e.g. country), we can filter the table to decide which subset of records to apply the various aggregation operations we have performed so far on.

For this purpose, we can use a slicer. Go to Pivot Table Analyze -> Insert Slicer, and select Country. You can then select particular countries from the slicer drop down to get the metric values only for sales records related to that particular country.

With the slicer selected and then selecting the Slicer option in the main menu, you can fine tune the slicer by adjusting the number of columns, as well as their height and width (you can just click and drag on the slicer itself to manually change this) to make them easier to work with:
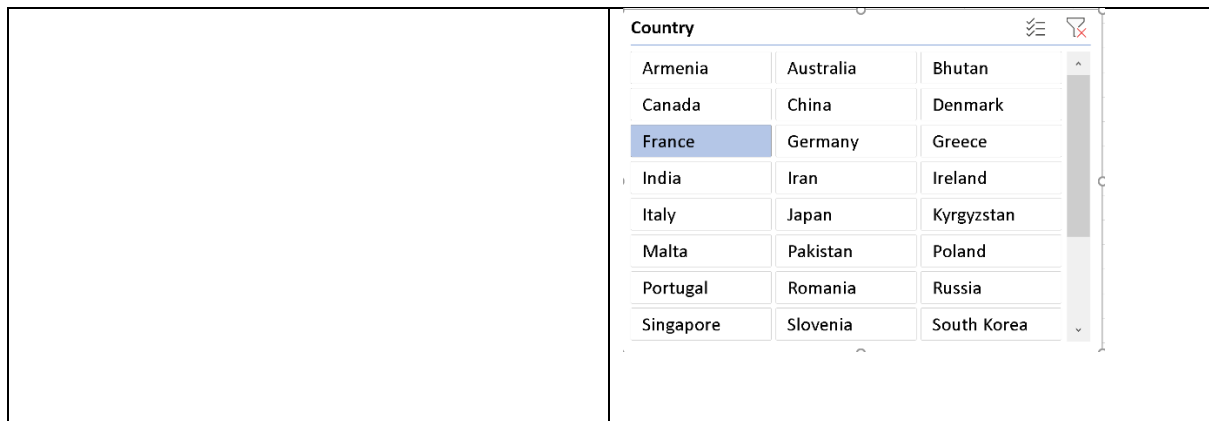
You can use the Ctrl+Left mouse click to select multiple countries at the same time, so the aggregation operations will apply to all these countries. You can then select the clear Filter button to clear the filter and make the aggregation apply to the entire dataset.
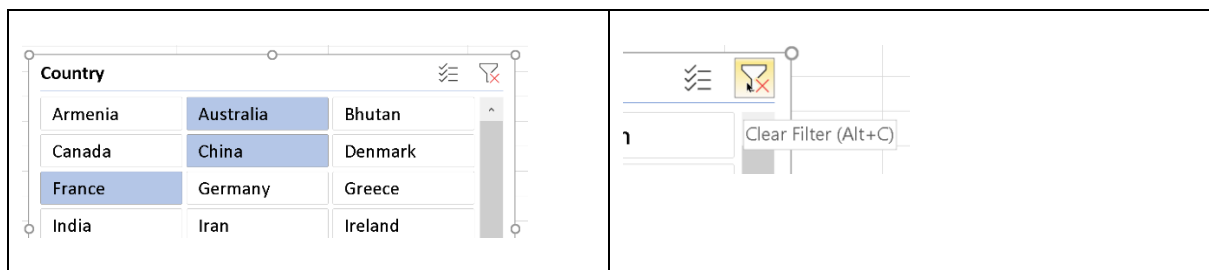
You can now also insert a Timeline and decide which particular period (Years, Quarters, Months) to examine.
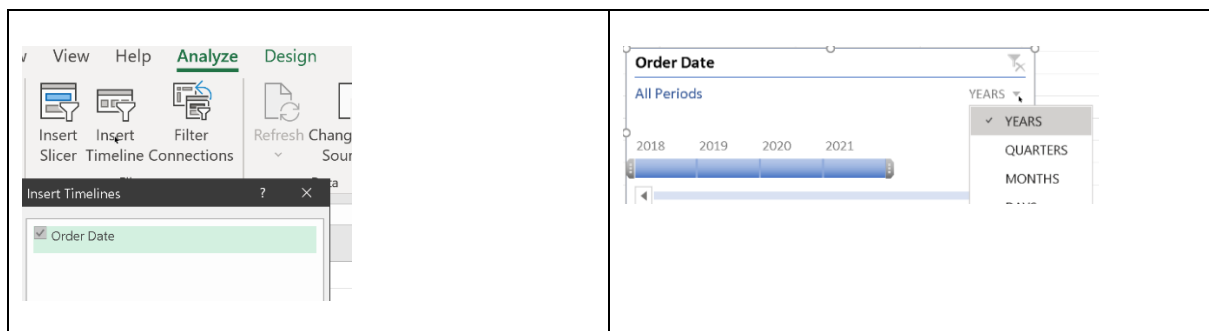
You can select a particular period or a range of periods (Shift-Click):
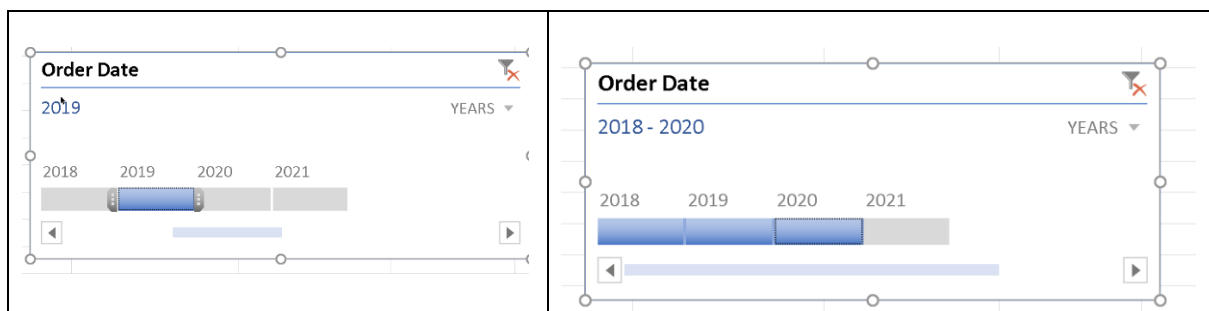
You can also select items from both filters to combine their filtering action together, for e.g. to perform aggregation operations on a specific country (or group of countries) over a specific period of time.
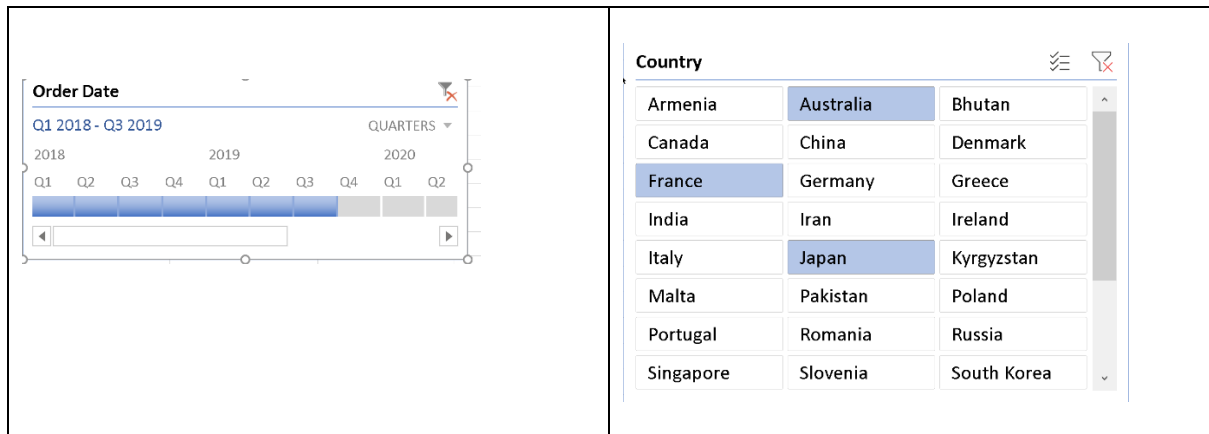
You can introduce another slicer for Region, and notice that selection of items in either slicer for Country or Region will influence the other: based on which countries are included in which region. To see this properly in effect, you will have to clear the filter in either one or both of the slicers.
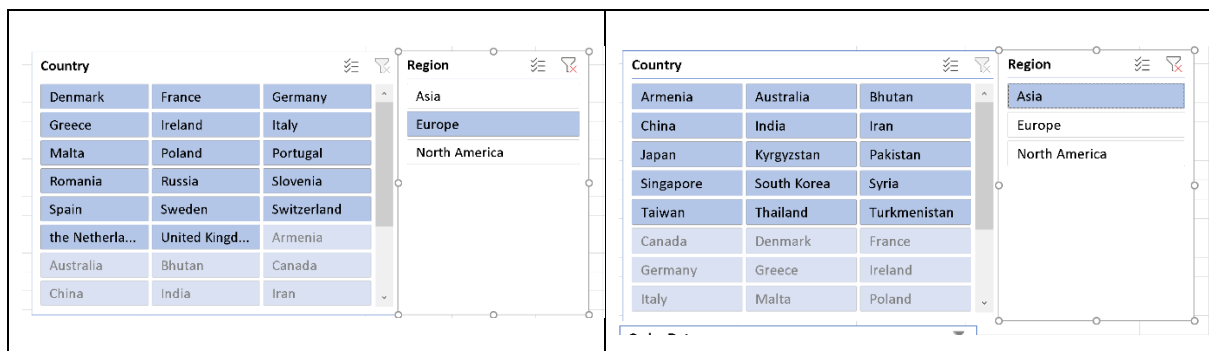
You can also set the Slicer Settings to only clearly show the countries within a particular region, and not just grey out the countries that are not in that region (to make the analysis even more clearer):
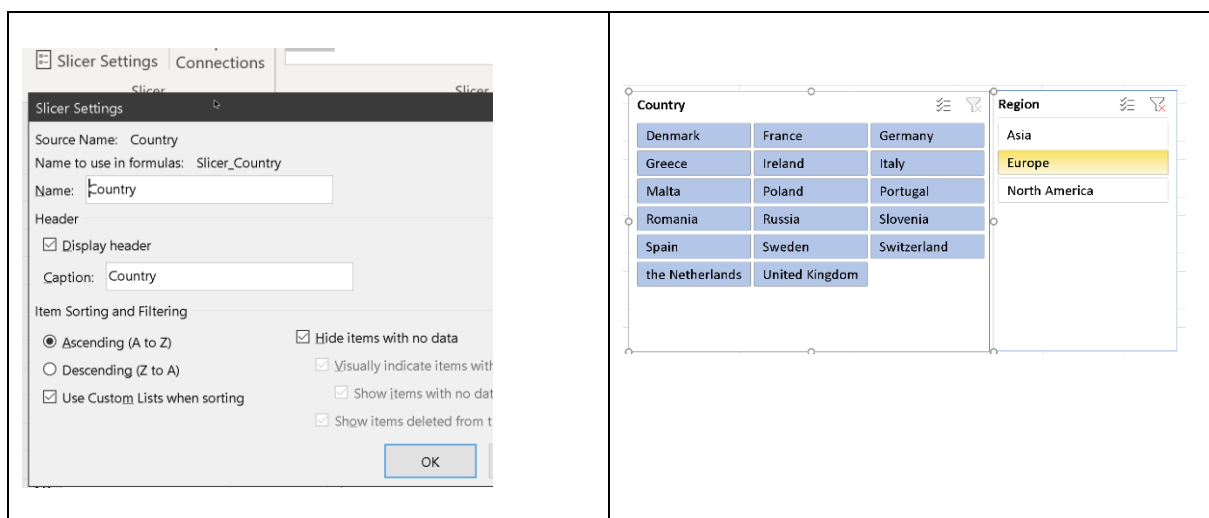
1.1   Practical Exercise for Key Metrics Analysis
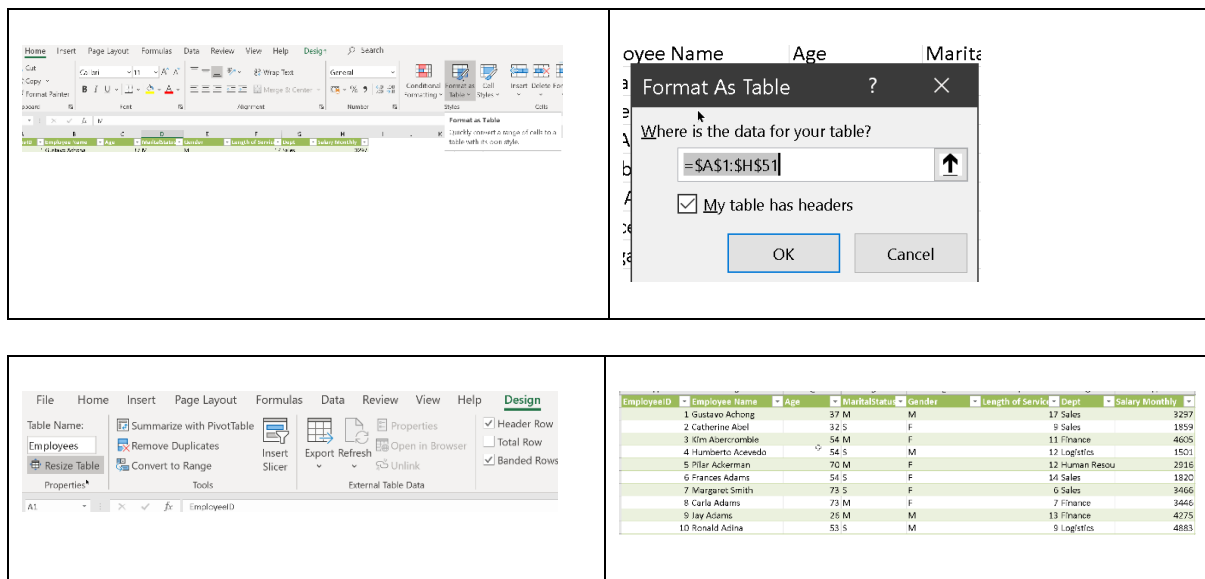
File to use: `Employee-v1.xlsx`

Analytical activities to perform:

1.   How many employees are in John's company?

2.   What is the average age of the employees?

3.   What is the highest age at the company?

4.   What is the lowest age at the company?

5.   What is the average Length of Service?

6.   What is the longest Length of Service?

Drill down further into this information based on the following categories: Dept and Gender

You can initially format the range of cells of the original data set as a table to make it easier to work with, and to create a Pivot Table. To do this, select a cell in the data range, go to Home -> Format as Table, format it and give it an appropriate name.

Once done, you can generate a Pivot Table in the usual way, but this time referencing the Table.

2   Comparison Analysis

Comparison analysis is probably one of the most popular forms of analysis you're going to do. The focus is finding the magnitude of difference between values for different fields. Visualizing your data set makes it much easier to appreciate this as opposed to directly examining the values themselves.
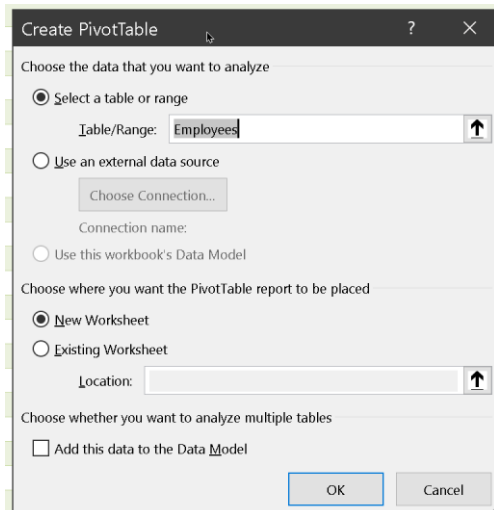
Sample analysis

  a)   compare the sales and profit by year
  b)   compare the sales by manufacturer
  c)   compare the profit by product categories
  d)   compare the sales by product category and by channel
  e)   compare the profit by product subcategory and review that by different countries.
  f)   compare the sales and profit by the countries, and review that by different product sub categories

File to use: `SalesData-v2.xlsx`

Start off again by generating a Pivot Table in the usual manner that references the Sales table.
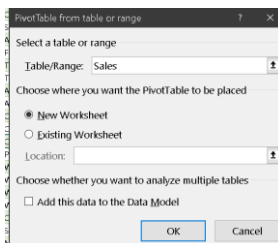
In this new Pivot table, obtain the Sum of Sales and Profit by Years (Rows) and format the cells with an appropriate numeric format (select all the cells -> Format Cells)

We can also sort on the Sum of Sales or Profit (either Smallest to Largest or vice versa)

The sorting can also be done via the Sort option in the Data tab

Sort in descending order for Sum of Sales, and visualize this via a Pivot Chart (choose Clustered Column)

You can format what fields to be shown in the chart, including hiding everything:
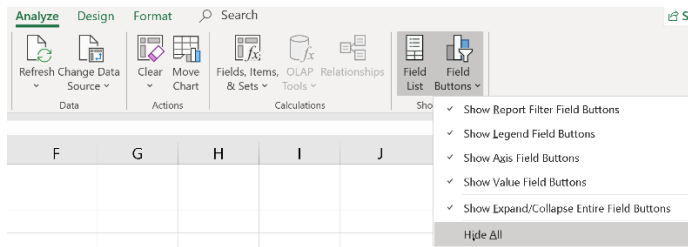
You can also change the Chart Elements appropriately, for e.g. add in Title and moving the legend to the Top.
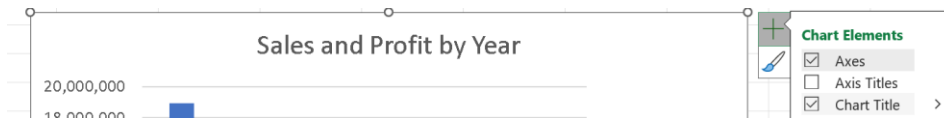
We can now create a new Pivot Table and repeat the earlier process to get the Sum of Sales by Manufacturer (Rows) and format the cells with an appropriate numeric format.

Row Labels	Sum of Sales
A. Datum Corporation	4404982.008
Adventure Works	7467703.692
Contoso, Ltd	9996750.019
Fabrikam, Inc.	12054281.48
Litware, Inc.	2134836.834
Northwind Traders	136571.9663
Proseware, Inc.	5996829.028
Southridge Video	3112730.481
The Phone Company	4386164.26
Wide World Importers	5700910.017
Grand Total	55391759.79

We can now insert another Clustered Column chart, and format it in the same way as we did previously.
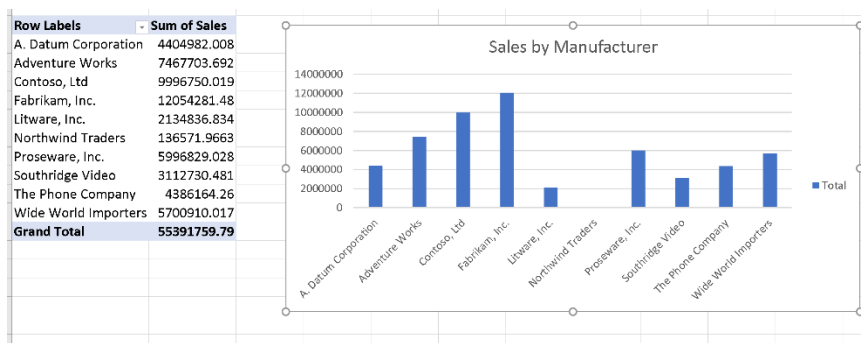
We can also opt to change the Chart Type to another type, for example Clustered Bar.
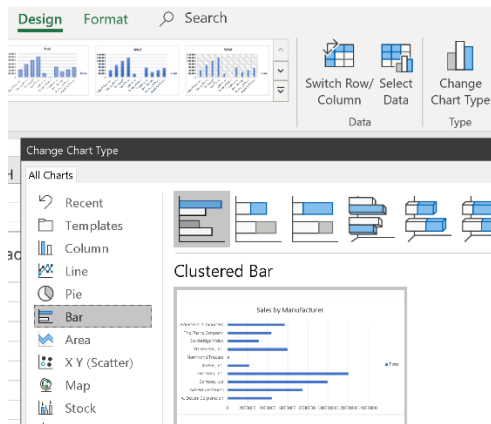
For the next analysis, we can again create a new Pivot Table and repeat the earlier process to get the Sum of Profit by Product Category (Rows) and format the cells with an appropriate numeric format

We can then generate the Pivot Chart (either Clustered Column or Clustered Bar).
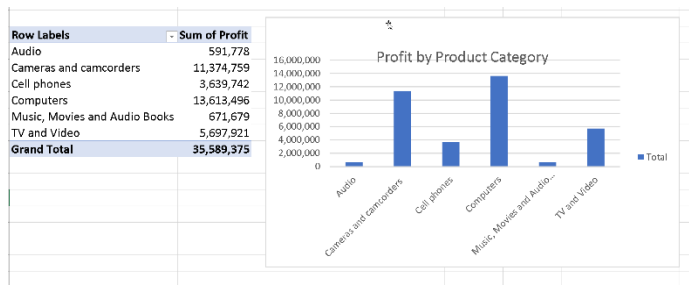
For the next analysis, we can again create a new pivot table and repeat the earlier process to get the Sum of Sales by Product Category (Rows) and Channel (Columns) and format the cells with an appropriate numeric format.

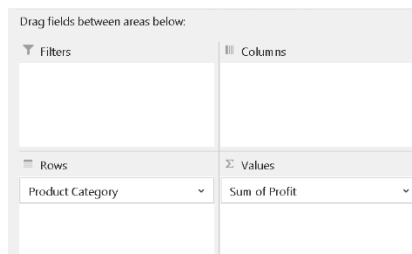| Sum of Sales	Column Labels				
Row Labels	Catalog	Online	Reseller	Store	Grand Total
Audio	53,274	248,765	159,539	567,803	1,029,381
Cameras and camcorders	1,641,341	3,514,850	1,933,395	9,959,345	17,048,932
Cell phones	527,490	1,114,245	845,416	3,354,004	5,841,156
Computers	1,682,876	4,290,928	3,045,515	12,287,652	21,306,971
Music, Movies and Audio Books	111,079	185,645	181,045	579,712	1,057,480
TV and Video	984,231	2,145,775	1,005,703	4,972,131	9,107,839
Grand Total	5,000,291	11,500,207	7,170,614	31,720,647	55,391,760

Then we could use either or both a Clustered Column chart and / or Stacked Column chart to view the differences.

We could also use a 100% stacked column as well, which is useful for seeing how each of the different channels contribute for all products, regardless of the product total - this is useful for products whose sales totals are very small compared to other products and therefore will not come out clearly in the previous graphs.
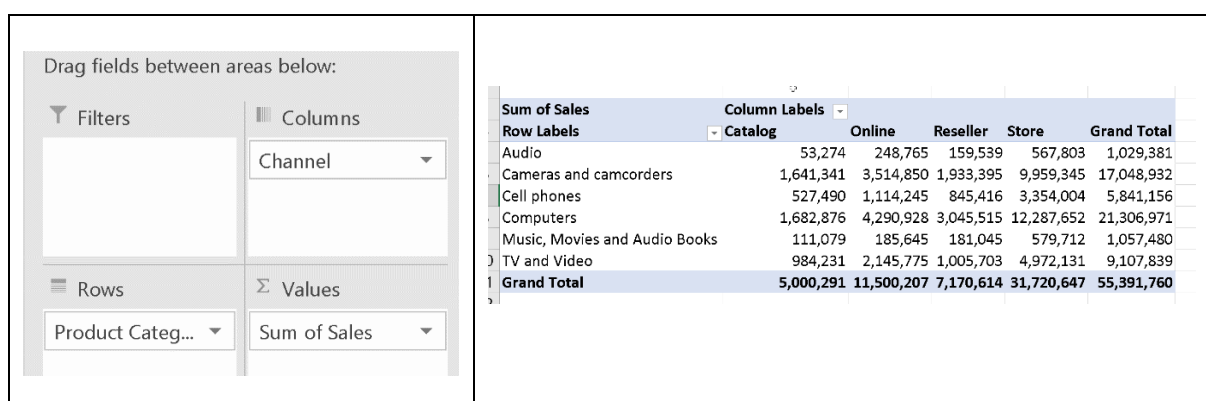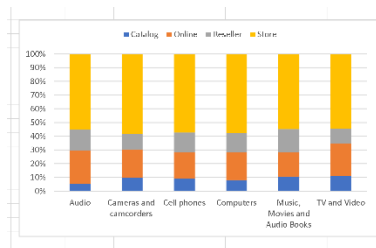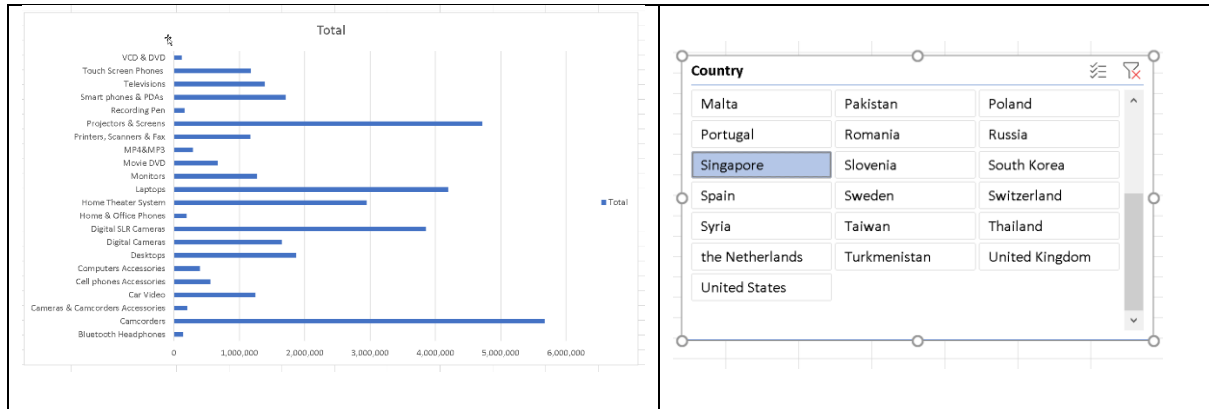

For the next analysis, we can again create a new pivot table and repeat the earlier process to get the Sum of Profit by Product Sub Category (Rows) and format the cells with an appropriate numeric format.


Row Labels	Sum of Profit
Bluetooth Headphones	142,887
Camcorders	5,670,098
Cameras & Camcorders Accessories	202,945
Car Video	1,243,947
Cell phones Accessories	558,835
Computers Accessories	395,441
Desktops	1,870,700
Digital Cameras	1,647,076
Digital SLR Cameras	3,854,640
Home & Office Phones	198,426
Home Theater System	2,946,634
Laptops	4,193,350
Monitors	1,268,828
Movie DVD	671,679
MP4&MP3	288,293
Printers, Scanners & Fax	1,167,342
Projectors & Screens	4,717,835
Recording Pen	160,597
Smart phones & PDAs	1,708,954
Televisions	1,388,547
Touch Screen Phones	1,173,528
VCD & DVD	118,793
Grand Total	35,589,375

Then we generate a standard Clustered Bar Chart as we have done before, and now we can also choose to use a slicer based on the Country field for the pivot table which will also dynamically affect the content of the chart which is linked to the same table.

For the next analysis, we can again create a new pivot table and repeat this earlier process to get the Sum of Sales and Profit by Country (Rows) and format the cells with an appropriate numeric format. Then insert a new Slicer based on Product Subcategory
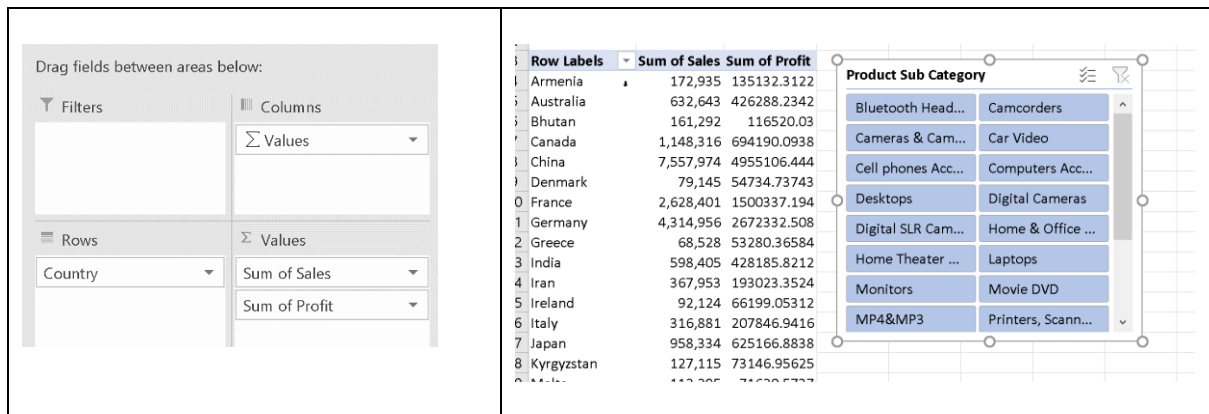
2.1   Practical exercise for comparison Analysis

File to use: File to use: `SalesData-v3.xlsx`

Analytical activities to perform:

1. What is the Total Order Quantity by Manufacturer?
2. What is the average Profit by Brand
3. Create a column graph displaying Total Order Quantity sold by Channel - review by Country
4. Create a bar graph displaying the Total Order quantity sold by Product Category - review by Region

Create two pivot graphs as per the instructions below. Filter both graphs by the Country and Year fields using a slicer.

5. Create a column graph displaying Total Profit by Product Category
6. Create a bar graph displaying Total Profit by Manufacturer

Note: While creating slicers, you can link a single slicer to multiple pivot tables on the same worksheet (if you decide to have more than one), using Report Connections in the Slicer Options.
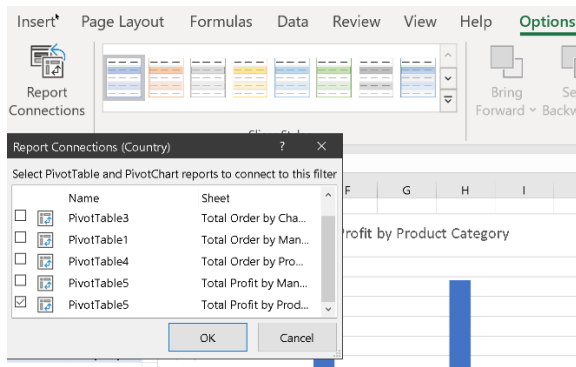
3   Trend Analysis

Trend analysis is all about understanding how data changes over time: is it moving up, is it moving down, is it staying stable or is it volatile. This can be accomplished mainly via trendlines. We also want to study these changes over different time frames: for e.g. changes over years, quarters, months, days. Again, this is best understood with some form of visualization.

Activities to be undertaken:
a) Understand the trend for sales by year and month
b) Understand the trend for sales by year and month by different product categories
c) Develop a seasonality graph displaying the sales by month

File to use: File to use: `SalesData-v4.xlsx`

Start off again by generating a Pivot Table in the usual manner that references the Sales table.
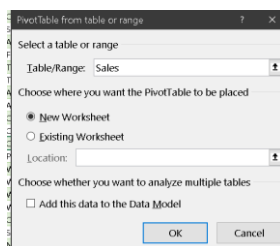
Adding Order Date to the Rows area results in Excel creating a hierarchy grouping involving Years and Quarters or Years and Months.
NOTE: You may get a slightly different result from that shown below depending on the Grouping option that you had created for this column from a previous exercise – that is ok because we will ungroup and regroup again with our custom grouping later.

You can ungroup the hierarchy first before creating your own custom grouping based only on years.

We can then continue to get Sum of Sales for each year

Row Labels	Sum of Sales
2018	18,919,151
2019	17,741,637
2020	17,284,124
2021	1,446,849
Grand Total	55,391,760

We can also alternatively change to another custom grouping (years / months)

Grouping	Row Labels	Sum of Sales

We will insert a Pivot Chart in the form of a Line Chart.
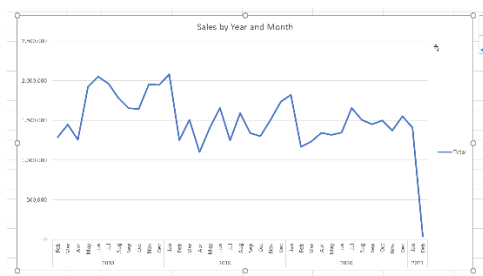
We can also filter out the first 2 months of 2021 to avoid issues with insufficient data at that point which will affect the trend analysis. Make sure the Row Labels are only showing the years (and not the months as well) and click on the drop down arrow.

With a more proper graph, you can examine the various trend line options possible:
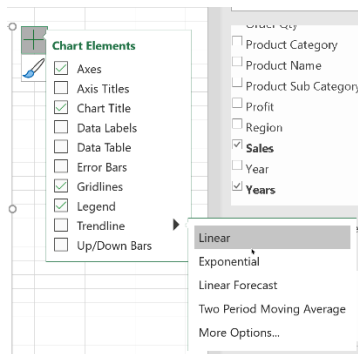
The first option is the default option which is linear. The rest (exponential, linear forecast) are quite similar, but not exactly the same. They are both using different algorithms to calculate what the trend line is. The two period moving average is basically taking two periods and it averages that value between the two periods. It tends to smooth the lines. Its not a trend line as such but tracking the shape of the sales lines.

Selecting More Options leads to the Format Trend line menu, which provides a variety of options such as polynomial, exponential, etc.
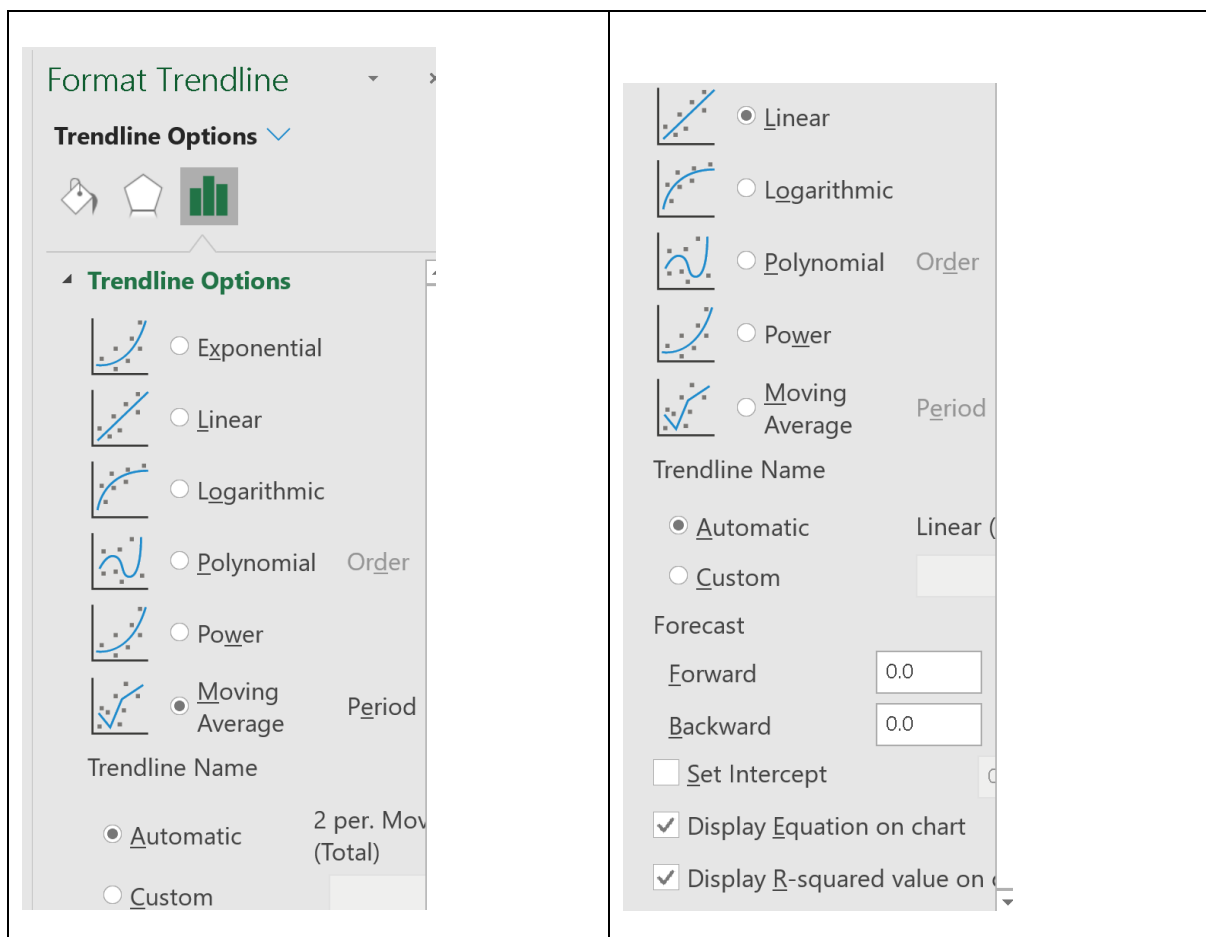
For a linear equation, you can display the trend line equation on the chart and calculate the R-squared, which is a measure of how accurate the predictions would be from the trend line equation.

The next step is to refine this further to understand the trend for sales by year and month by different product categories. We can simply copy the existing pivot table to a new worksheet and work from there.

We can add a PivotChart that is a line chart and stacked line chart to get an initial view of this data.
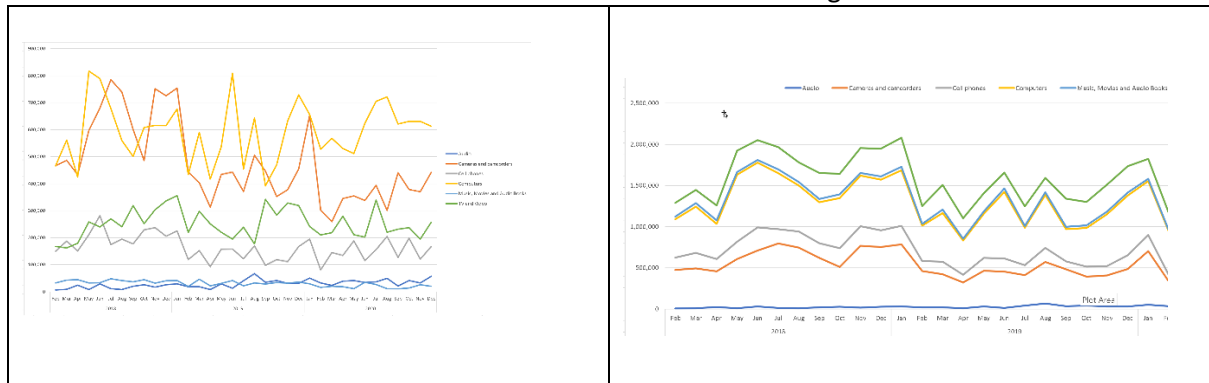
You can select which particular product category to project a trend line on if you wish.

We can also change this to either an area graph and stacked area graph to get a better view of the contributions of each product category trending over time.

You could also take your Product Category out of the Columns and utilize it as a slicer instead, which then allows you to view the trends for each of these categories across the different chart types.
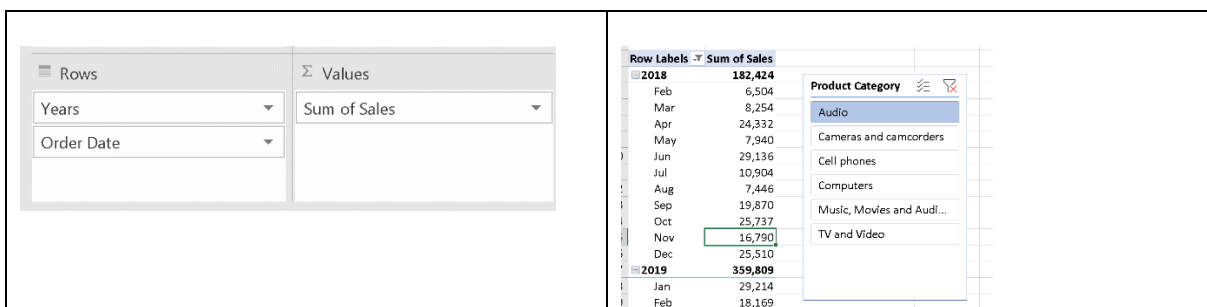
The next thing we will look is a seasonality graph, which we can create from another copy of the pivot table on a new worksheet. For seasonality, we typically want to take into account the years which have data for all the months, and filter out those that do not.

Row Labels	Sum of Sales
2019	**17,741,637**
Jan	2,081,908
Feb	1,252,298
Mar	1,507,751
Apr	1,103,451
May	1,408,131
Jun	1,657,980
Jul	1,248,653
Aug	1,593,694
Sep	1,340,433
Oct	1,300,579
Nov	1,511,809
Dec	1,734,950
2020	**17,284,124**
Jan	1,824,856
Feb	1,168,697
Mar	1,234,561
Apr	1,345,859
May	1,318,429
Jun	1,347,578
Jul	1,656,039
Aug	1,507,838
Sep	1,452,020
Oct	1,499,587
Nov	1,374,350
Dec	1,554,311
Grand Total	**35,025,760**

Next change custom grouping to Month. This will combine data from all the years concerned (2019, 2020).

Row Labels	Sum of Sales
Jan	5,315,113
Feb	3,749,046
Mar	4,190,56?
Apr	3,706,018
May	4,650,233
Jun	5,057,531
Jul	4,870,610
Aug	4,883,301
Sep	4,446,563
Oct	4,443,978
Nov	4,841,686
Dec	5,237,119
Grand Total	**55.391.760**

Then we can generate a line graph to demonstrate the Seasonality.
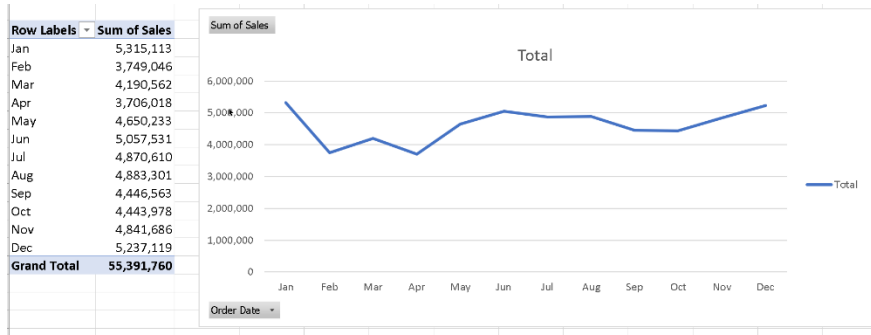
Seasonality is really useful to understand so that we can understand where resource / sales demand peaks in specific periods of a year and therefore can significantly help with resource allocation planning. Its important to make sure you have a complete set of data for the entire time duration you are interested to perform analysis on (for e.g. all the months of a year, rather than partial months) because that will skew the visualization of the graph.

3.1   Practical exercise for Trend Analysis

Analytical activities to perform:

a)   Create a line graph displaying the Profit by Year and Quarter

b)   Add a polynomial trend line to the line graph

c)   Add a Slicer for Product Category

d)   Create an area graph displaying the Profit by Year and Quarter for the Product Category

e)   Add a Slicer for Country

File to use: File to use: `SalesData-v5.xlsx`

For first solution a), we can specifically target the quarters which are outliers in the dataset (suggesting that the data might be incomplete) and remove them using a Date filter with a Between.
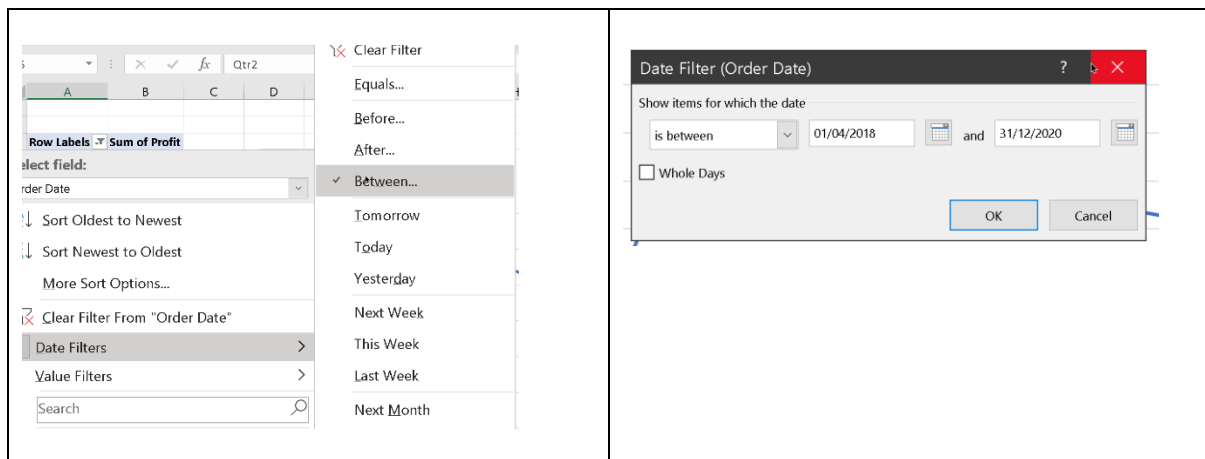
Row Labels	Sum of Profit
2018	**10,266,628**
Qtr2	3,352,068
Qtr3	3,431,388
Qtr4	3,483,171
2019	**11,378,957**
Qtr1	3,172,056
Qtr2	2,541,105
Qtr3	2,731,591
Qtr4	2,934,205
2020	**11,237,878**
Qtr1	2,705,240
Qtr2	2,588,481
Qtr3	3,037,253
Qtr4	2,906,905
Grand Total	**32,883,463**