Summary: Data Cleaning with Power BI

SESSION OVERVIEW:

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

- Understand various ways to do data cleaning and manipulation in Power BI
- Understand the features available under the Add Column, View, Tools and Help tabs in the ribbon
- Use various AI tools in Power BI

KEY TOPICS AND EXAMPLES:

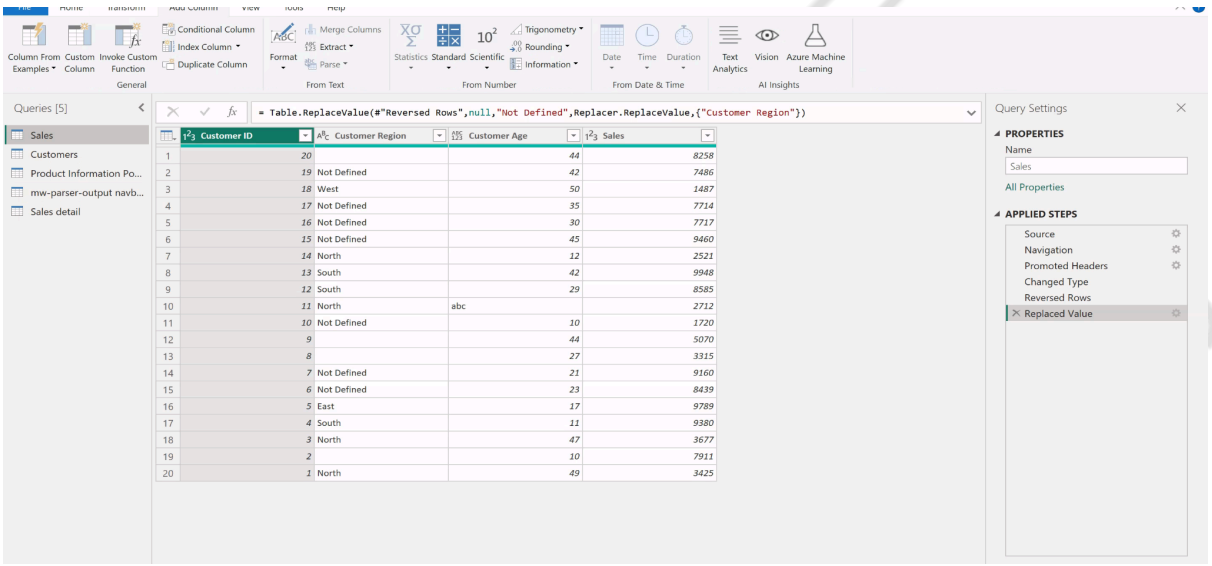
Using Power Query Editor for Data Cleaning

While doing data cleaning through Power Query, it is essential to realize that Power BI just reads data, it does not write data back to the source. In other words, the original dataset remains unchanged while you transform/clean data on Power BI.

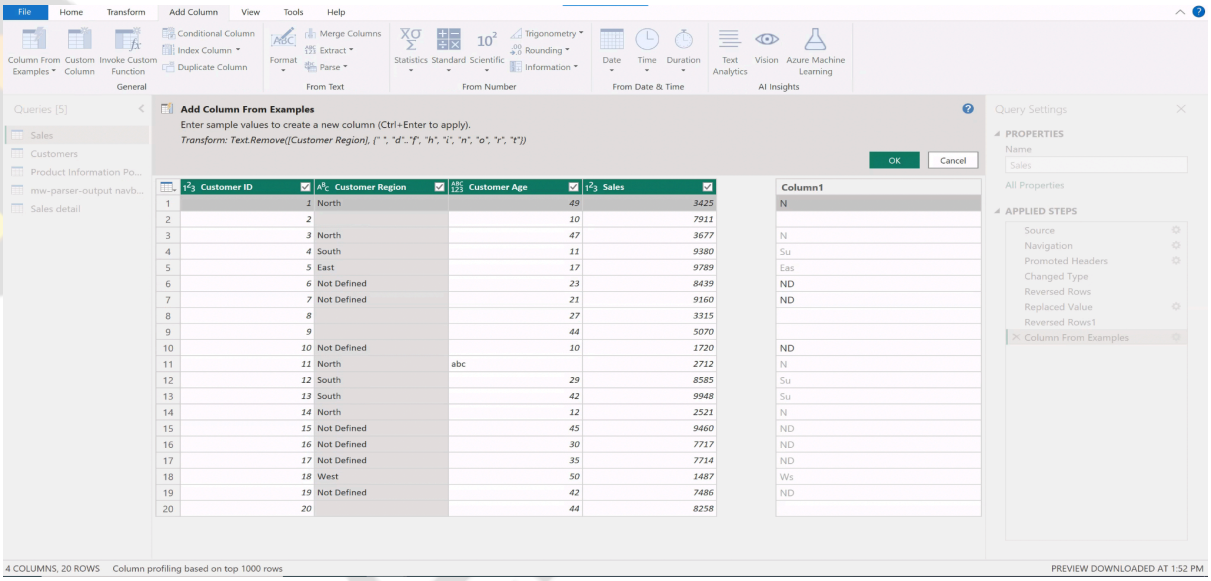
Whenever we perform any cleaning/transformation function, it gets added as a new **step** in the window on the right side: in the Query Settings window. Each step can be deleted to reverse the previous action. This helps you manage your data manipulations (for transformation/cleaning).

Power Query Editor provides some of the features and functionality given below for cleaning data:

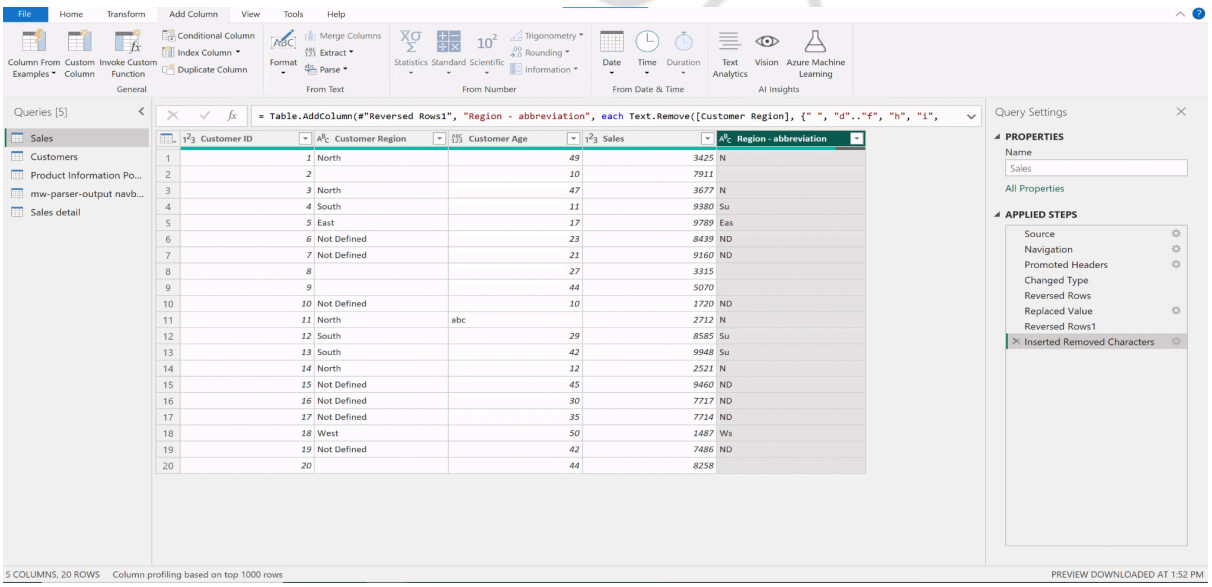
1. The **Add Column** tab has multiple options to clean the data. These are explained below.



- **Add Column From Examples**: This is used to create a new column from an existing column. Example: From date we can get month, year etc.
  Example: The steps and image below demonstrates its use from this Dataset.



i. Select the column from which you intend to generate a new column.
ii. Select '<u>Column From Examples</u>' → '<u>From Selection</u>'.
iii. Enter sample values into the new column. **Power BI will intelligently recognise the formula that may help you get those values**.
iv. Click OK or press Ctrl+Enter to apply.



- **Custom Column:** We can also create a new column based on a custom formula. These formulas are written in M Query. **M Query is Power BI language used in Power Query editor to transform, clean and manipulate data.** A detailed list of M Query functions is given here: https://learn.microsoft.com/en-us/powerquery-m/power-query-m-function-reference

  Example: In the Product Information dataset, we can calculate the sales for each customer by multiplying the **Qty bought** and **Price per item**. This will give us the sales value for each customer ID.The image below demonstrates this formula.

  Some other Formulas:
  To create a new column **"Bulk Purchase"**, which will mark orders as **"High"** if Qty bought is greater than 50, otherwise mark them as **"Low"**.

```
if [Qty bought] > 50 then "High" else "Low"
```

  To categorize the "Bulk Purchase" column into **3 categories**.

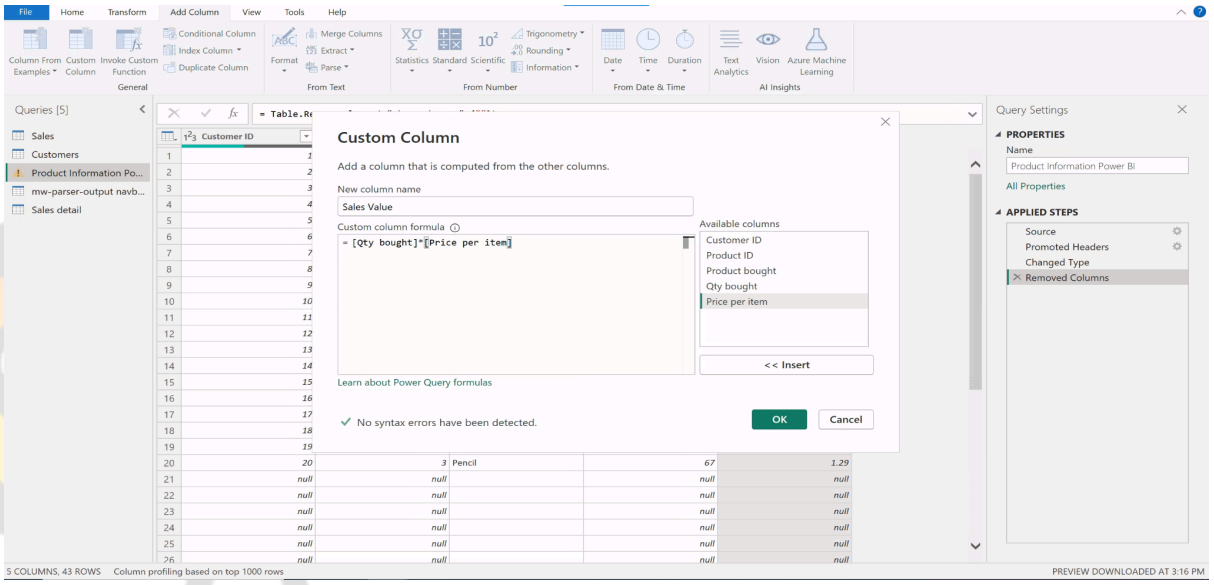
```
if [Qty bought] > 50 then "High"
else if [Qty bought] >= 30 and [Qty bought] <= 50 then "Medium"
else "Low"
```

  Query for adding a new column named "**Price**" that **concatenates** "**$**" with the existing "**Price per item**" column.

```
"$" & Number.ToText([Price per item])
```

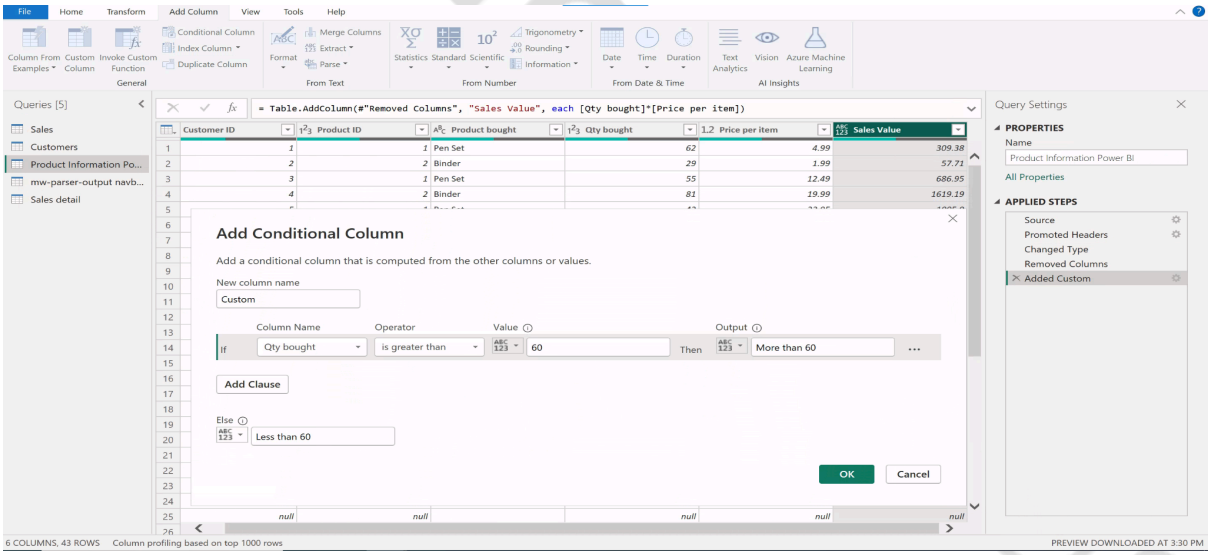
  To categorize products based on their **price per item**, capitalize the **category labels**, and concatenate them with their respective **price per item** values to create a more descriptive and readable output.

```
= if [Price per item] > 10 then Text.Proper("expensive") & " - $" &
Number.ToText([Price per item])

  else if [Price per item] >= 5 and [Price per item] <= 10 then
Text.Proper("affordable") & " - $" & Number.ToText([Price per item])

  else Text.Proper("cheap") & " - $" & Number.ToText([Price per
item])
```



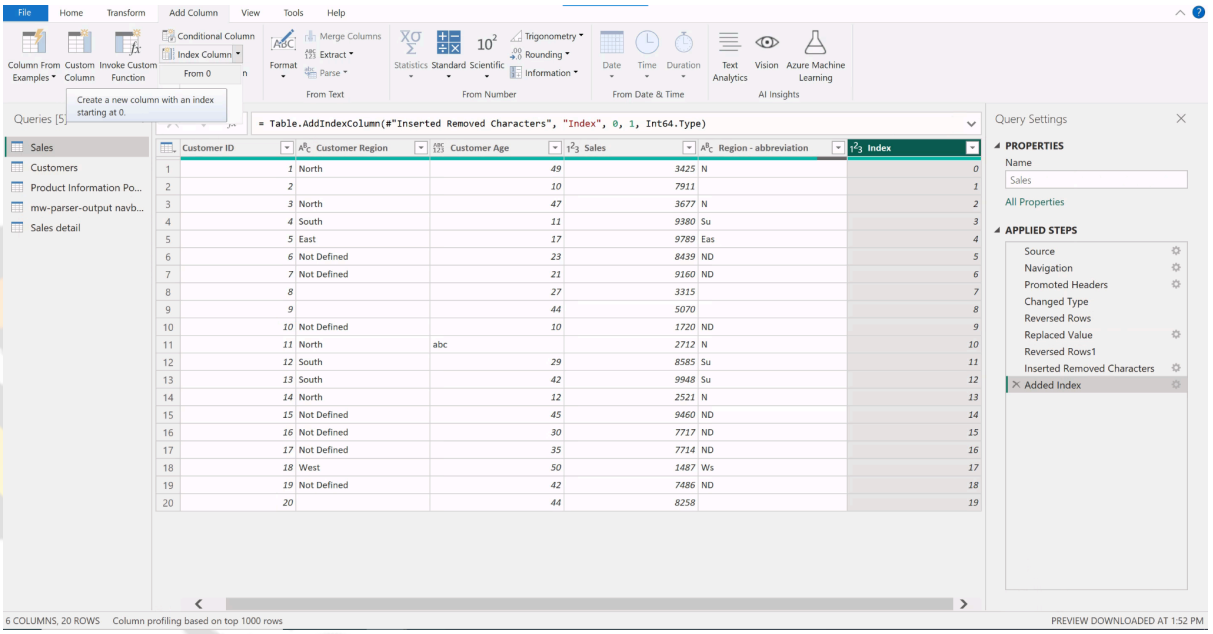
- **Conditional Column:** It is used to create a new column using a condition that is based on another column.

  Example: In the [Product Information dataset], if we want to create a bucket for customers that bought more than 60 items, then we can use a conditional column for that.

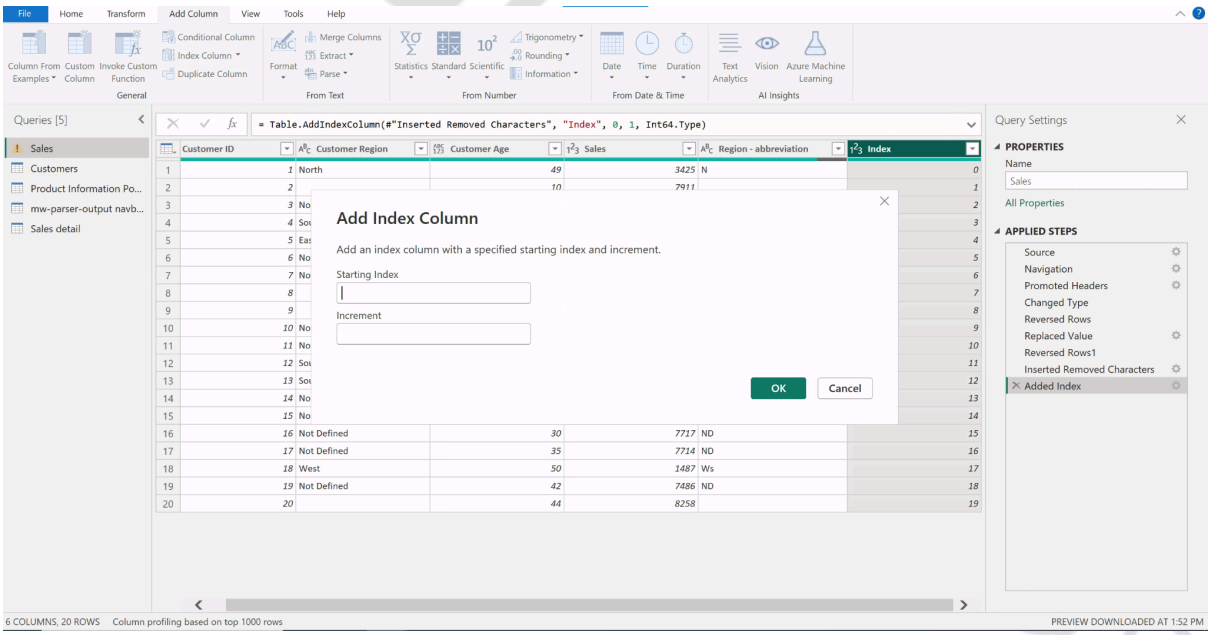


- **Index Column**: It creates a new column (from 0 or 1).

  Example: In our Sales dataset, we do not have an index column. It can be added from the ‘Add Column’ tab → ‘Index Column’ → ‘From 0’.



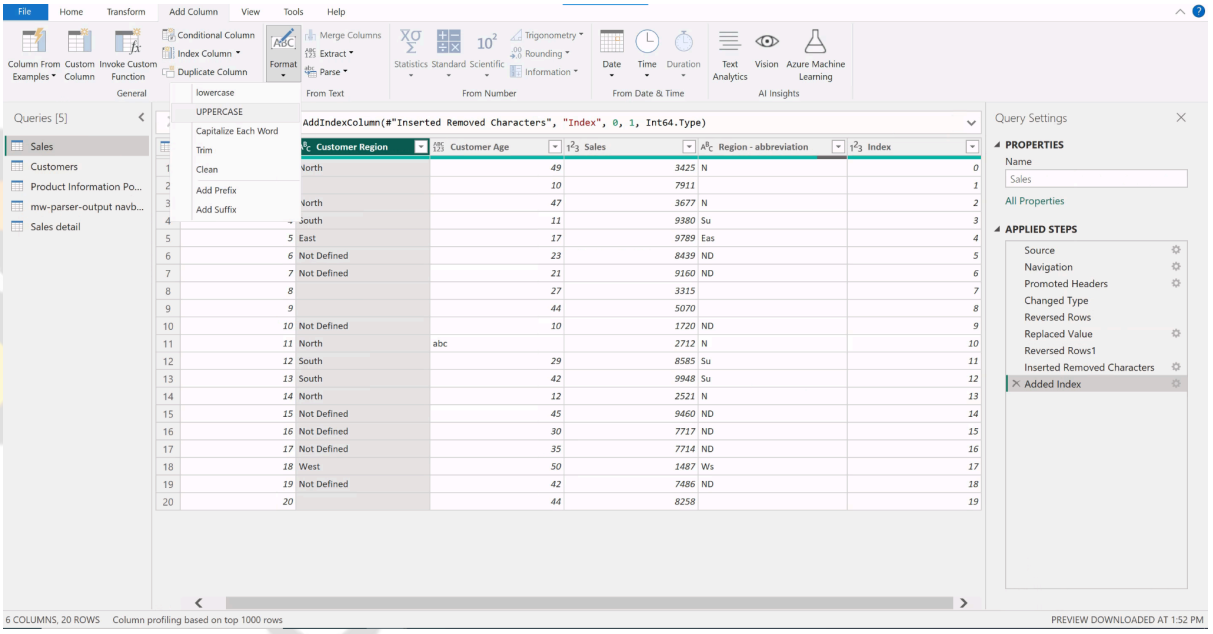
Index column option provides us with the flexibility to start from 0, 1 or any custom number with a custom increment (see image below).



Note: By default, Power BI adds an Index column at the end but it can be moved to the beginning by simply dragging the header of the column.

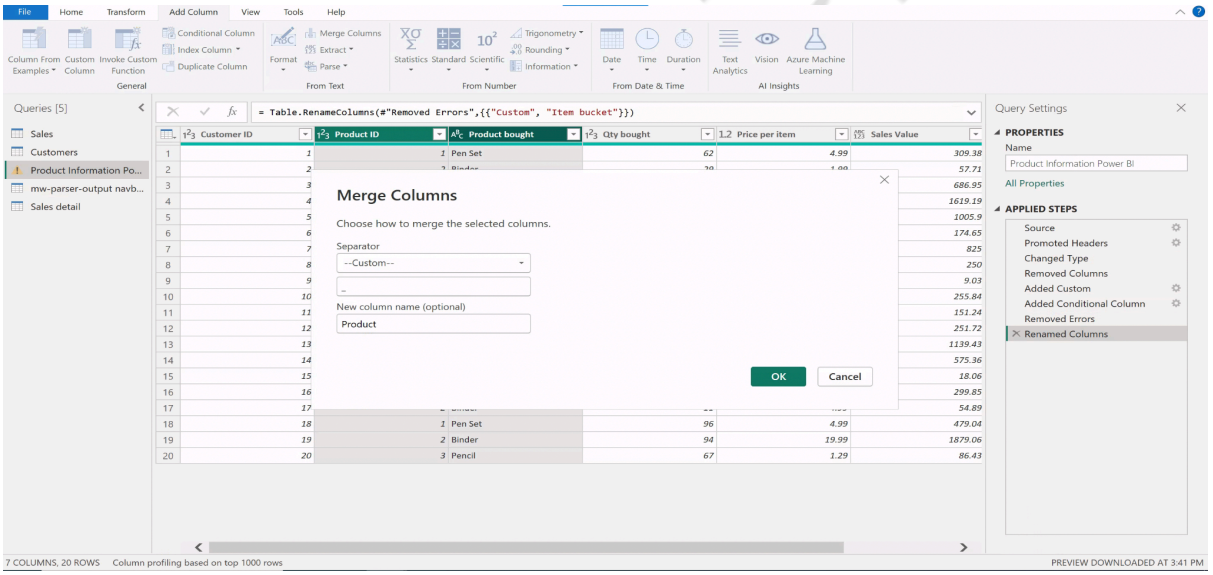
- **Format:** It is used to change the case of the text. It can be used to convert it to lowercase, uppercase, capitalize each word, trim, clean or add prefix/suffix.

Example: We can choose from the options below in the dataset. We can convert the ‘Customer Region’ column to UPPERCASE.

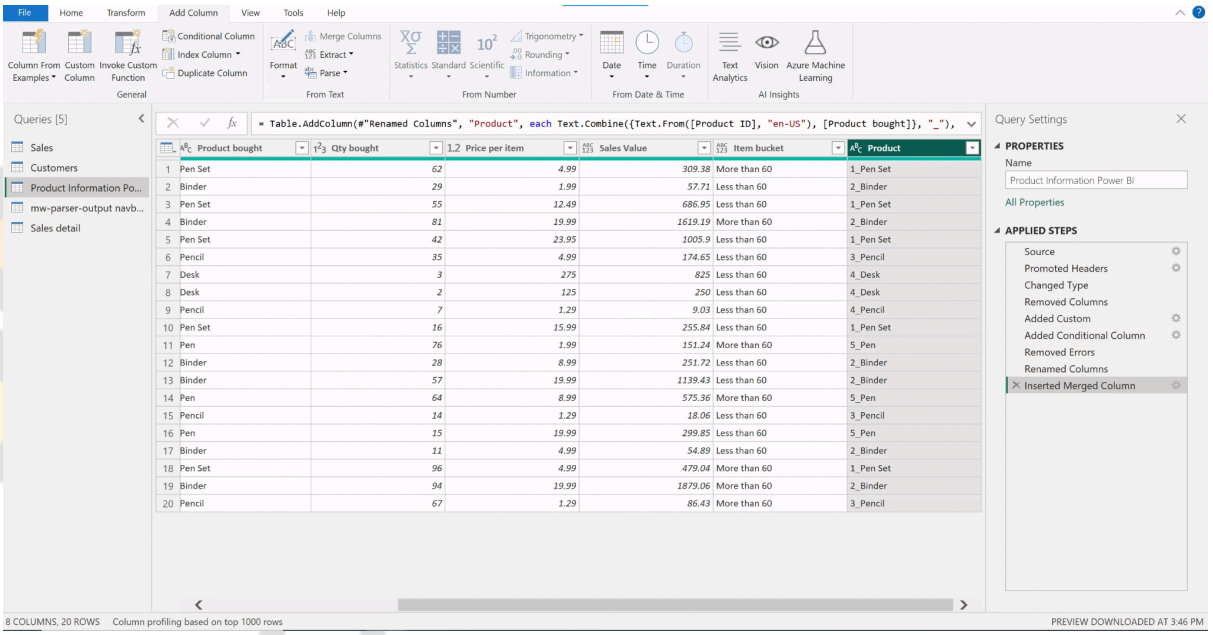


- **Merge Columns:** This is used to merge multiple columns. To perform this operation we need to select multiple columns, click on the Merge Columns option and then select the type of separator we require.

  Example: Suppose, in the Product Information dataset we want to merge the ‘Product ID’ and ‘Product bought’ column to create a single column that will be in this format: ‘ProductID_ProductBought’. To do this, we can use the ‘Merge Columns’ operation, as shown below.

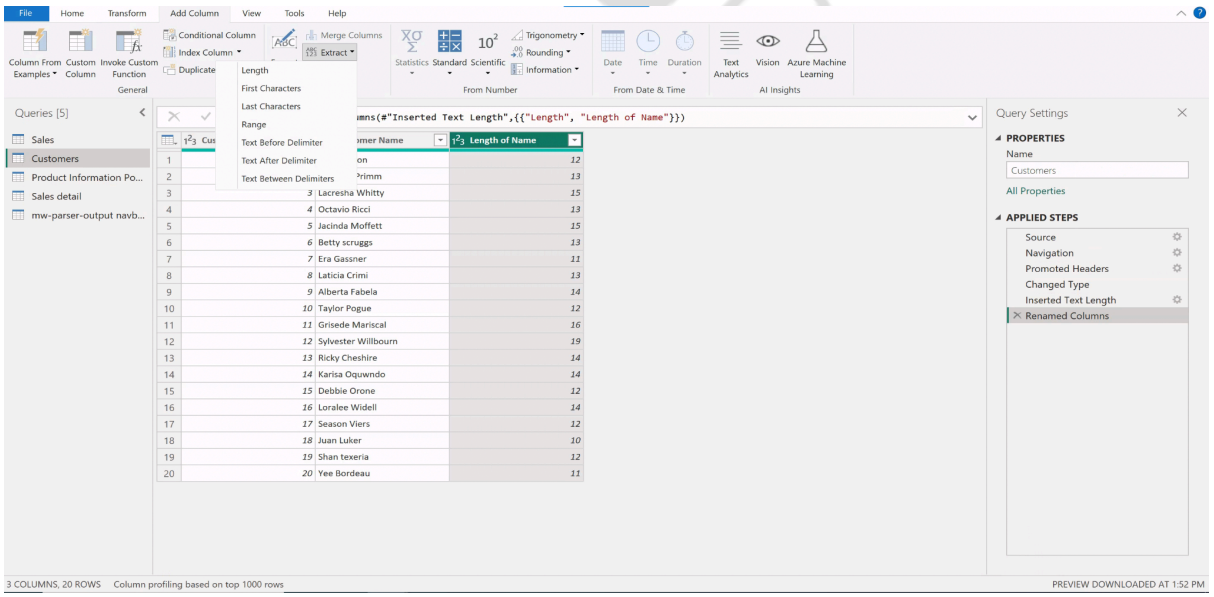


The final, merged column will look like this:



- **Extract:** Using **Extract** we can extract the first characters, last characters, length, Range, Text before delimiter, Text after delimiter and Text Between Delimiters.

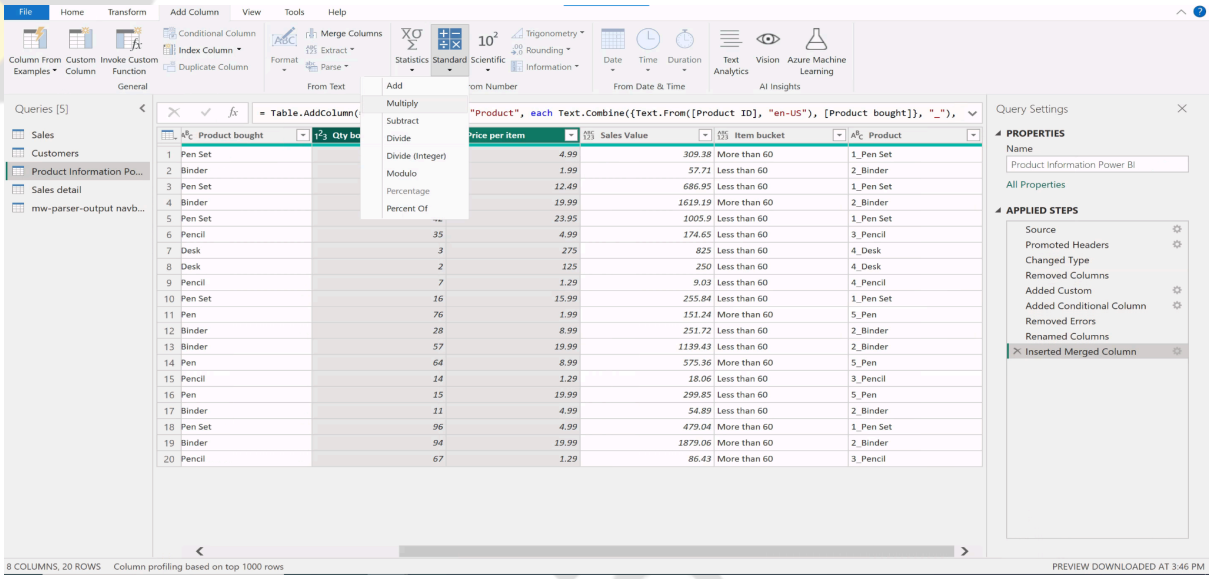
  Example: Using the Customers table in the Sales dataset, if we want to find the length of the name of the customer, we can use the following steps: click 'Extract' → Select 'Length' → rename the field as per your convenience. The final result as well as the options Extract provides are shown in the image below.



- **Statistics:** We have numerical operations like ‘Statistics’ that can be used to create outputs that are based on statistical measures like mean, median, mode etc.

- **Standard:** It enables you to perform basic mathematical operations on your data, like Add, Subtract, Multiply, Divide etc. We can either perform these operations on the selected column with a fixed value or with another column.

  Example: In our Product Information dataset, we can multiply “Qty bought” with “Price per item” using this operation. Image below shows this operation being performed.



- **Scientific:** It is used to perform scientific operations like square root, power, absolute value and exponent.

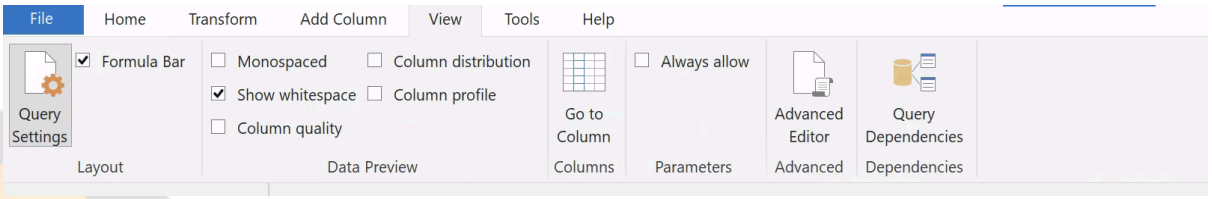
AI TOOLS IN POWER BI:

- **Text Analytics:** Helps in detecting the language of a column, extract key phrases and generate sentiment scores.

- **Vision:** It helps us analyse the images and return tags that are based on what the picture contains.

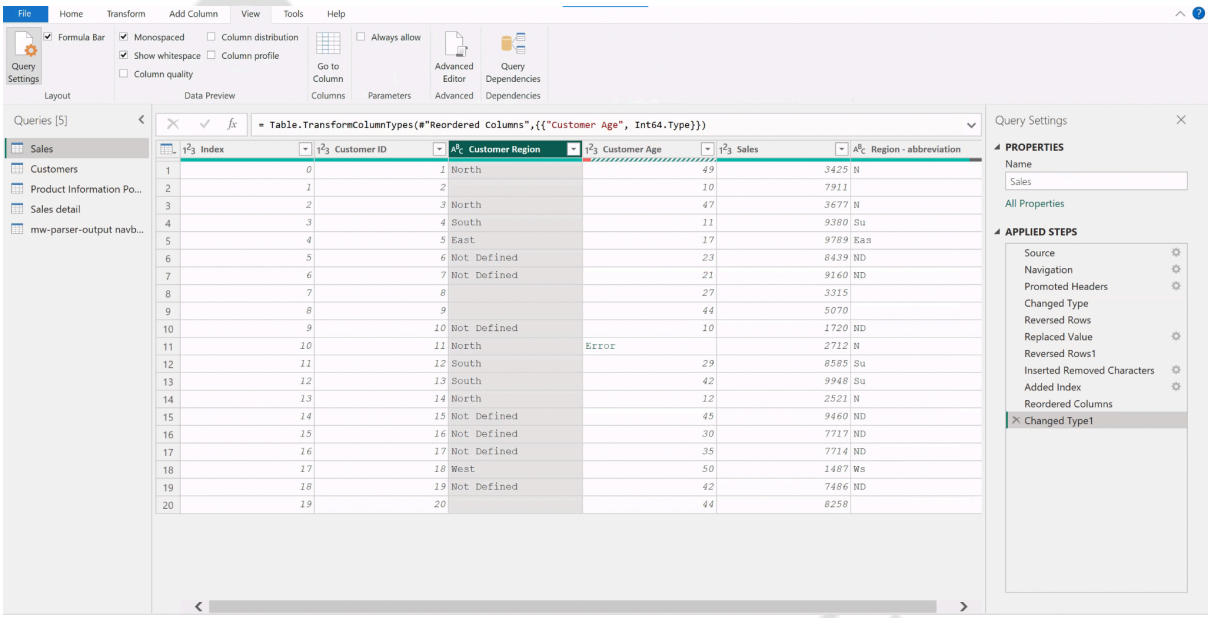
- **Azure Machine Learning:** If your organisation has subscribed for Azure Machine Learning, then some of its options can be found here through which the BI analyst can directly use those models in Power Query Editor.

2. The **View** tab enables us to add/disable some settings and helps us view data in a better format. It has some useful options for cleaning data as listed below.



- **Monospaced:** It displays data in a fixed-width font, making each character the same width.

  Example: The image below from the Sales dataset shows data with the monospace check-box ticked.

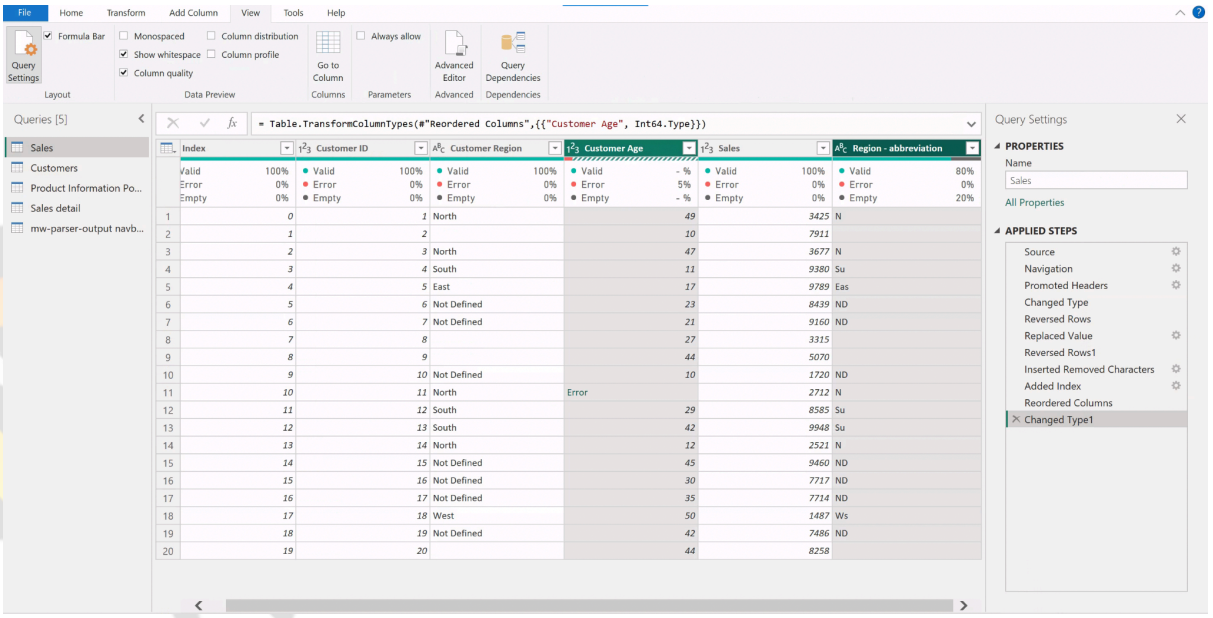


- **Column Quality:** It shows the percentage of valid, error, or empty values in a column. It helps us identify errors easily and fix them.

  Example: The image below shows the column quality for some of the columns in the dataset.

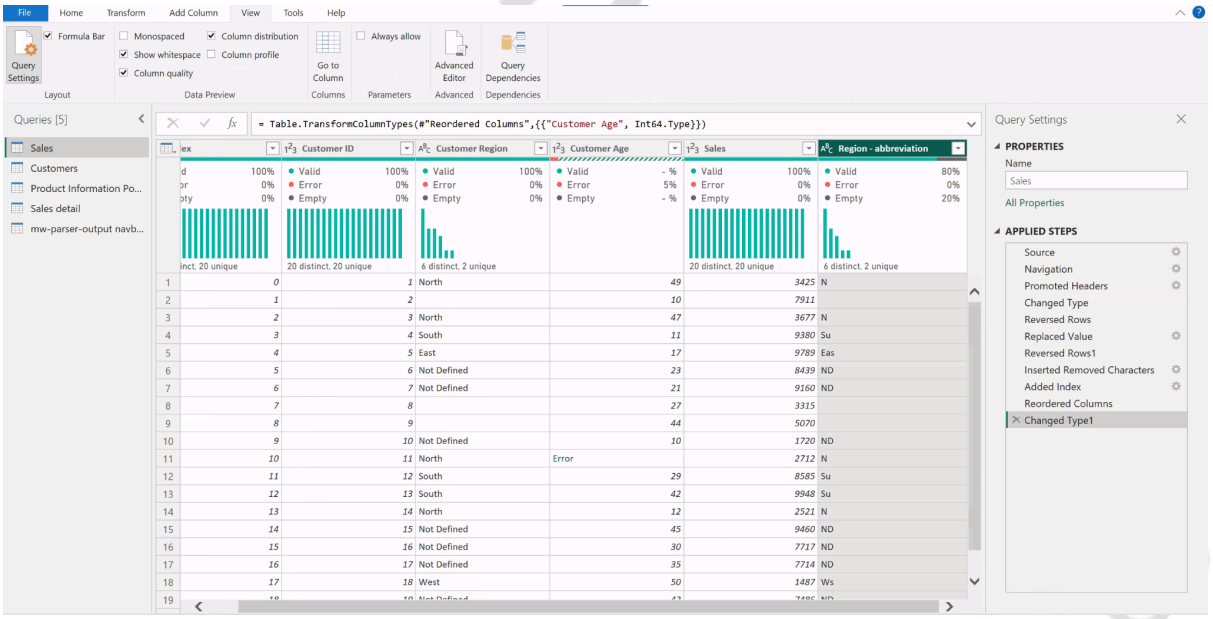
  ➔ The Region-abbreviation column has 20% empty values, so we need to handle these missing values separately. For example, we can right click on the column and replace empty values by some fixed value.

  ➔ The customer age column has 5% error values. We can remove errors by right clicking on the column and replacing errors by a fixed value.



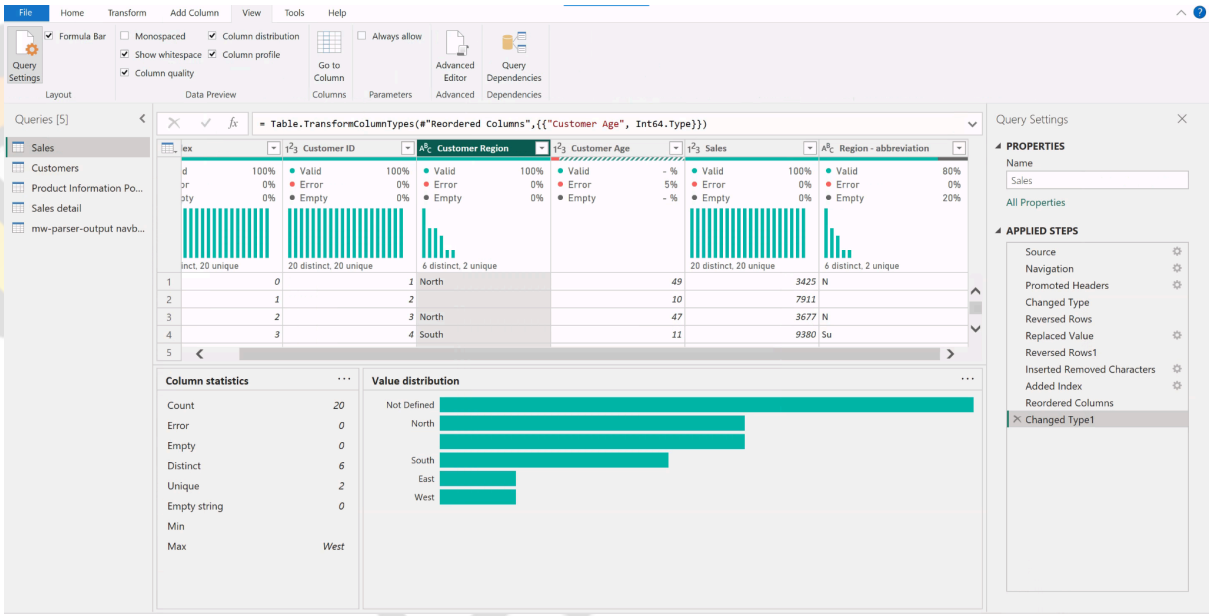
- **Column Distribution**: It is used to view the distribution of the data in each column. In particular, it displays the number of distinct and unique values in each column.

  Example: In the Sales dataset, it enables one to identify the different regions and the number of distinct/unique values.



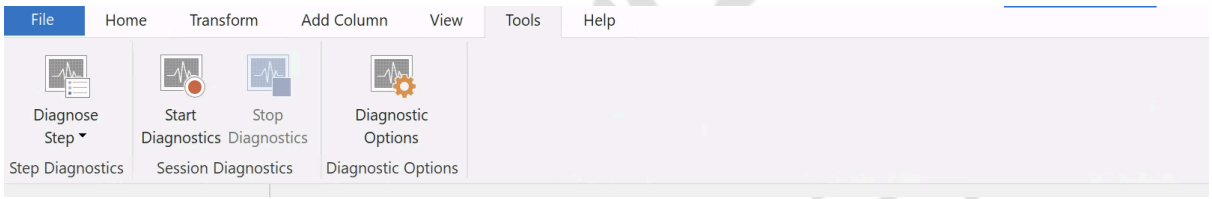
- **Column Profile:** It provides statistical information about the distribution of data, including errors, empty values, distinct values, minimum values, and maximum values. It provides a snapshot of the column and can be very useful to understand the column data distribution.

Example: In the Sales dataset, if we want to understand the distribution of each region as well as the errors, empty values, distinct values, minimum values, and maximum values, **Column profile** would be the best option, as displayed below. It also displays the quantum in each of these categories of region.



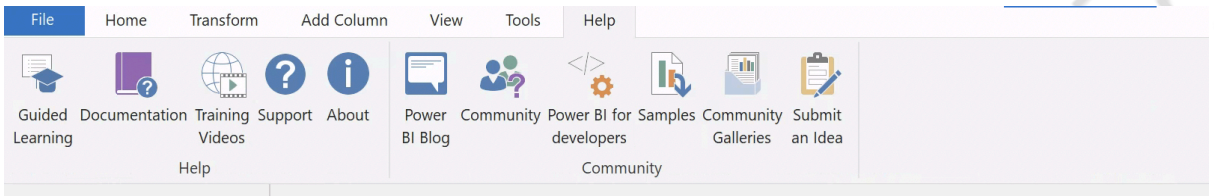
Note: Notice that on the right side we can see all the applied steps from the beginning. At any point we can remove any of these steps or go back and change the manner in which we performed any of these steps.

3. **The Tools tab:** It helps us diagnose and troubleshoot data transformations.



The image above shows the options available in the Tools tab.

4. **The Help tab:** It provides us with documentation and other information to help us understand the features of Power Query Editor.



The image above shows the options available in the Help tab.