

Internshala Trainings

Data Science

PowerBI Assignment 1

Prepared By,
Utkarsh Anand

Task 1 – Data Cleaning

The screenshot shows the Power BI Desktop interface with the 'Transform' tab selected in the ribbon. In the main area, a table named 'Table.TransformColumns(#"Removed Duplicates", {"Price", each Number.Round(_, 0), type number})' is displayed. The 'Price' column has been rounded to the nearest integer. A 'Round' dialog box is open, asking to specify decimal places, with '0' entered. The 'APPLIED STEPS' pane on the right shows the step 'Rounded Off' with a gear icon, indicating it's a new step. The status bar at the bottom right shows 'PREVIEW DOWNLOADED AT 14:09' and the date '05-09-2025'.

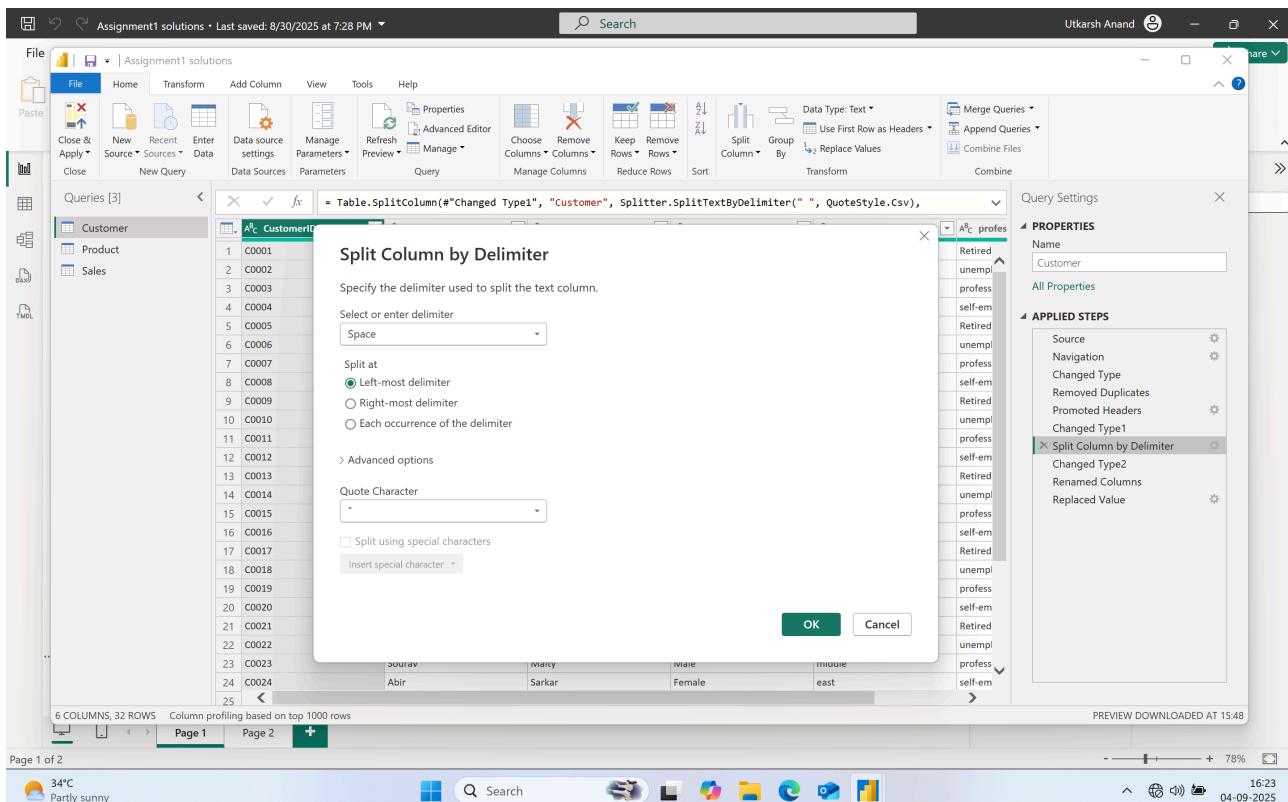
Part 1: Round the ‘Price’ column (Product) to the nearest integer

Steps taken to round the ‘Price’ column in the Product dataset to the nearest integer for simplicity:

1. In Power BI Desktop, on the Home ribbon, click Transform data to open Power Query Editor.
2. In the Queries pane (left), click the Product table.
3. Click the Price column header to select the entire column.
4. Go to the Transform tab (top) → in the Number Column group, click Rounding ▼ → choose Round.
5. If a gear icon appears next to the new step in Applied Steps (right), click it and set Decimal places = 0, then OK (this guarantees whole numbers).

6. Confirm that the Price values now display as whole numbers (no decimals).
7. Click Home (in Power Query) → Close & Apply to save changes back to the model.

Part 2: Split the ‘Customer’ column (Customer) into ‘FirstName’ and ‘LastName’



Steps taken to split the ‘Customer’ column in the Customer table into two columns — ‘FirstName’ and ‘LastName’:

1. On the Home ribbon in Power BI Desktop, click Transform data to open Power Query Editor (if not already open).
2. In Queries (left), click the Customer table.
3. Click the Customer column header.
4. Go to Home tab → Split Column ▾ → By Delimiter...

5. In the dialog:

- Select or enter delimiter: choose Space.
 - Open Advanced options → set Split at: Left-most delimiter (creates only two parts).
 - Click OK.
6. Rename the resulting columns: double-click the headers and set them to FirstName and LastName.
7. Click Home → Close & Apply.

Task 2 – Standardization

The screenshot shows the Microsoft Power Query Editor interface. The main area displays a table with columns: ProductID, Category, Name, Size, and Price. The 'Category' column has been transformed using the formula `= Table.TransformColumns(#"Rounded Off", {{"Category", Text.Upper, type text}})`. The 'APPLIED STEPS' pane on the right lists the following transformations: Source, Navigation, Promoted Headers, Changed Type, Removed Duplicates, Rounded Off, Uppercased Text (which is currently selected), and Capitalized Each Word.

ProductID	Category	Name	Size	Price
P001	SNACKS	Product1	Small	67
P002	DRINKS	Product2	Large	85
P003	DRINKS	Product3	medium	35
P004	SNACKS	Product4	Small	44
P005	CHOCOLATES	Product5	medium	65
P006	JELLY	Product6	Small	20
P007	JELLY	Product7	Large	11
P008	JELLY	Product8	Large	65
P009	SNACKS	Product9	Small	91
P010	CHOCOLATES	Product10	medium	43
P011	CHOCOLATES	Product11	medium	95
P012	DRINKS	Product12	Large	92
P013	DRINKS	Product13	Small	72
P014	SNACKS	Product14	Large	12
P015	DRINKS	Product15	medium	26
P016	JELLY	Product16	Small	94
P017	SNACKS	Product17	medium	52
P018	DRINKS	Product18	Small	9
P019	DRINKS	Product19	Large	57
P020	SNACKS	Product20	Large	62
P021	CHOCOLATES	Product21	Small	9
P022	JELLY	Product22	medium	19
P023	JELLY	Product23	medium	27
P024	JELLY	Product24	Large	51
P025	SNACKS	Product25	Small	17
P026	CHOCOLATES	Product26	Large	95

Part 1: Convert ‘Category’ (Product) to UPPERCASE

Steps taken to convert all entries in the ‘Category’ column in the Product table to uppercase:

1. Open Power Query Editor (Home → Transform data).
2. Select the Product table → click the Category column header.
3. Go to Transform tab → Format ▼ → click UPPERCASE.
4. Verify that each Category value is now fully capitalized.
5. Click Home → Close & Apply.

Part 2: Replace ‘unemployment’ with ‘Unemployed’ in ‘Profession’ (Customer)

The screenshot shows the Microsoft Power Query Editor interface. The 'File' ribbon is selected. In the center, a 'Replace Values' dialog box is open over a table named 'Customer'. The dialog box has fields for 'Value to Find' (containing 'unemployment') and 'Replace With' (containing 'Unemployed'). Advanced options like 'Match entire cell contents' are checked. The table below shows the 'First Name' column with various names, and the 'Profession' column with some entries like 'Unemployed' and 'self-employed'. The right side of the screen displays the 'Query Settings' pane, which includes sections for 'PROPERTIES' (Name set to 'Customer') and 'APPLIED STEPS' (listing various transformations applied to the query). The status bar at the bottom indicates 'PREVIEW DOWNLOADED AT 15:48' and shows system icons.

Steps taken to replace all occurrences of ‘unemployment’ with ‘Unemployed’ in the ‘Profession’ column of the Customer table:

1. In Power Query Editor, select the Customer table.
2. Click the Profession column header.
3. Go to Home tab → click Replace Values.
4. In the dialog:
 - Value To Find: unemployment
 - Replace With: Unemployed
 - Click Advanced options and tick Match entire cell contents (ensures only the exact word is replaced).
 - Click OK.
5. Review a few rows to confirm the change, then Close & Apply.

Task 3 – Data Types & Consistency

The screenshot shows the Microsoft Power Query Editor interface. The 'Customer' table has the following data:

CustomerID	First Name	Last Name	Gender	Area	profession
C0001	Sujata	Mohanty	Male	middle	Retired
C0002	Suraj	Rajput	Male	east	Unemployed
C0003	Pramod	Bhavar	Male	east	profession
C0004	Satish	Ojha	Male	west	self-employed
C0005	Sintu	Kumar	Male	middle	Retired
C0006	Krutiika	Shelar	Male	middle	Unemployed
C0007	Arjun	Shaw	Male	east	profession
C0008	Shrikant	Badge	Female	west	self-employed
C0009	Jitender	Kumar	Male	south	Retired
C0010	Dharmendar	Rana	Male	middle	Unemployed
C0011	Adnan	Soukat	Female	south	profession
C0012	Sheetal	Nishad	Male	middle	self-employed
C0013	Monika	Pawar	Female	east	Retired
C0014	Meena	Mourya	Male	east	Unemployed
C0015	Ashu	Sharma	Male	west	profession
C0016	Harivansh	Gautam	Male	middle	self-employed
C0017	Vini	Saini	Female	middle	Retired
C0018	Anand	Singh	Male	east	Unemployed
C0019	Jaishri	Saxena	Male	west	profession
C0020	Virender	Sroha	Male	south	self-employed
C0021	Shrikant	Badge	Female	middle	Retired
C0022	Harivansh	Gautam	Male	south	Unemployed
C0023	Sourav	Matty	Male	middle	profession
C0024	Abir	Sarkar	Female	east	self-employed
C0025	Sandeep	Aswal	Male	east	Retired
C0026	Karan	Kapoor	Male	west	Unemployed
C0027	Kishor	Panara	Female	middle	profession
C0028	Manpreet	Kaur	Male	middle	self-employed
C0029	Aakanksha	Srivastava	Male	east	Retired

PREVIEW DOWNLOADED AT 15:48

17:12 04-09-2025

Part 1: Ensure appropriate data types (Date, Price, etc.)

Steps taken to ensure all columns in the datasets have appropriate data types (e.g., ‘Date’ as Date, ‘Price’ as Decimal/Whole Number):

1. Open Power Query Editor.

2. Sales table:

- Click the Date column → on Home tab (Power Query), open the Data Type dropdown (icon left of column name or on ribbon) → choose Date → Replace current.

3. Product table:

- Click Price → set Data Type to Whole Number (since we rounded to integers) or Decimal Number (if you kept decimals).

4. Repeat for any other columns (IDs as Whole Number/Text as appropriate).

5. Confirm the ABC/123/calendar icons reflect the correct types, then Close & Apply.

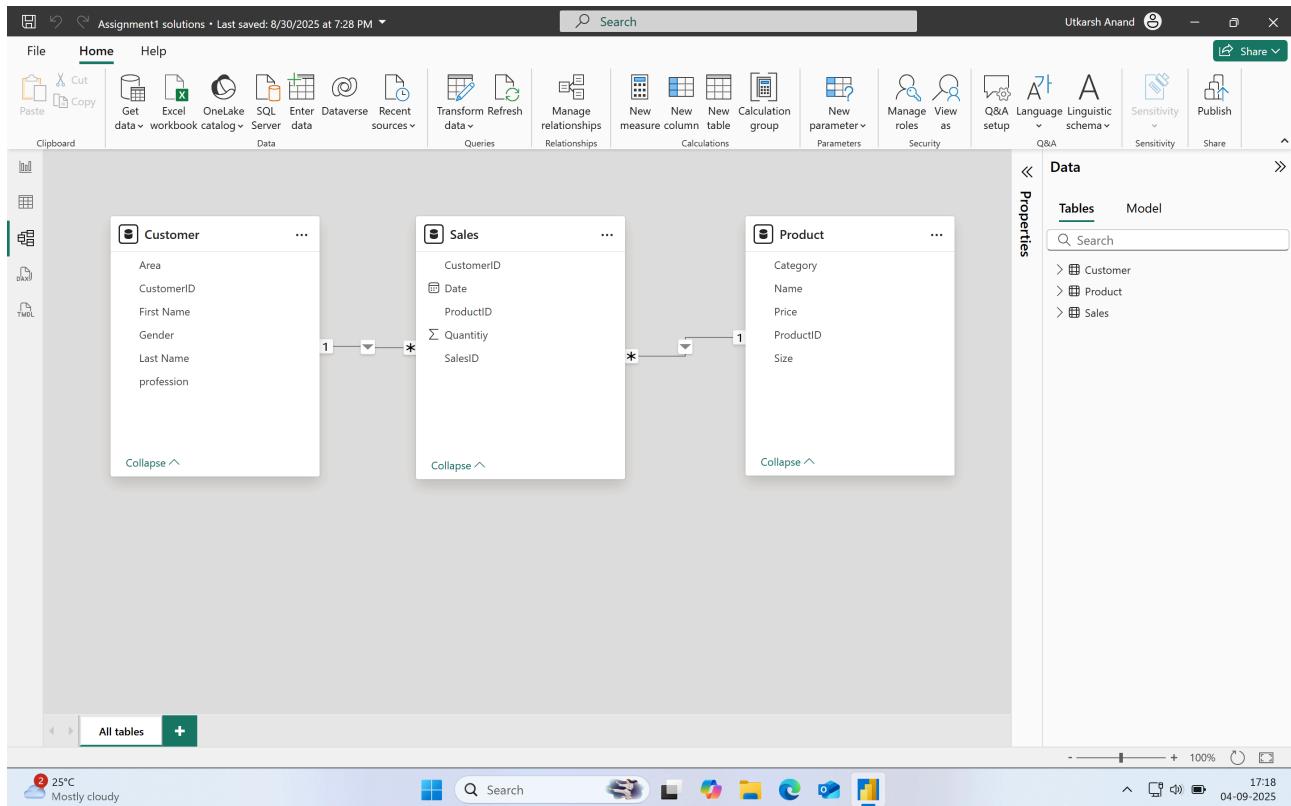
Part 2: Standardize ‘Size’ (Product) — fix inconsistent values

The screenshot shows the Microsoft Power Query Editor interface. A 'Replace Values' dialog box is open, prompting the user to replace the value 'medium' with 'Medium'. The main area displays a table with columns: ProductID, Category, Name, and Size. The 'Size' column contains values like 'Small', 'Large', and 'Medium'. The 'Category' column includes 'DRINKS', 'SNACKS', and 'CHOCOLATES'. The 'Name' column lists various product names. The 'Size' column has a dropdown menu open, showing '1.2 Price' as the current setting. The 'Properties' pane on the right shows the 'Name' is set to 'Product'. The 'Applied Steps' pane at the bottom lists steps such as 'Source', 'Navigation', 'Promoted Headers', 'Changed Type', 'Removed Duplicates', 'Rounded Off', 'Uppercased Text', and 'Capitalized Each Word'. The status bar at the bottom right indicates 'PREVIEW DOWNLOADED AT 15:50'.

Steps taken to identify and replace inconsistent values in the ‘Size’ column of the Product dataset to ensure uniformity:

1. In Power Query Editor, select the Product table → click the Size column.
2. Quick normalization: Transform tab → Format ▾ → Capitalize Each Word (converts “medium”, “SMALL” to “Medium”, “Small”).
3. Targeted cleanup (if needed): Home → Replace Values to convert any remaining variants (e.g., Med → Medium, Lrg → Large).
4. Review the Value Distribution (column header profiling if enabled) to ensure only the intended forms remain.
5. Close & Apply.

Task 4 – Modeling & Deduplication



Part 1: Create relationships via 'CustomerID' and 'ProductID'

Steps taken to create relationships between the tables using 'CustomerID' and 'ProductID' as keys:

1. In Power BI Desktop, click the Model view icon (left sidebar diagram).
2. Drag Customer[CustomerID] onto Sales[CustomerID] → confirm Cardinality: One to many (1:*)
Cross filter: Single, Make this relationship active: On → OK.
3. Drag Product[ProductID] onto Sales[ProductID] → same settings as above → OK.
4. Verify both relationships show **1 → *** from Customer/Product to Sales.

Part 2: Remove duplicates in Customer and Product

The screenshot shows the Power BI Desktop interface in the 'Model' view. The 'Queries [3]' pane on the left lists 'Customer', 'Product', and 'Sales'. The 'Customer' table is selected. The main area displays the 'Sales' table with the following data:

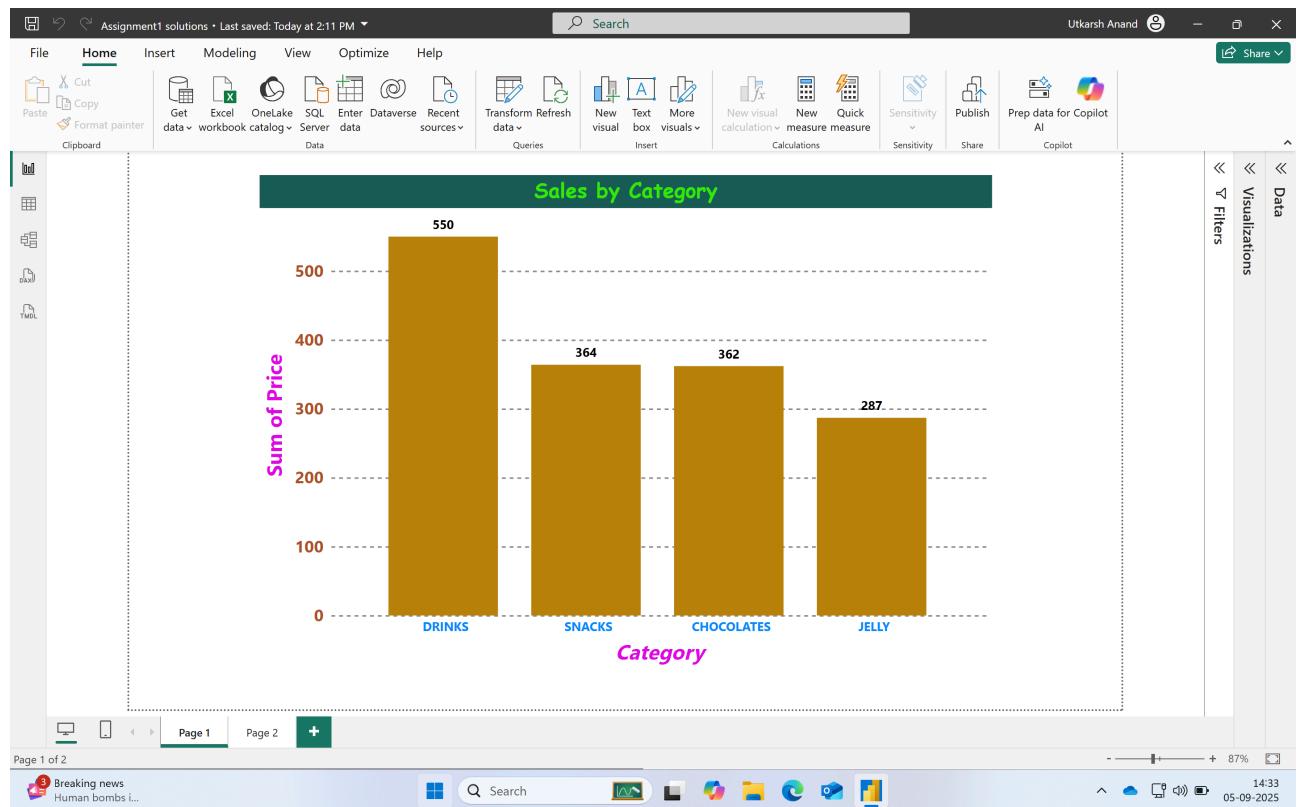
CustomerID	Customer	Gender	Area	profession
C0001	Sujata Mohanty	Male	middle	Retired
C0002	Suraj Rajput	Male	east	unemployment
C0003	Pramod Bhavsar	Male	east	profession
C0004	Satish Ojha	Male	west	self-employed
C0005	Sintu Kumar	Male	middle	Retired
C0006	Krutika Shelar	Male	middle	unemployment
C0007	Arjun Shaw	Male	east	profession
C0008	Shrikant Badge	Female	west	self-employed
C0009	Jitender Kumar	Male	south	Retired
C0010	Dharmendar Rana	Male	middle	unemployment
C0011	Adnan Soukat	Female	south	profession
C0012	Sheetal Nishad	Male	middle	self-employed
C0013	Monika Pawar	Female	east	Retired
C0014	Meena Mourya	Male	east	unemployment
C0015	Ashu Sharma	Male	west	profession
C0016	Harivansh Gautam	Male	middle	self-employed
C0017	Vini Saini	Female	middle	Retired
C0018	Anand Singh Rajput	Male	east	unemployment
C0019	Jaishri Saxena	Male	west	profession
C0020	Virender Sroha	Male	south	self-employed
C0021	Shrikant Badge	Female	middle	Retired
C0022	Harivansh Gautam	Male	south	unemployment
C0023	Sourav Maity	Male	middle	profession
C0024	Abir Sarkar	Female	east	self-employed
C0025	Sandeep Awral	Male	east	Retired
C0026	Karan Kapoor	Male	west	unemployment
C0027	Kishor Panara	Female	middle	profession
C0028	Manpreet Kaur	Male	middle	self-employed

The 'Query Settings' pane on the right shows the 'Customer' table has been renamed and includes steps for removing duplicates. The status bar at the bottom indicates 5 columns, 33 rows, and a preview download date of 04-09-2025.

Steps taken to clean data by removing any duplicate entries in the Customer and Product tables:

1. Open Power Query Editor.
2. Customer table:
 - Select the CustomerID column (or select multiple columns if the uniqueness rule demands).
 - Home tab → Remove Rows ▼ → Remove Duplicates.
3. Product table:
 - Select ProductID → Home → Remove Rows ▼ → Remove Duplicates.
4. Confirm the Applied Steps reflect the duplicate removal, then Close & Apply.

Task 5 — Visualizations



Part 1: Sales by Category (Total Sales / TotalPrice)

Steps taken to create and format a chart showing total sales (TotalPrice) by product category, with colors, title, and data labels:

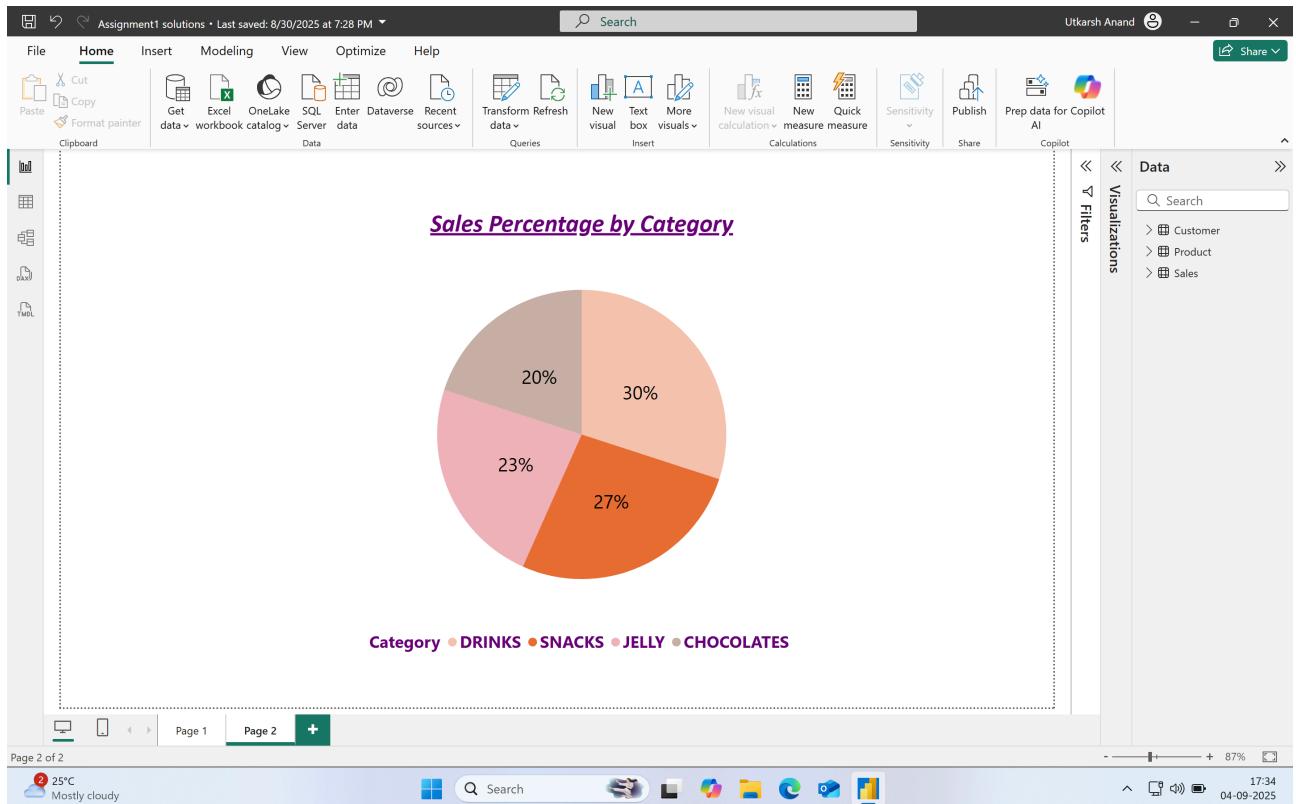
1. Switch to Report view (left sidebar report icon).
2. In the Visualizations pane, click Clustered column chart (or Stacked column chart if you prefer).
3. In Fields:
 - Drag Product[Category] to X-axis.
 - Drag Sales[TotalPrice] to Y-axis and ensure Summarization = Sum.
 - If you don't have a TotalPrice field: create a quick measure:
 - Modeling ribbon → New measure → enter

Total Sales := SUMX(Sales, Sales[Price] * Sales[Quantity])

• Use Total Sales in Y-axis instead.

4. Format the chart: click the chart → Format (paint-roller icon) →
 - General → Title: On → Title text: "Total Sales by Category (₹)".
 - Data labels: On → Display units: None → Decimal places: 0.
 - Columns/Colors: set distinct colors for readability.
 - X-axis/Y-axis: adjust Text size and Gridlines as needed.
5. Resize and align the visual neatly on the canvas.

Part 2: Sales Percentage by Category (Price)



Steps taken to create a chart showing the percentage share of sales (Price) for each product category, kept compact and clear:

1. In Report view, choose Pie chart (or Donut chart for a compact ring layout).
2. In Fields:
 - Drag Product[Category] to Legend.
 - Drag Sales[Price] to Values (summarization: Sum).
 - (If allowed by your rubric, using TotalPrice here is usually more accurate for share of sales.)

3. Format: Format (paint-roller) →

- Data labels: On → Label style: Data value and percent of total.
- Legend: On → position to Bottom or Right to save space.
- Title: On → “Sales % by Category”.
- Detail labels: reduce Text size so the visual remains compact and easy to read.

4. Position the chart next to the bar/column chart for a clean dashboard look.

Prepared By,

Utkarsh Anand