Power BI

Prepare the Data

AI and Data Analytics

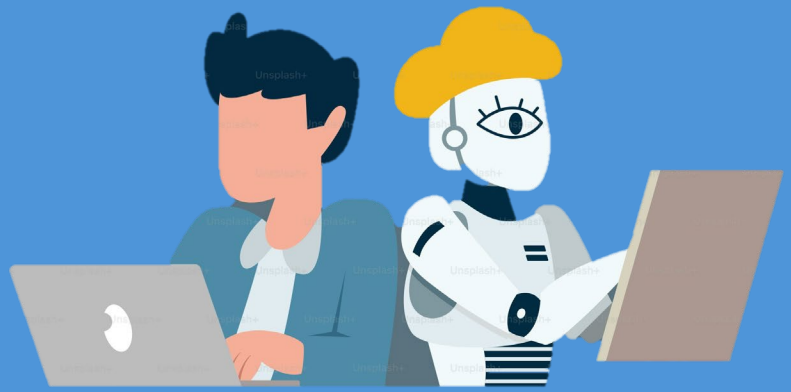
Power BI

Prepare the Data

- Get data from data sources
- Profile and clean the data
- Transform and load the data

Prepare the Data

Get data from data sources



Prepare the Data

Get data from data sources

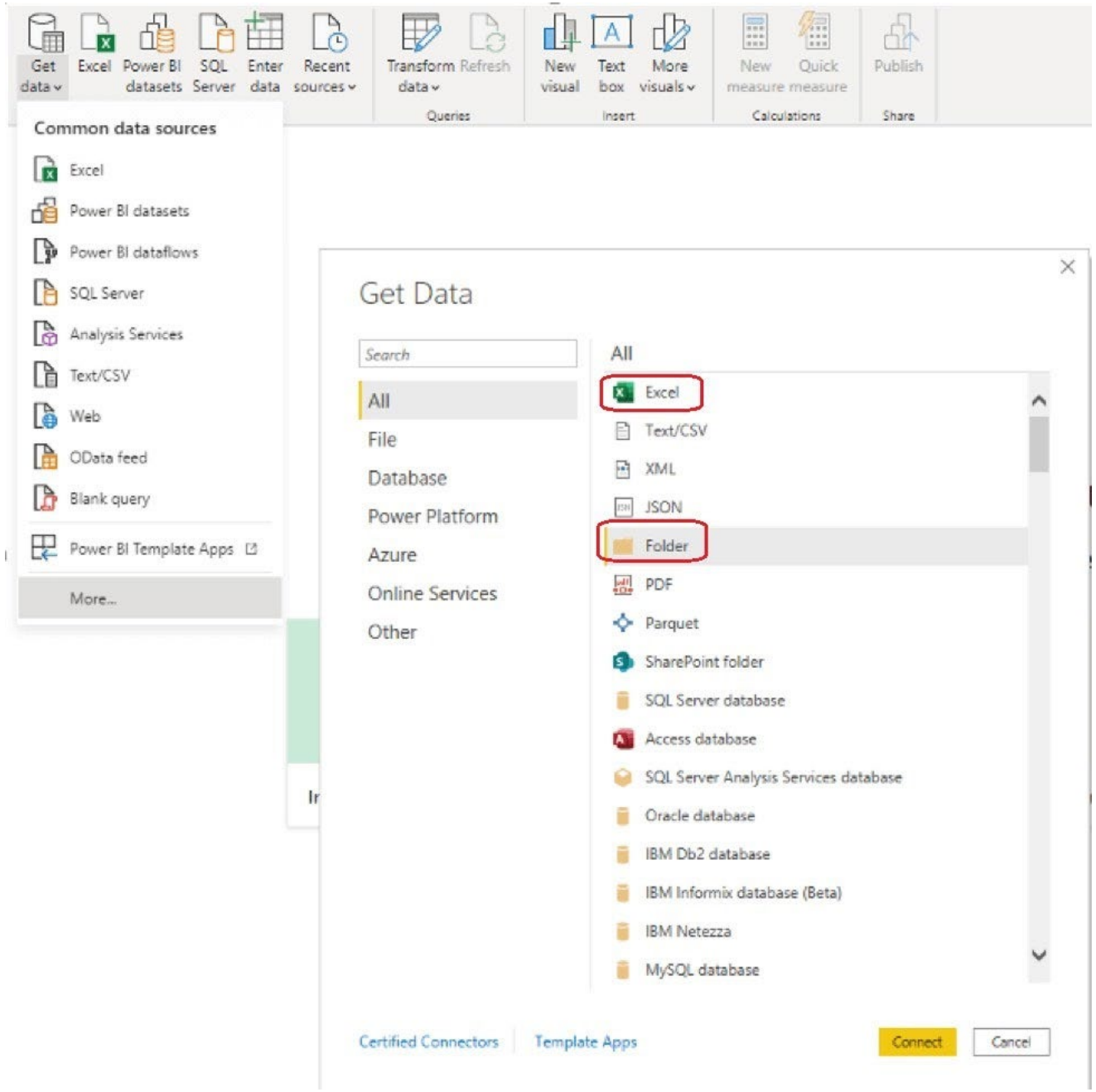
- Identify and connect to a data source
- Change data source settings, including credentials, privacy levels, and data source locations
- Choose between **DirectQuery**, **Import**, and **Dual** mode
- Clean and modify parameters

Get Data

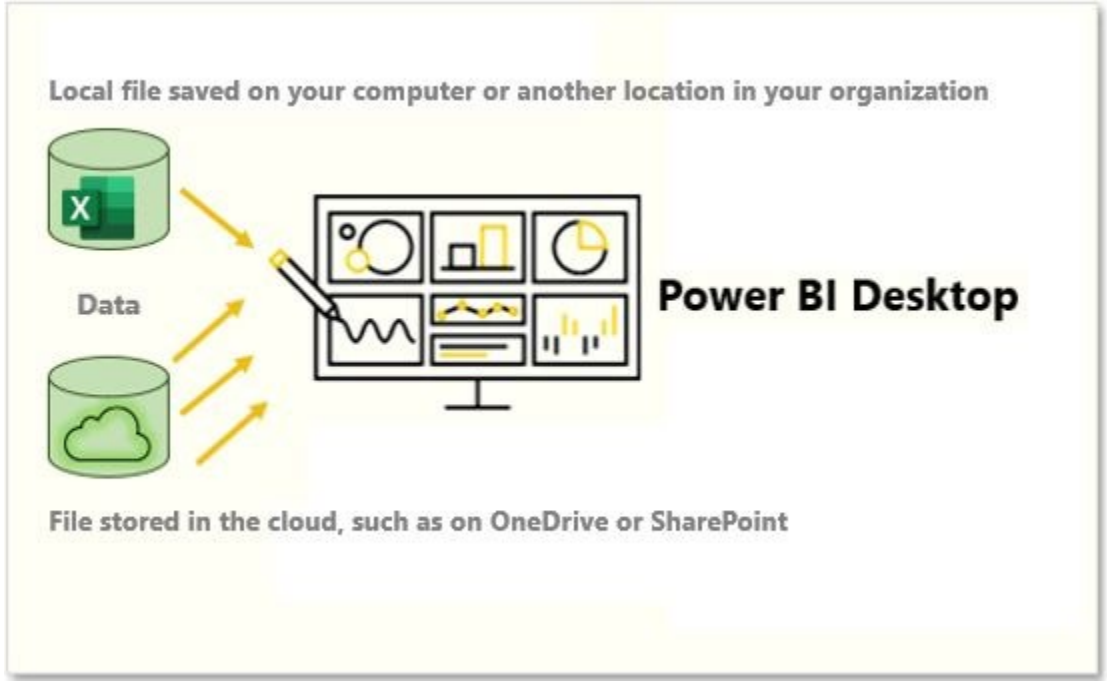
With PowerBI, you can get data from

1. Files
2. Relational data
3. NoSQL

etc.



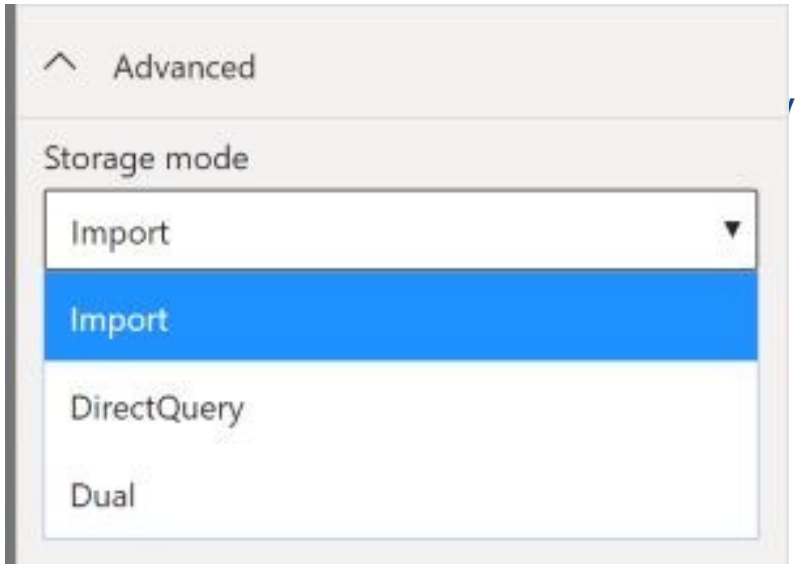
Get Data from File



- Using a **cloud option** such as *OneDrive* or *SharePoint Team Sites* is the most effective way to keep your file and your semantic model, reports, and dashboards in Power BI in-sync.
- Saving files on a **local computer** is a suitable option if your data doesn't change regularly,

Storage Mode

Import, DirectQuery and Dual



Import mode

- Data is cached. Create a local copy of your models from your data source.
- Data refreshes can be scheduled or on-demand.
- The default mode for new Power BI reports

DirectQuery mode

- Data isn't cached. You don't want to save local copies of your data
- With a direct connection to the data source, ensures that you're always getting the most up-to-date data, and that all security requirements are satisfied.
- Query the specific tables and the required data will be retrieved from the underlying data source.
- This mode is also suited for when you have large semantic models to pull data from. Instead of slowing down performance, you can query the data, solving data latency issues as well.

Dual (Composite mode)

- Identify some data to be directly imported and other data that must be queried.
- Any table that is brought into your report is a product of both Import and DirectQuery modes.
- Allows Power BI to choose the most efficient form of data retrieval.

Import vs Direct Query

Feature	Power BI Import	Power BI DirectQuery
Data storage	Data copied into Power BI's memory.	Queries data directly from the source.
Real-time access	No, requires scheduled refreshes.	Yes, displays data in real-time.
Volume of data	Subject to Power BI's data size limits.	Not limited by Power BI's data size limits, but dependent on source system performance.
Data security	Data duplicated in Power BI, uses its security model.	Utilizes the source system's security model, no data duplication.
Data transformation	Extensive within Power BI.	Limited, depends on source system capabilities.
System compatibility	Broad, supports many data sources.	Dependent on Power BI's ability to establish live connections.
Frequency of refresh	Requires manual or scheduled refreshes.	Not applicable, as data is always up-to-date.

Dual – Support Table Relationships

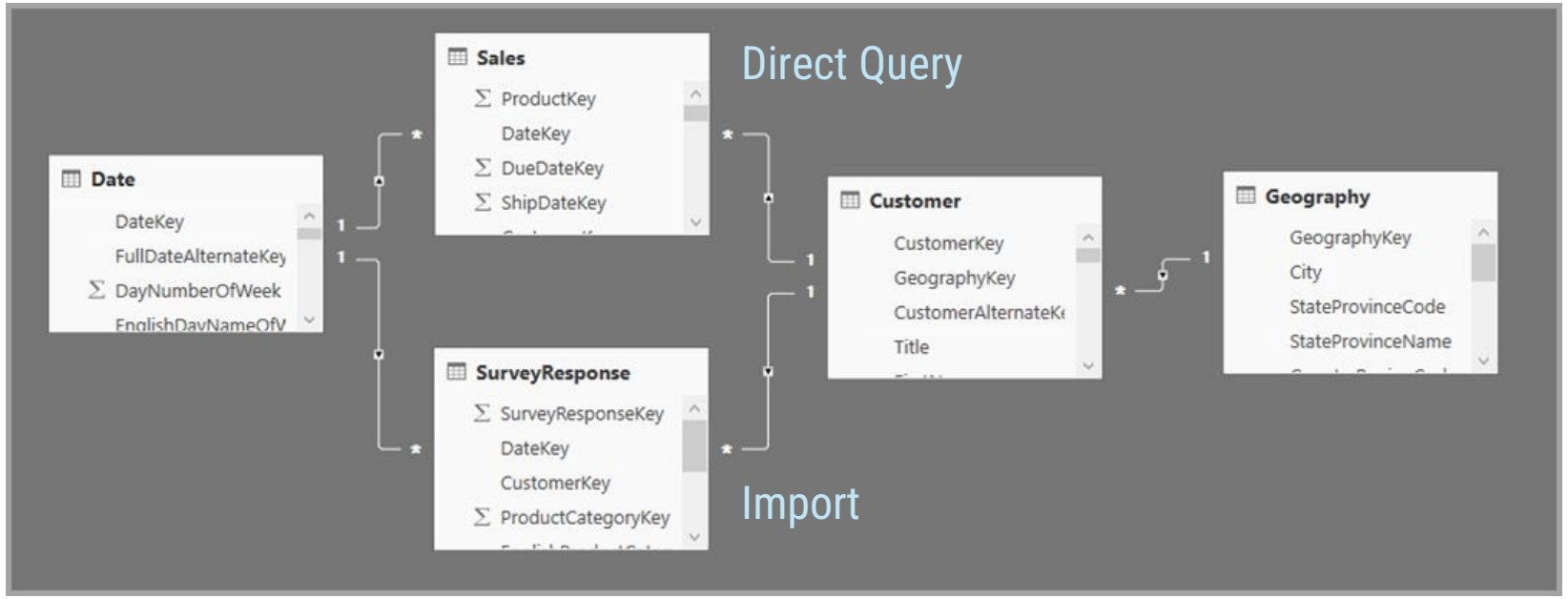


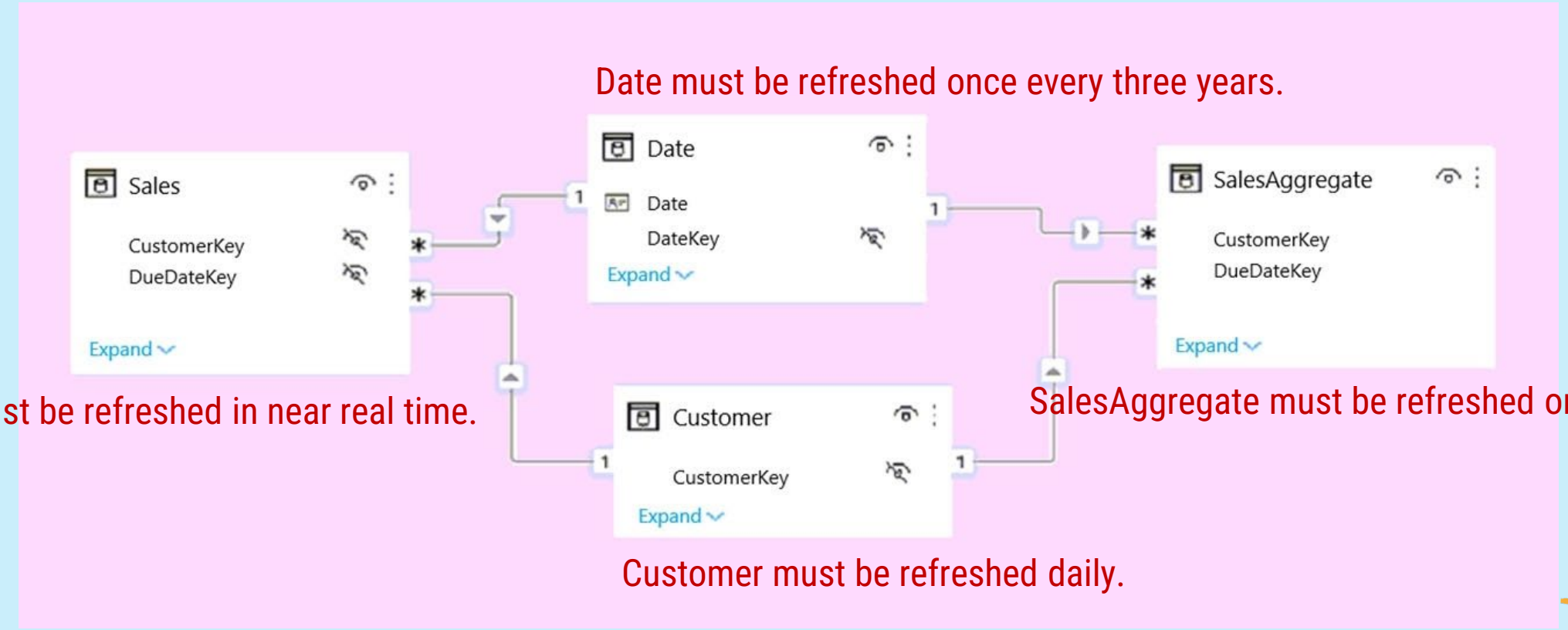
Table	Storage mode
Sales	DirectQuery
SurveyResponse	Import
Date	Dual
Customer	Dual
Geography	Dual

If **Sales** is Direct Query, and **SurveyResponse** is Import,
Then **Date**, **Customer** and **Geography** must be Dual.

Knowledge Check

Import, DirectQuery and Dual

You plan to create the Power BI model shown in the exhibit.



Import, DirectQuery and Dual

The data has the following refresh requirements:
- Customer must be refreshed daily.
- Date must be refreshed once every three years.
- Sales must be refreshed in near real time.
- SalesAggregate must be refreshed once per week.

You need to select the storage modes for the tables. The solution must meet the following requirements:
- Minimize the load times of visuals.
- Ensure that the data is loaded to the model based on the refresh requirements.

Which storage mode should you select for each table?

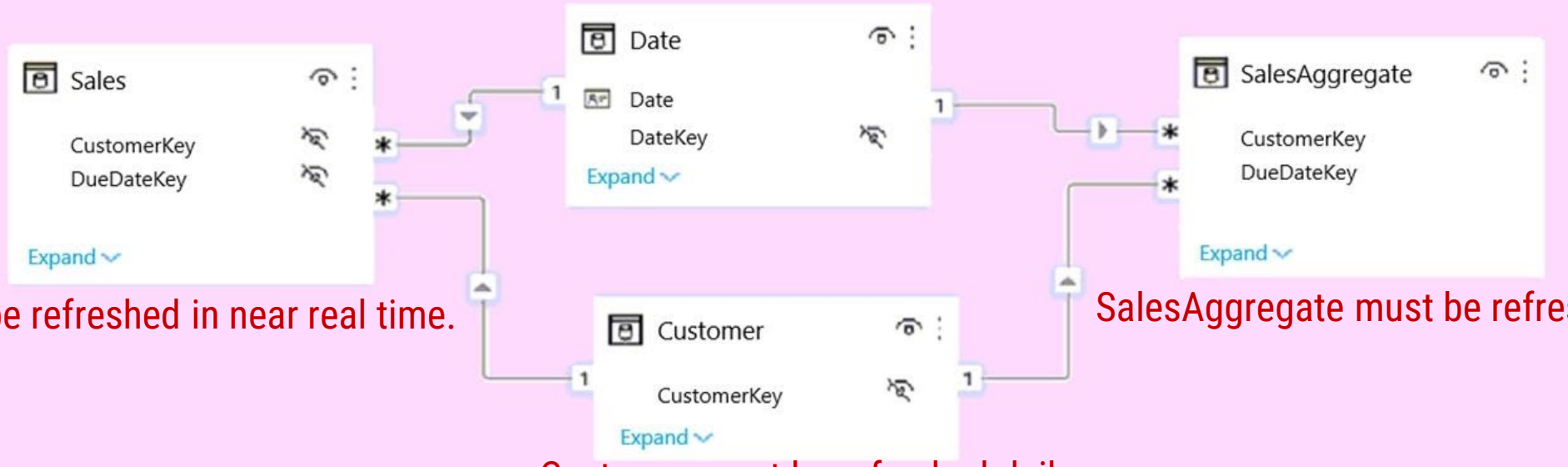
Answer Area

Customer:	DirectQuery / Dual / Import
Date:	DirectQuery / Dual / Import
Sales:	DirectQuery / Dual / Import
SalesAggregate:	DirectQuery / Dual / Import

Knowledge Check

Import, DirectQuery and Dual

Date must be refreshed once every three years.



Sales must be refreshed in near real time.

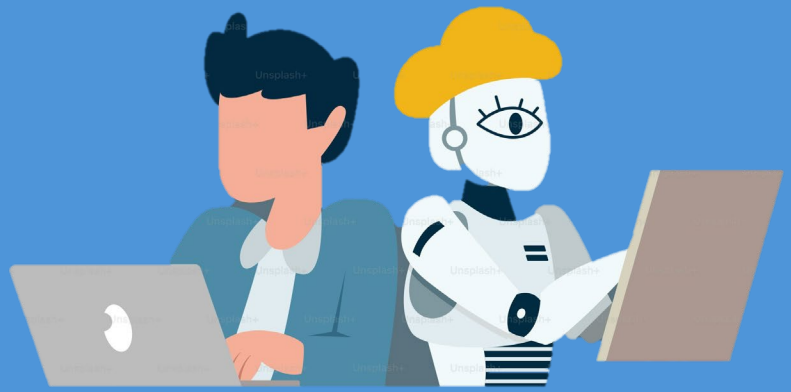
SalesAggregate must be refreshed once per wee

Customer must be refreshed daily.

Both **Date** and **Customer** has relationship with both Sales and SalesAggregate tables. Set to **"Dual"** to support performance for **DirectQuery** (**Sales**) and Import(**SalesAggregate**)

Prepare the Data

Profile and Clean the data



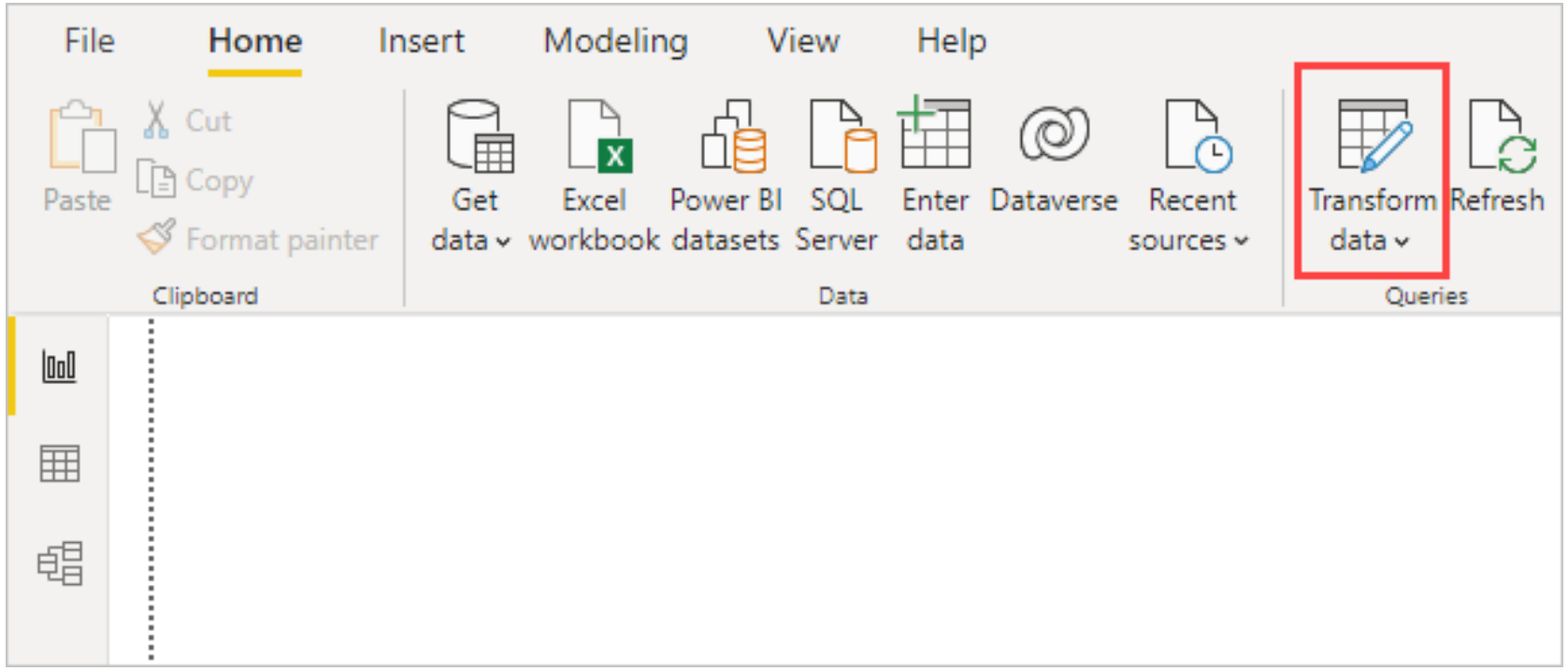
Prepare the Data

Profile and Clean the data

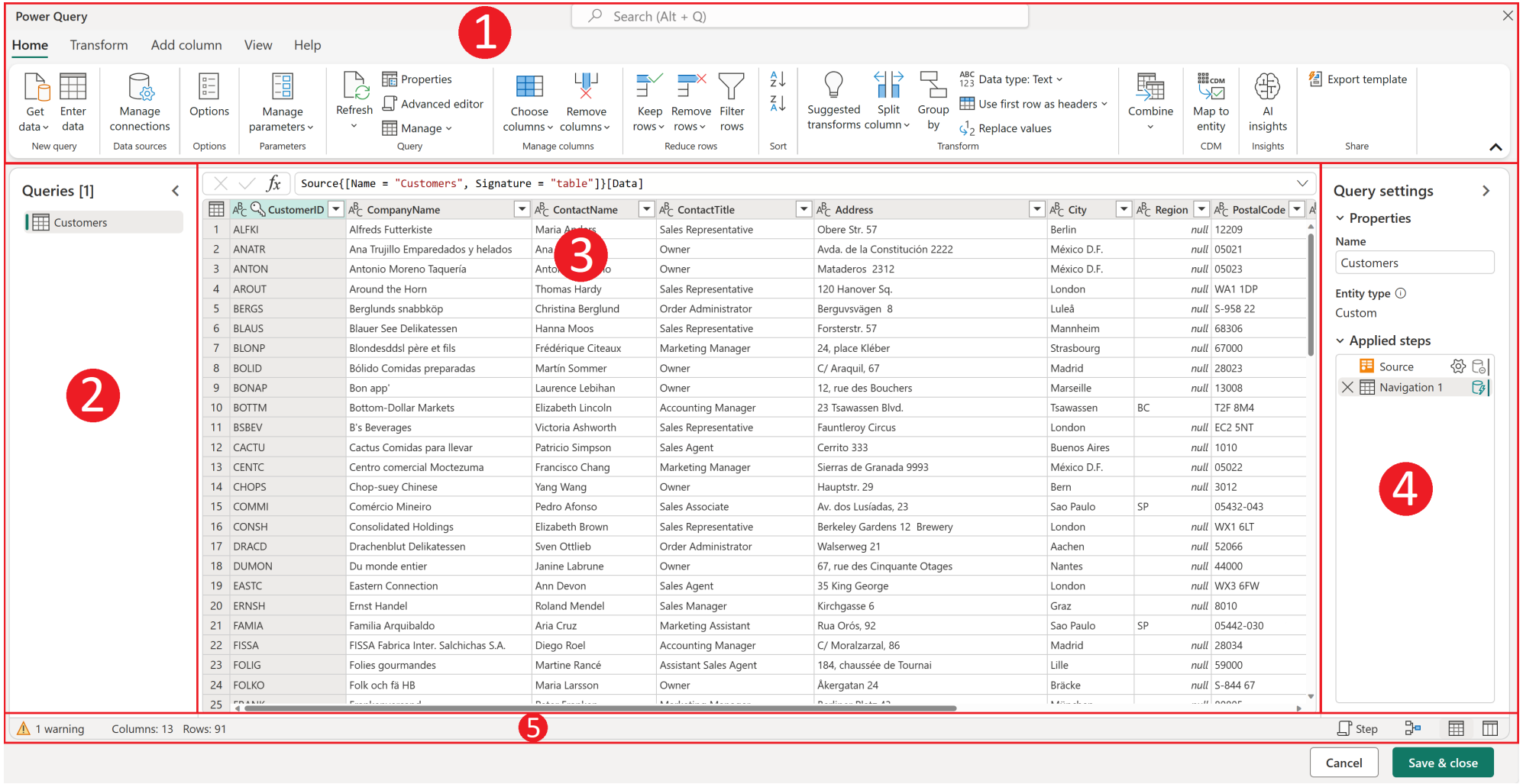
- Evaluate data, including **data statistics** and **column properties**
- Resolve **inconsistencies**, **unexpected** or **null** values, and **data quality** issues
- Resolve data import errors

Power Query

To get to Power Query Editor, select **Transform data** from the **Home** tab of Power BI Desktop.

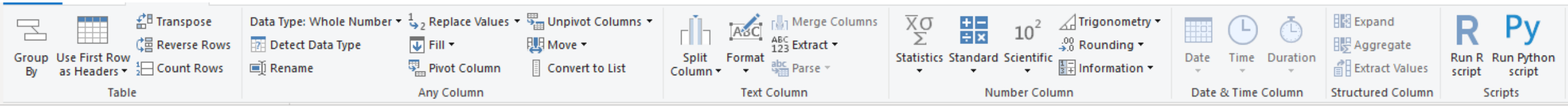


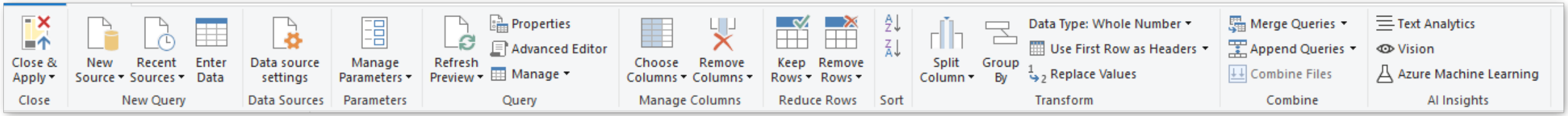
Power Query

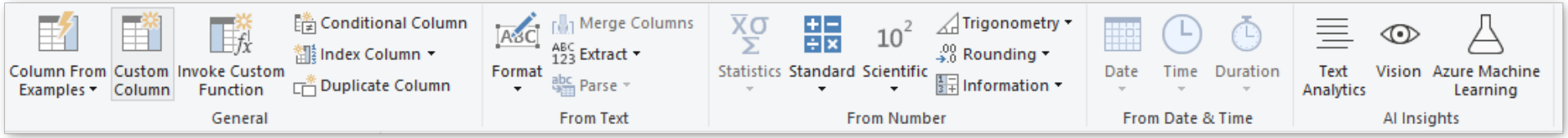

Power Query
Search (Alt + Q)
Home
Transform
Add column
View
Help
Get data
Enter data
Manage connections
Options
Manage parameters
Refresh
Properties
Advanced editor
Manage
Choose columns
Remove columns
Keep rows
Remove rows
Filter rows
Suggested transforms
Split column
Group by
Data type: Text
Use first row as headers
Replace values
Combine
Map to entity
AI insights
Export template
New query
Data sources
Options
Parameters
Query
Manage columns
Reduce rows
Sort
Transform
CDM
Insights
Share
Queries [1]
Customers
Source{[Name = "Customers", Signature = "table"]}[Data]
CustomerID
CompanyName
ContactName
ContactTitle
Address
City
Region
PostalCode
1 ALFKI Alfreds Futterkiste Maria Anders Sales Representative Obere Str. 57 Berlin null 12209
2 ANATR Ana Trujillo Emparedados y helados Ana Trujillo Owner Avda. de la Constitución 2222 México D.F. null 05021
3 ANTON Antonio Moreno Taquería Antonio Moreno Owner Mataderos 2312 México D.F. null 05023
4 AROUT Around the Horn Thomas Hardy Sales Representative 120 Hanover Sq. London null WA1 1DP
5 BERGS Berglunds snabbköp Christina Berglund Order Administrator Berguvsvägen 8 Luleå null S-958 22
6 BLAUS Blauer See Delikatessen Hanna Moos Sales Representative Forsterstr. 57 Mannheim null 68306
7 BLONP Blondesddsl père et fils Frédérique Citeaux Marketing Manager 24, place Kléber Strasbourg null 67000
8 BOLID Bólido Comidas preparadas Martín Sommer Owner C/ Araquil, 67 Madrid null 28023
9 BONAP Bon app' Laurence Lebihan Owner 12, rue des Bouchers Marseille null 13008
10 BOTTM Bottom-Dollar Markets Elizabeth Lincoln Accounting Manager 23 Tsawassen Blvd. Tsawassen BC T2F 8M4
11 BSBEV B's Beverages Victoria Ashworth Sales Representative Fauntleroy Circus London null EC2 5NT
12 CACTU Cactus Comidas para llevar Patricio Simpson Sales Agent Cerrito 333 Buenos Aires null 1010
13 CENTC Centro comercial Moctezuma Francisco Chang Marketing Manager Sierras de Granada 9993 México D.F. null 05022
14 CHOPS Chop-suey Chinese Yang Wang Owner Hauptstr. 29 Bern null 3012
15 COMMI Comércio Mineiro Pedro Afonso Sales Associate Av. dos Lusíadas, 23 Sao Paulo SP 05432-043
16 CONSH Consolidated Holdings Elizabeth Brown Sales Representative Berkeley Gardens 12 Brewery London null WX1 6LT
17 DRACD Drachenblut Delikatessen Sven Ottlieb Order Administrator Walserweg 21 Aachen null 52066
18 DUMON Du monde entier Janine Labrune Owner 67, rue des Cinquante Otages Nantes null 44000
19 EASTC Eastern Connection Ann Devon Sales Agent 35 King George London null WX3 6FW
20 ERNSH Ernst Handel Roland Mendel Sales Manager Kirchgasse 6 Graz null 8010
21 FAMIA Familia Arquibaldo Aria Cruz Marketing Assistant Rua Orós, 92 Sao Paulo SP 05442-030
22 FISSA FISSA Fabrica Inter. Salchichas S.A. Diego Roel Accounting Manager C/ Moralzarzal, 86 Madrid null 28034
23 FOLIG Folies gourmandes Martine Rancé Assistant Sales Agent 184, chaussée de Tournai Lille null 59000
24 FOLKO Folk och fä HB Maria Larsson Owner Åkergatan 24 Bräcke null S-844 67
Query settings
Properties
Name
Customers
Entity type
Custom
Applied steps
Source
Navigation 1
1 warning
Columns: 13 Rows: 91
Step
Cancel
Save & close
1
2
3
4
5

Power Query

- Power Query is a data transformation and data preparation engine.

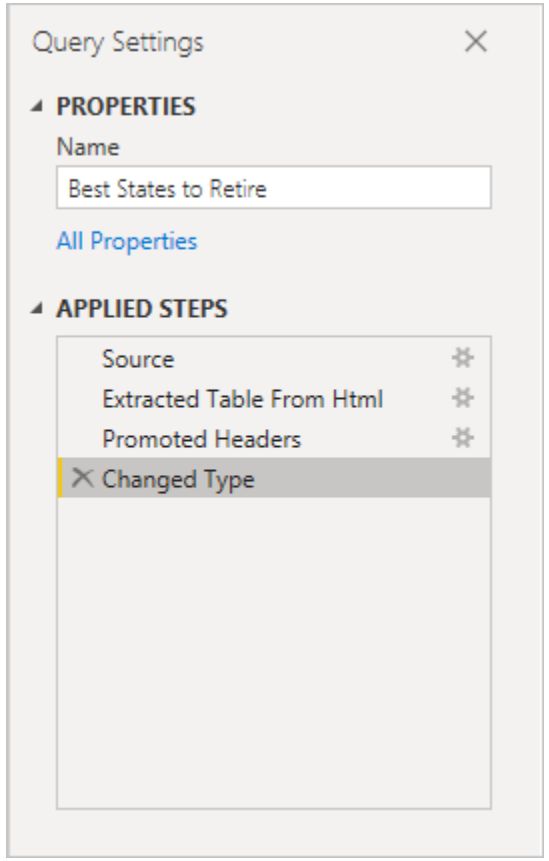






Power Query

Applied Steps



In Power Query

- underlying data *isn't* changed. Rather, Power Query Editor adjusts and shapes its view of the data.

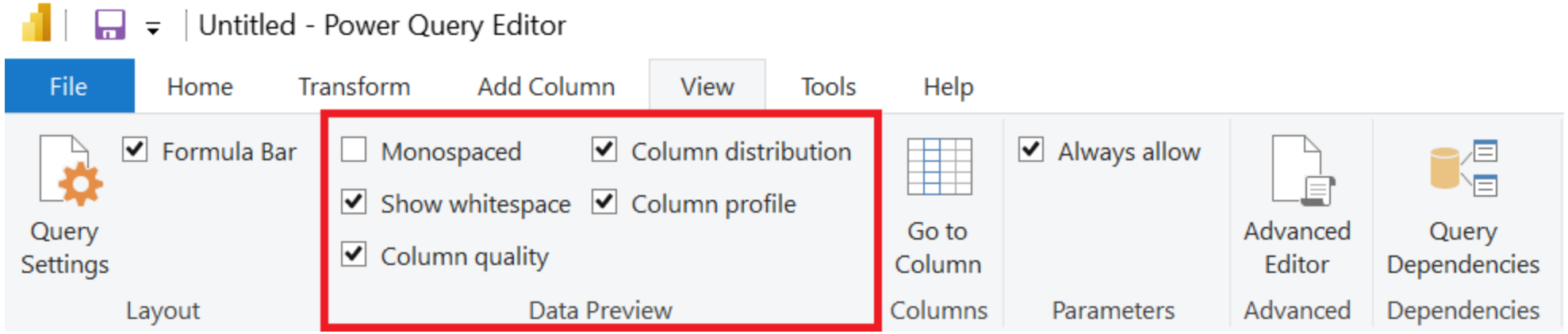
Go to the right pane **Query Settings** > **Applied Steps**

- You can rename steps, delete steps, or reorder the steps

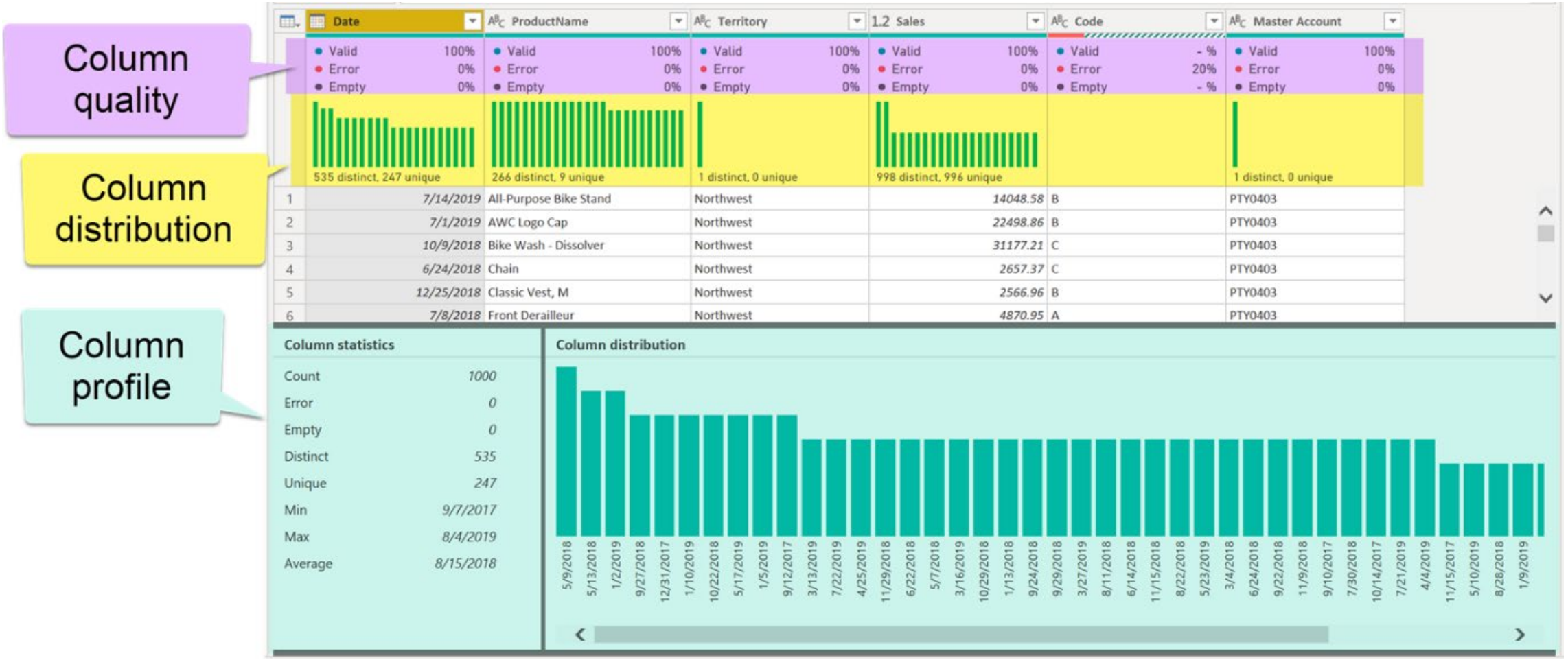
Data profile tools

The data profiling tools provide new and intuitive ways to clean, transform, and understand data in Power Query Editor. They include:

- Column quality
- Column distribution
- Column profile



Profile Data

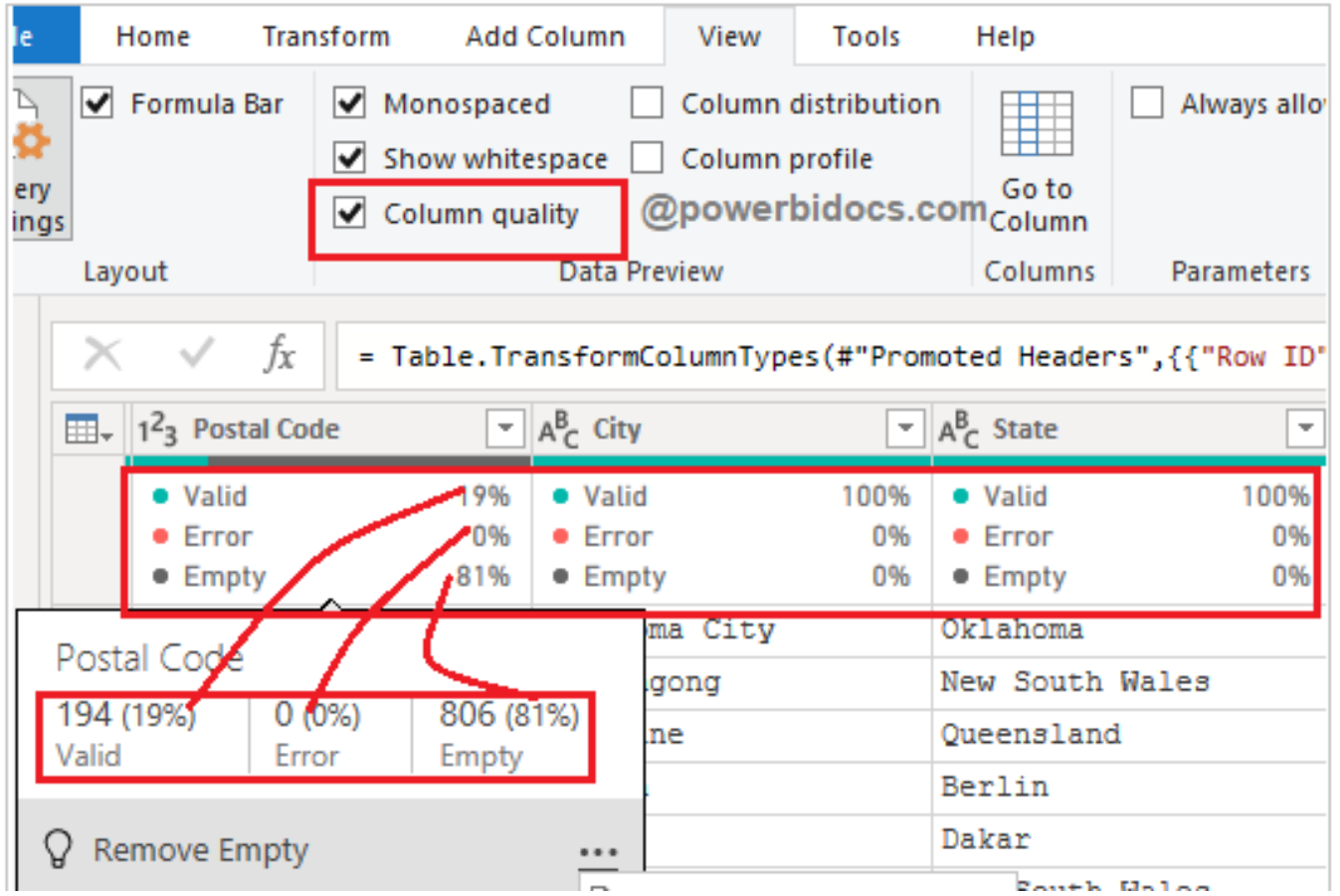


Column Quality

Shows number and percentage of

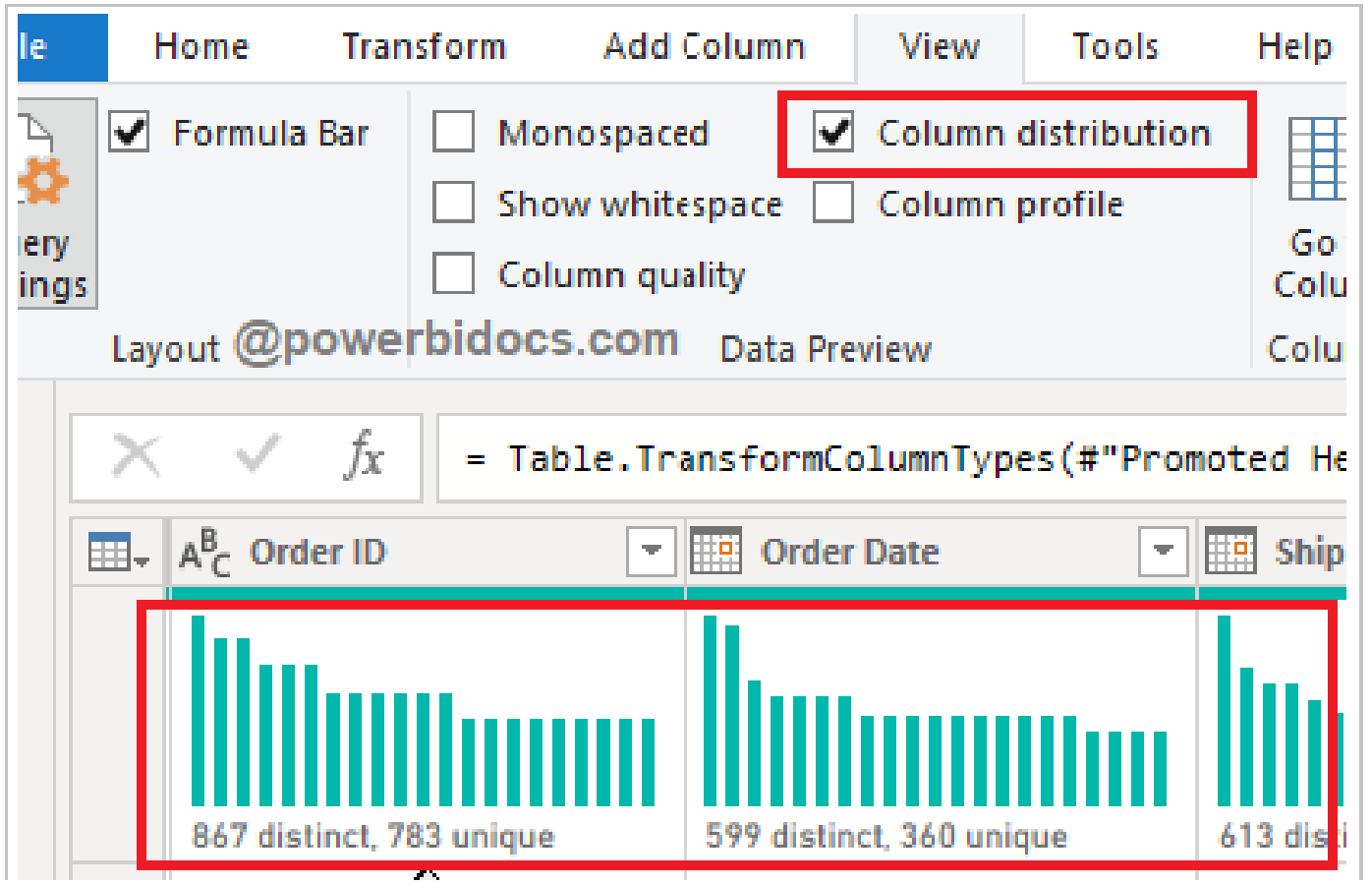
- **Valid**
- **Error**
- **Empty**

for *each column*



Column Distribution

- **Distinct** values
  - total number of different values found in a column, including duplicates and null values
- **Unique** values
  - total number of values that only appear once in a column.
  - Do not include duplicates and nulls



Distinct vs Unique

Example

Column – Computer Brand

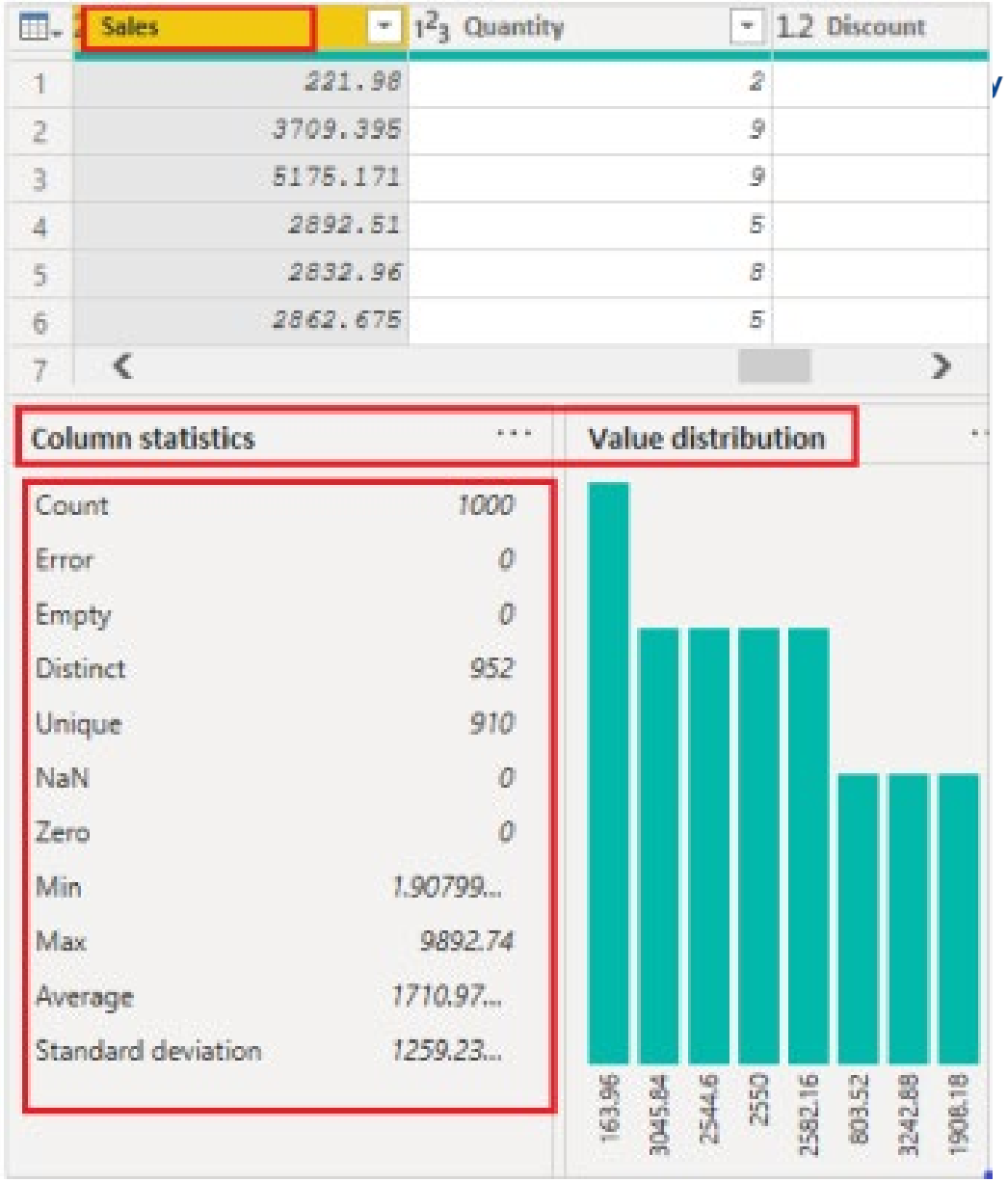
- HP – 12
- Lenovo – 8
- Asus – 4
- Mac – 1
- Dell - 1

No. of Distinct values – 5

No. of Unique values - 2

Column Profile

- **Column profile** gives you a more in-depth look into the statistics within the columns for the first 1,000 rows of data.
- **Column Statistics:** distinct, unique, empty, error, min, max, average, count for each column
- **Value Distribution:** counts for each distinct value in that specific column.



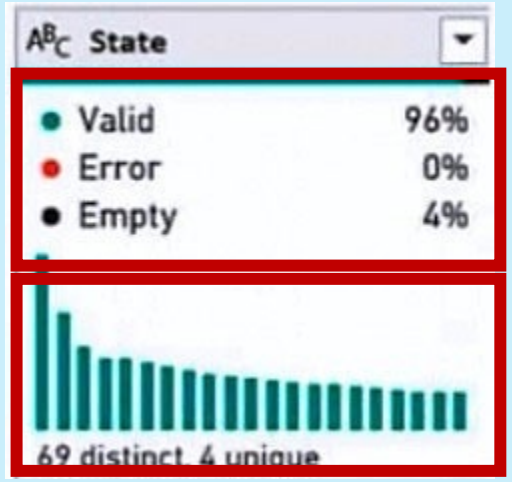
Knowledge Check

Data Profiling

You are profiling data by using **Power Query Editor**.
You have a table named **Reports** that contains a column named **State**. The column quality and distribution for the **State** column is shown.

Column Quality

Column Distribution



Answer Area

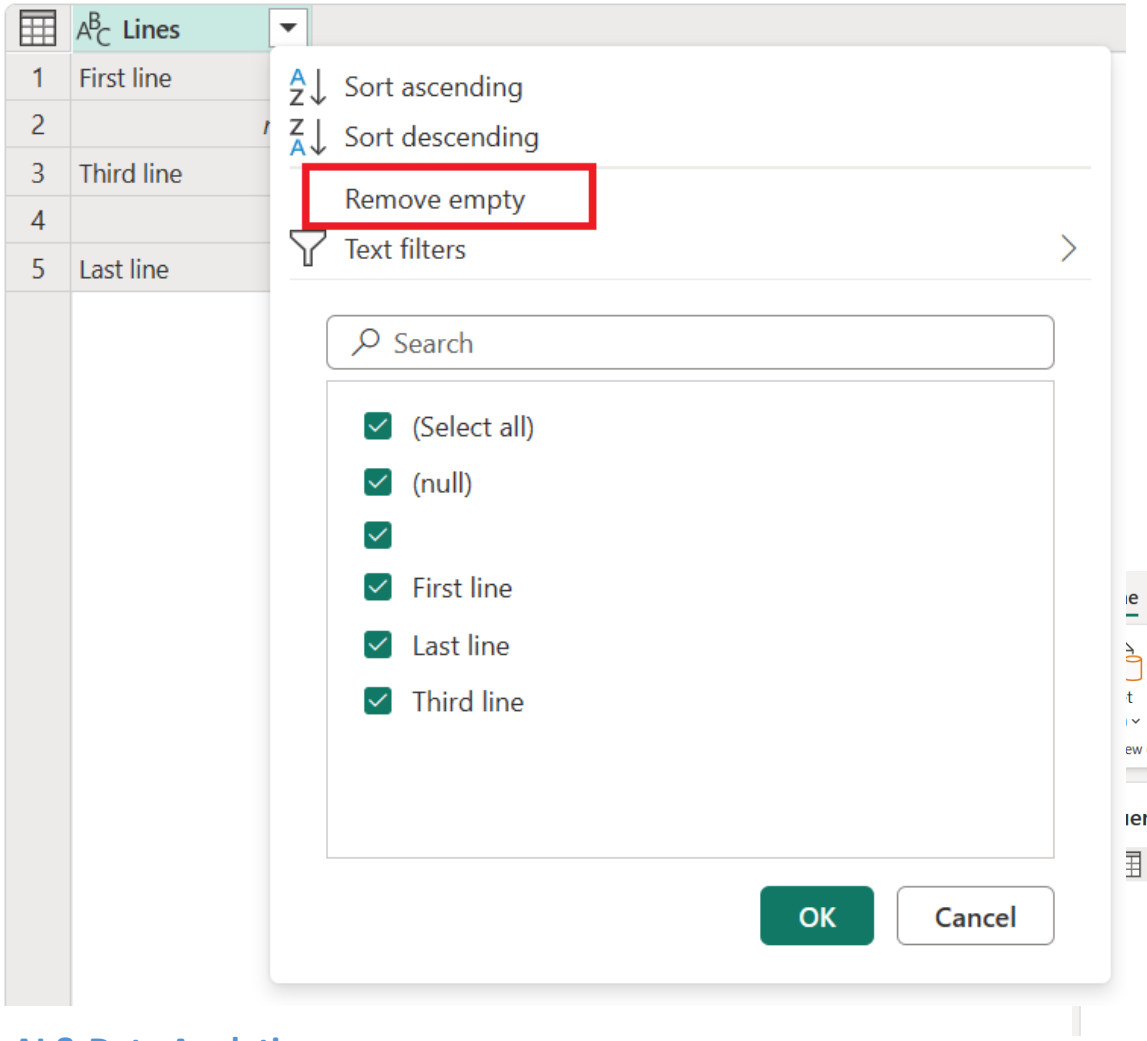
There are **[answer choice]** different values in State including nulls.

4
65
69
73

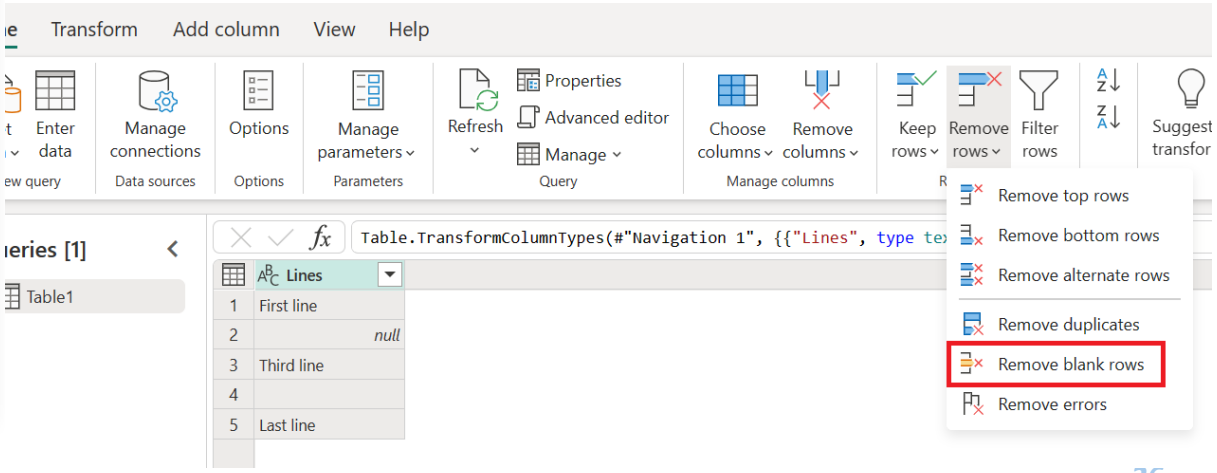
There are **[answer choice]** non-null values that occur only once in State.

4
65
69
73

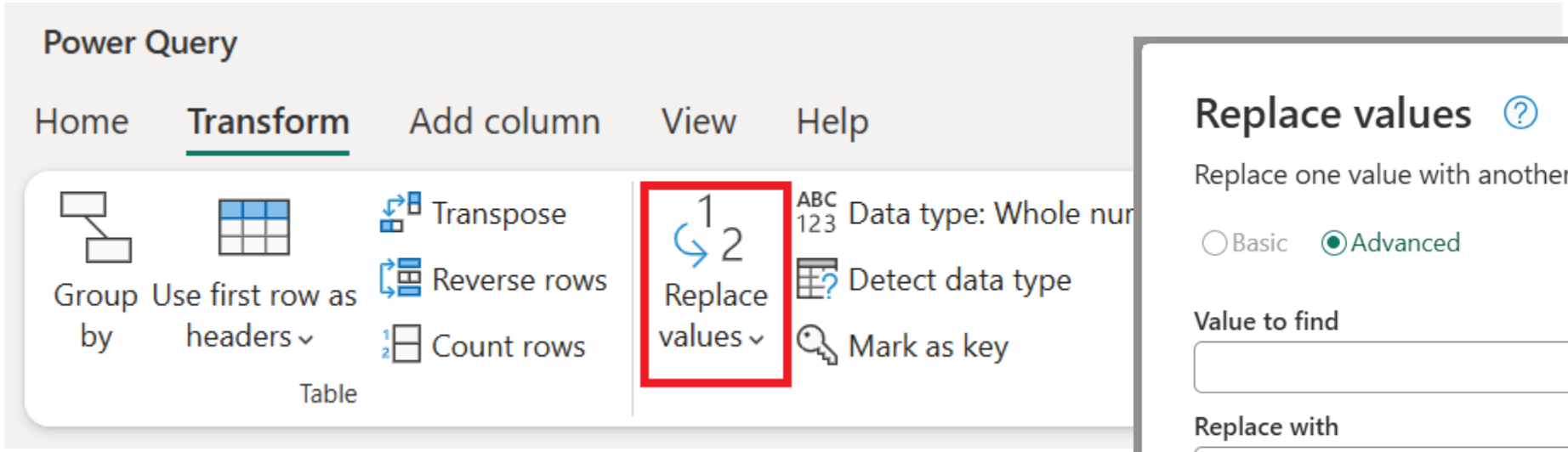
Filter by value

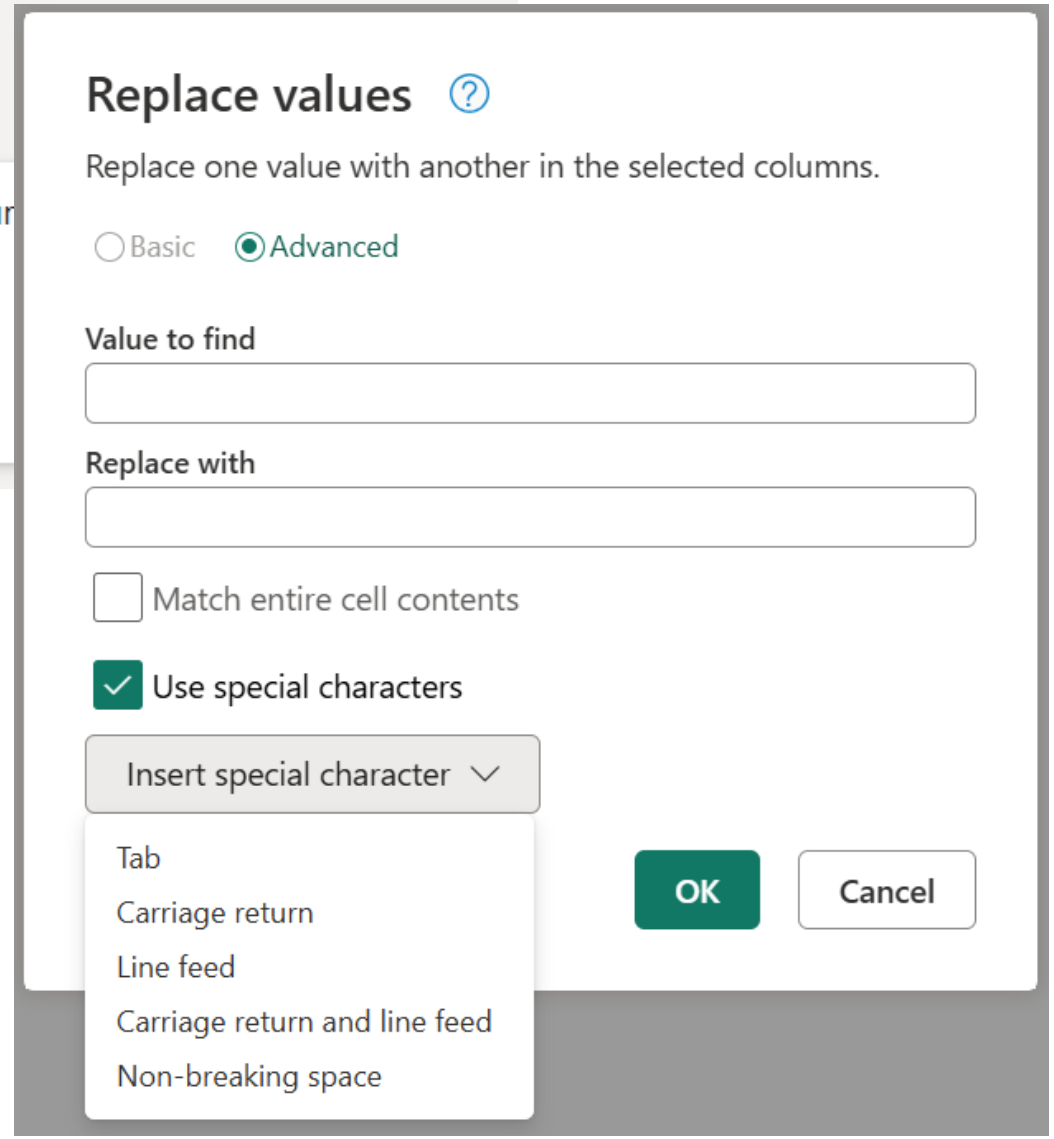


- You can exclude rows by using filter
- On the column **click the arrow** to Sort & filter.
- The **Remove empty** command applies two filter rules to your column. The first rule gets rid of any null values. The second rule gets rid of any blank values.
- A null value is a specific value in the Power Query language that represents no value.



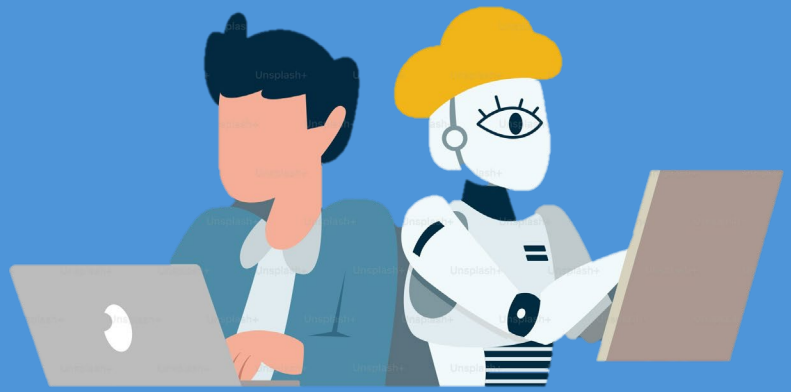
Replace Values





Prepare the Data

Transform and Load the Data



Prepare the Data

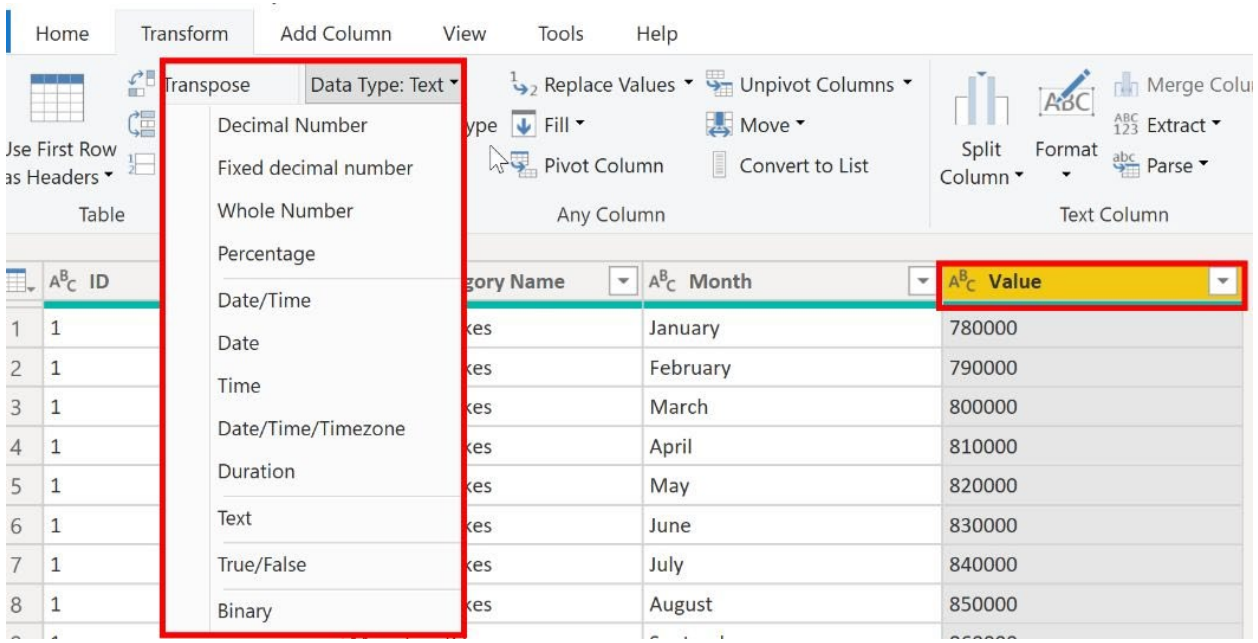
Transform and load the data

- Select appropriate **column data types**
- Create and **transform columns**
- **Group** and **aggregate** rows
- **Pivot**, **unpivot** and **transpose** data
- Convert semi-structured data to a table
- Create **fact tables** and **dimension tables**
- **Merge** and **append** queries
- Identify and create appropriate **keys** for relationships
- Identify when to use **reference** or **duplicate** queries and the resulting impact
- Configure **data loading** for queries

Change Column data types

You can change the data type of a column in two places:

1. **Power Query Editor**
2. **Power BI Desktop Report view** by using the column tools.



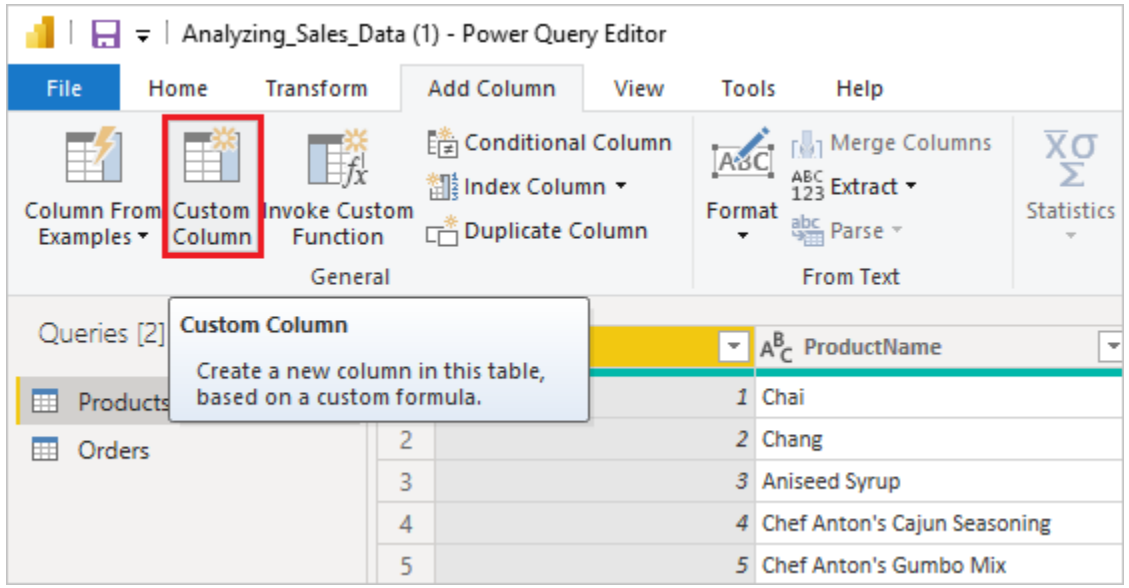
It is best to change the data type in the Power Query Editor before you load the data.

Change Column data type

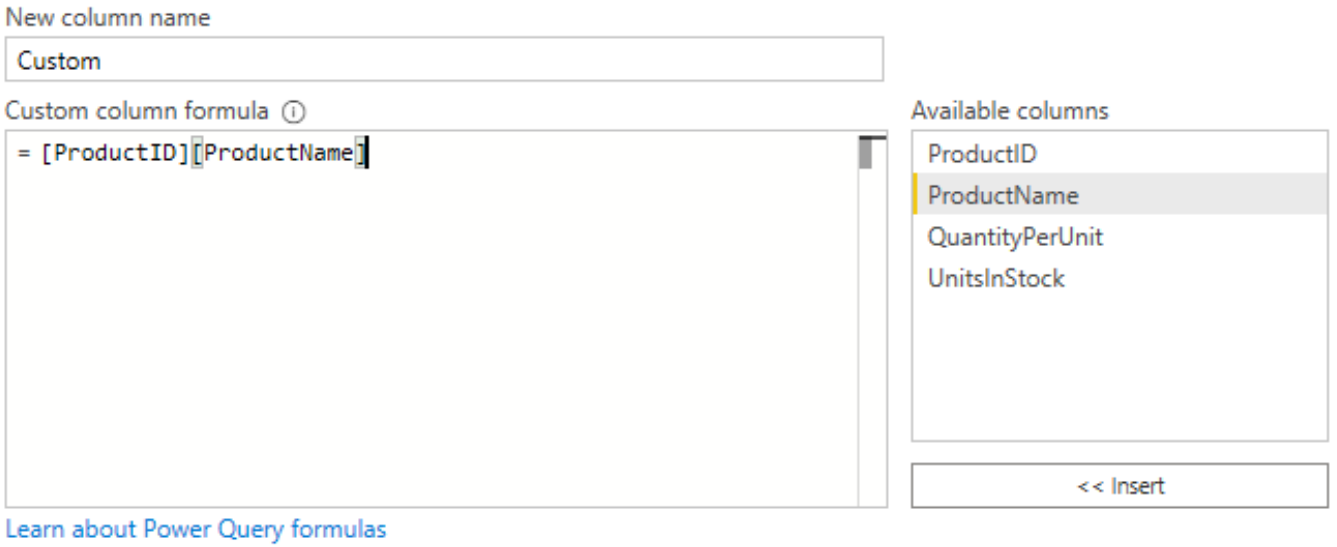
- By default, automatic data type detection occurs when connect to data such as Excel, CSV and databases.

Icon	Data type	Description
ABC 123	Any	Indicates no explicit data type definition.
	Binary	A binary value, such as Y/N or 0/1.
$	Fixed decimal number	Has a fixed format of four digits to the right and 19 digits to the left. Also known as the Currency type.
	Date	A date with no time and having a zero for the fractional value.
	Date/Time	A date and time value stored as a Decimal Number type.
	Date/Time/TimeZone	A UTC Date/Time with a time-zone offset.
	Duration	A length of time converted into a Decimal Number.
	True/False	A Boolean value of either True or False.
1.2	Decimal number	A 64-bit (eight-byte) floating point number.
%	Percentage	A fixed decimal number with a mask to format as a percentage.
ABC	Text	Strings, numbers, or dates represented in a text format.
	Time	A time with no date having no digits to the left of the decimal place.
123	Whole Number	A 64-bit (eight-byte) integer value.

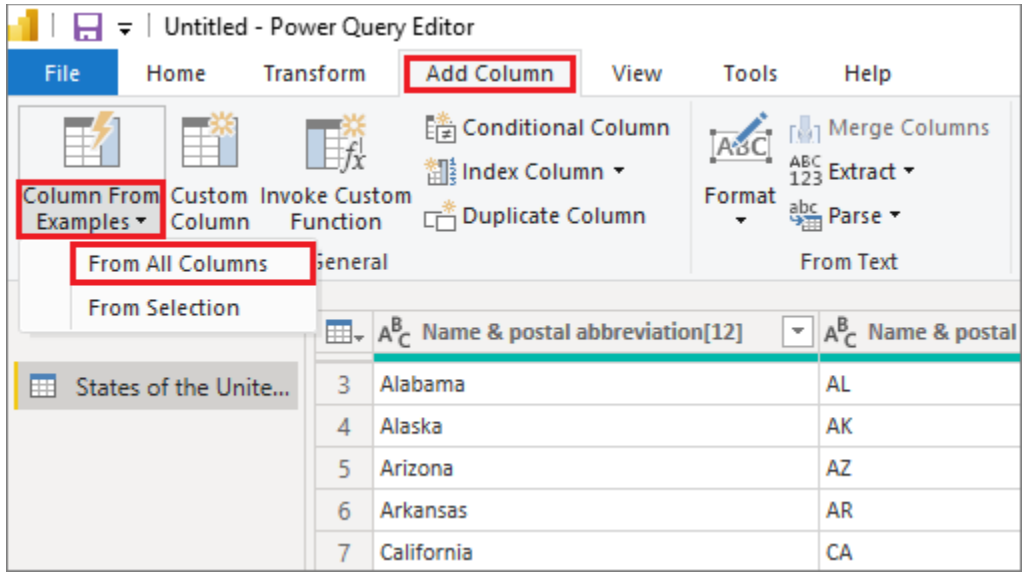
Custom Column



M Formula is used to create Custom Column



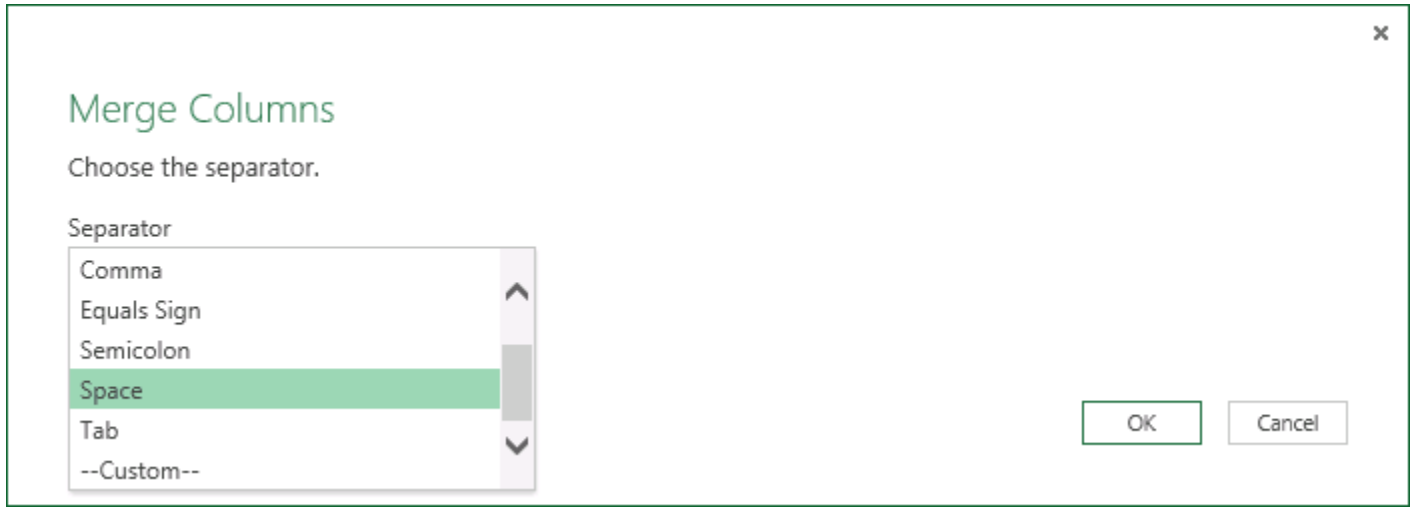
Column from Examples



- As you type your example in the new column, Power BI shows a preview of the rest of the column, based on the transformations it creates
- Examples:
  - Text between delimiters
  - Replace text
  - Day of Week Name
  - Age
  - Round up/down

Merge Column

- Select two or more columns that you need to merge. To select contiguous columns, press Shift+Click. To select discontiguous columns, press CTRL+Click.
- Select **Transform** > **Merge Columns**.



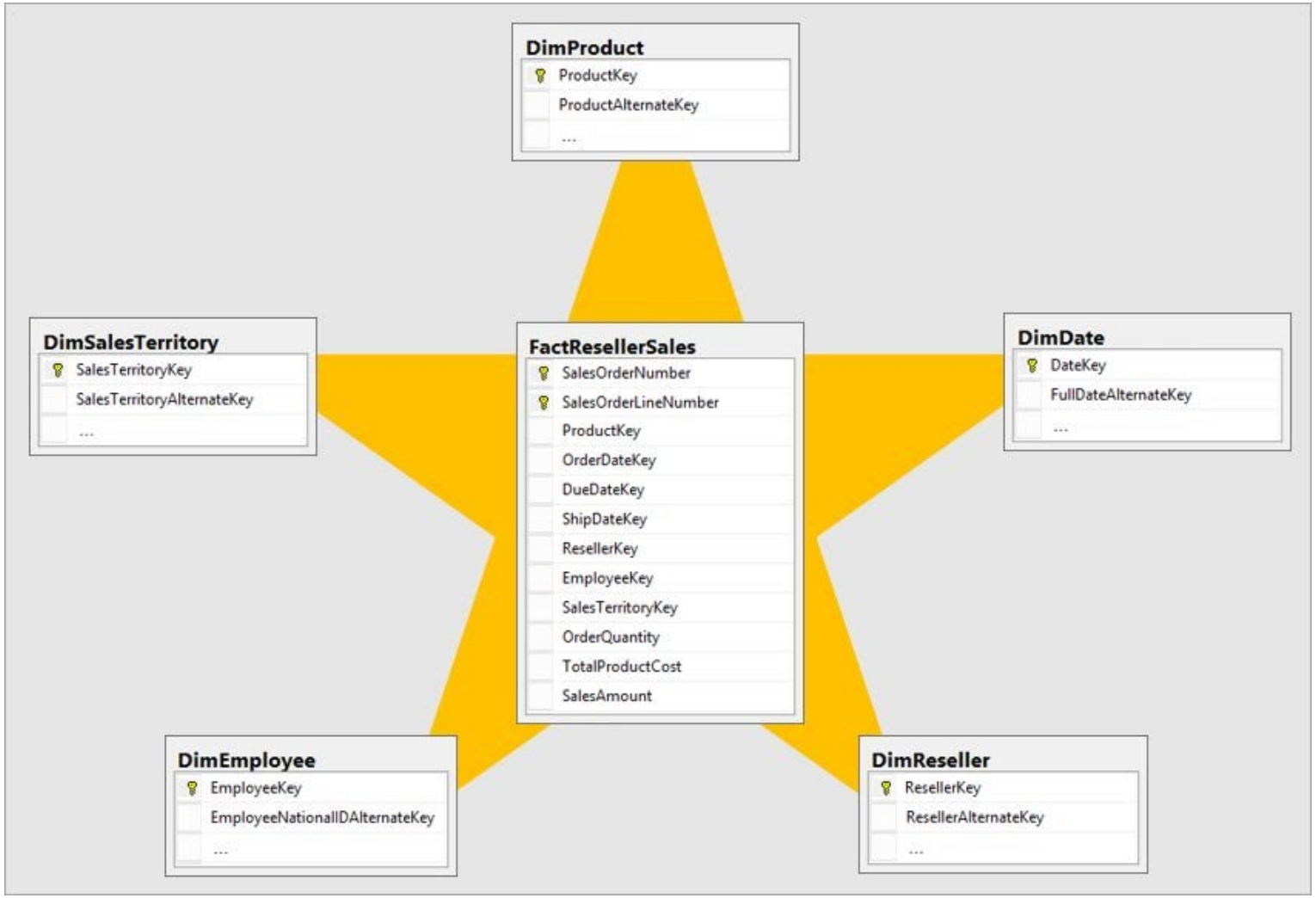
Dimension tables

- Descriptive data

Fact tables

- Values

Star Schema

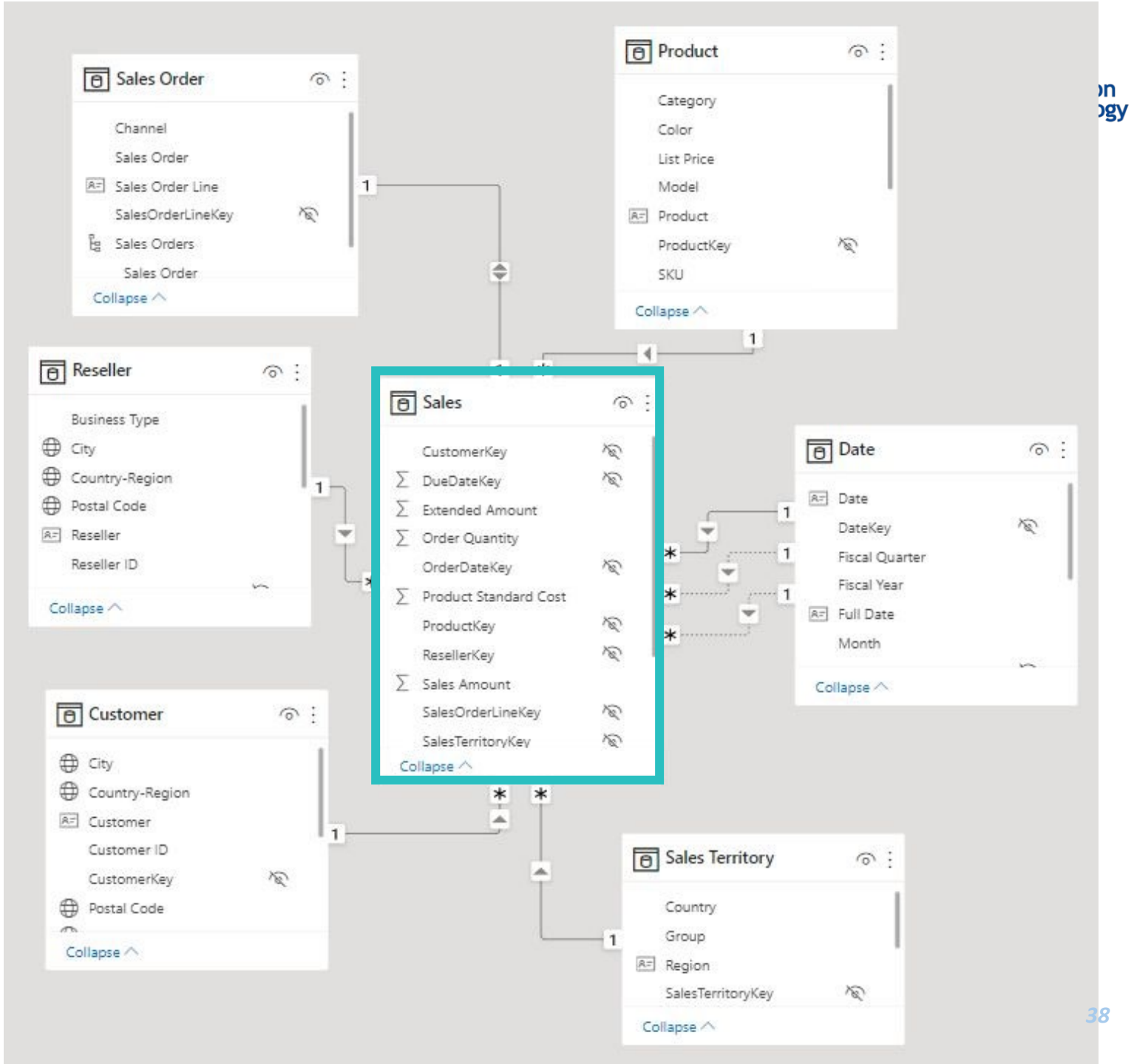

DimProduct
ProductKey
ProductAlternateKey
...
DimSalesTerritory
SalesTerritoryKey
SalesTerritoryAlternateKey
...
FactResellerSales
SalesOrderNumber
SalesOrderLineNumber
ProductKey
OrderDateKey
DueDateKey
ShipDateKey
ResellerKey
EmployeeKey
SalesTerritoryKey
OrderQuantity
TotalProductCost
SalesAmount
DimDate
DateKey
FullDateAlternateKey
...
DimEmployee
EmployeeKey
EmployeeNationalIDAlternateKey
...
DimReseller
ResellerKey
ResellerAlternateKey
...

Star Schema

- An optimal model adheres to *star schema* design principles.
  - a design approach that's commonly used by relational data warehouse designers
  - supports high-performance analytic queries.
- Classifies model tables as either *fact* or *dimension*.
  - **Fact** table forms the center of a star
  - **Dimension** tables, when placed around a fact table, represent the points of the star.

Star Schema

- Sales is the fact table
- All the other are dimension tables



Dimension tables vs Fact tables

- **Dimension tables**
  - describe business entities—the *things* you model. Entities can include products, people, places, and concepts including time itself.
  - Stores descriptive data
  - Contain a relatively small number of rows
- **Fact tables**
  - store observations or events, and can be sales orders, stock balances, exchange rates, temperatures, etc.
  - Stores values
  - Typically contain a large number of rows and continue to grow over time

Transpose Table

- Turn **rows into columns**
- Turn columns into rows

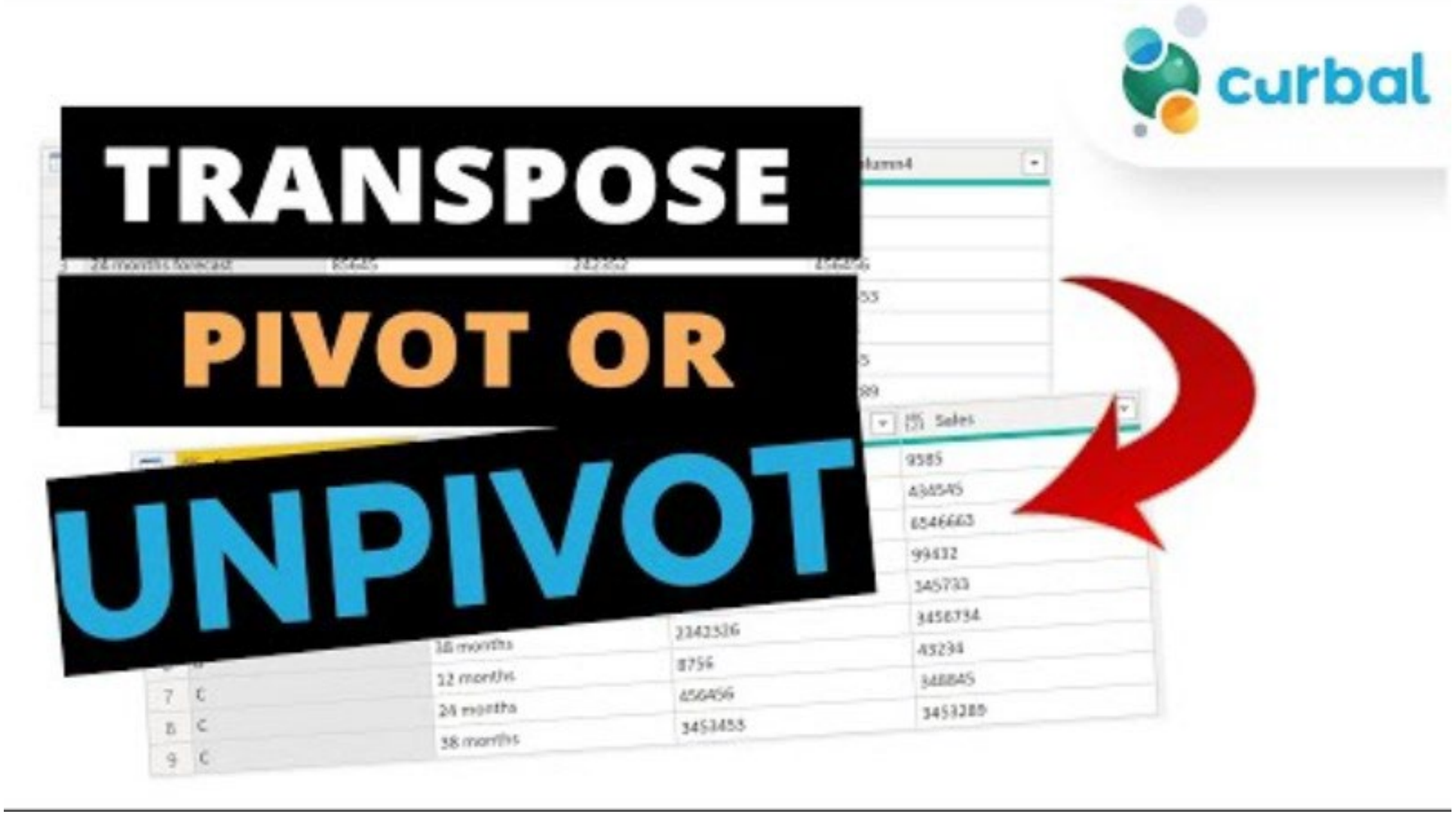
Pivot Column

- **Aggregrate** data

Unpivot Column

- **Flatten** data (Transform nested data into a flat, tabular format

Transpose, pivot or unpivot in Power Query?



Transpose Table

Transpose Table in Power Query rotates your table 90 degrees, turning your rows into columns and your columns into rows.

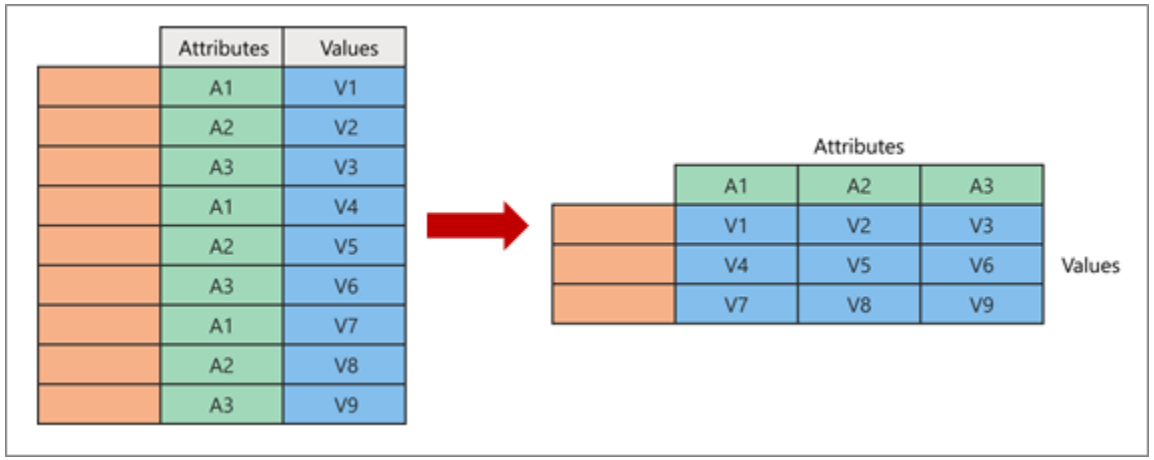
Steps:

1. On the **Transform** tab in the ribbon, select **Transpose**.
2. Select **Use first row as headers**.

	Column1	Column2	Column3	Column4
1	Events	Event 1	Event 2	Event 2
2	Participants	150	450	1250
3	Funds	4000	10000	15000

	Events	Participants	Funds
1	Event 1	150	4000
2	Event 2	450	10000
3	Event 2	1250	15000

Pivot Column

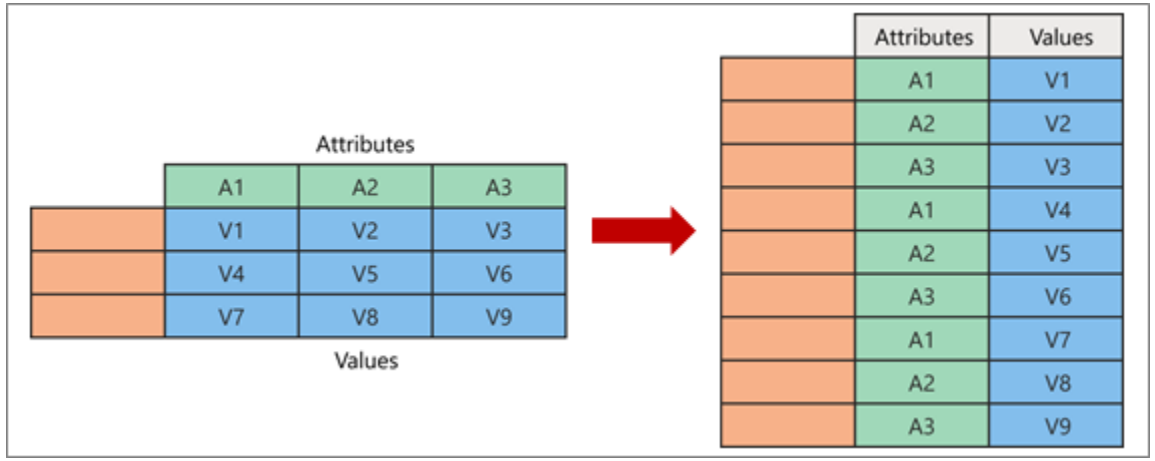


Pivot changes rows into columns

Aggregates matching values in a column to create a new table orientation.

The table is sorted in ascending order by the values in the first column.

Unpivot Columns



Unpivot changes columns into rows

Default Column Headers: *Attributes* and *Values*

Often, data is stored in a **nested** or stacked format, making it difficult to analyze and visualize effectively.

Unpivot helps to flatten and unstack your data for easier analysis.

Unpivot Columns

Steps:

1. Highlight the *2018* and *2019* columns, select the **Transform** tab in Power Query, and then select **Unpivot**. (Unpivot columns other than *Year* column)
2. Rename the *Year* column to *Month*

	Year	2018	2019
1	January	15370	16063
2	February	15950	12161
3	March	13862	14180
4	April	18530	6516
5	May	5203	19395
6	June	5928	19324
7	July	14736	15939
8	August	6243	15390
9	September	15178	17832
10	October	18148	5185
11	November	8014	9299
12	December	19470	14082

	Year	Attribute	Value
	January	2018	15370
	January	2019	16063
	February	2018	15950
	February	2019	12161
	March	2018	13862
	March	2019	14180
	April	2018	18530
	April	2019	6516
	May	2018	5203
0	May	2019	19395
1	June	2018	5928
2	June	2019	19324
3	July	2018	14736
4	July	2019	15939
5	August	2018	6243
6	August	2019	15390
7	September	2018	15178
8	September	2019	17832
9	October	2018	18148
0	October	2019	5185
1	November	2018	8014
2	November	2019	9299
3	December	2018	19470
4	December	2019	14082

Append

- Add **rows**

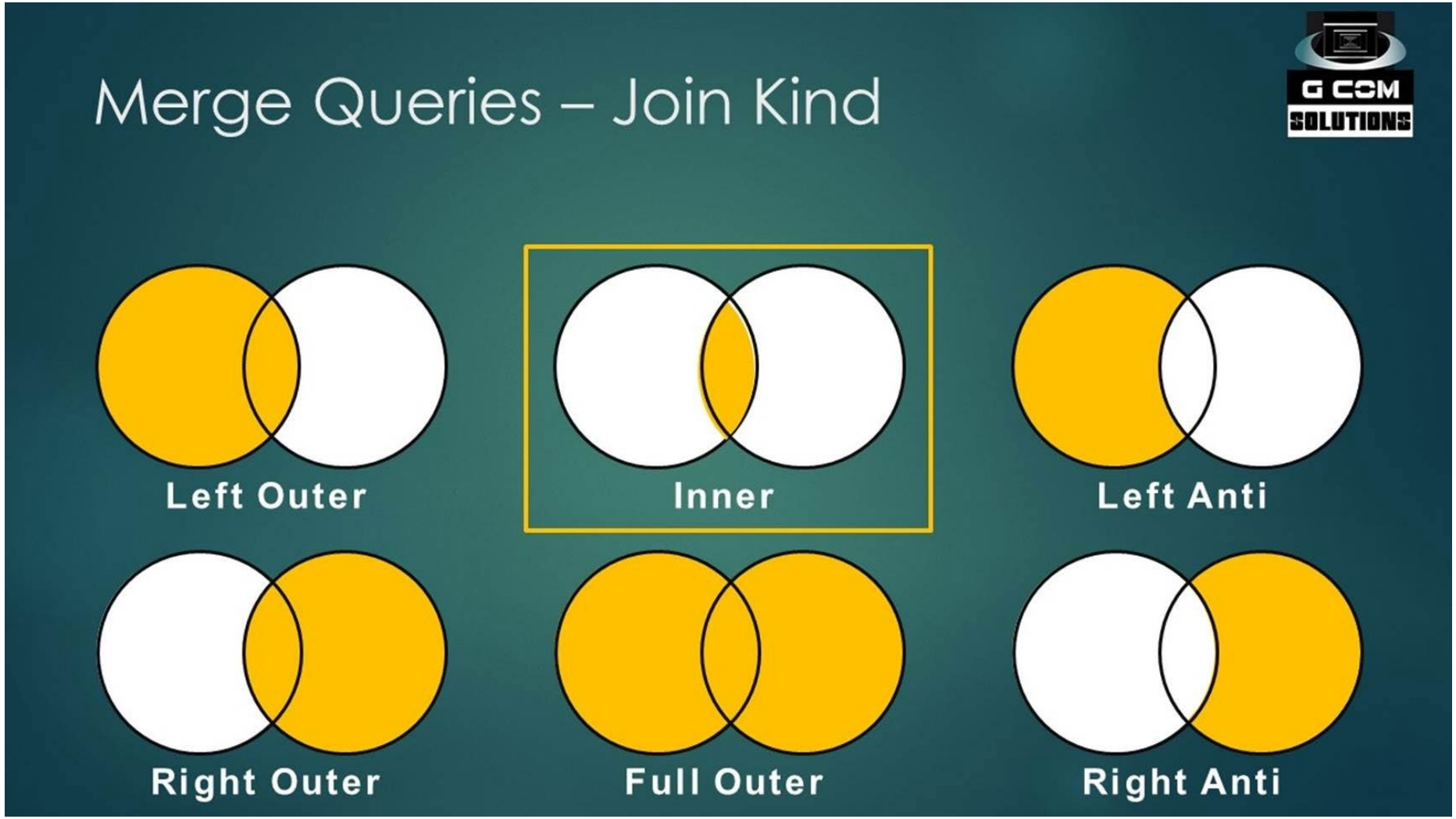
Merge

- Add **columns**

Append vs Merge

- **Append**
  - Adding rows of data to another table or query.
- **Merge**
  - Adding columns from one table (or query) into another.
  - To merge two tables, you must have a **column** that is the **key** between the two tables.
  - Similar to the JOIN clause in SQL using Foreign Key
    - Left Outer (all from first, matching from second)
    - Right Outer (all rows from second, matching from first)
    - Full Outer (all rows from both)
    - Inner (only matching rows)
    - Let Anti (rows only in first)
    - Right Anti (rows only in second)

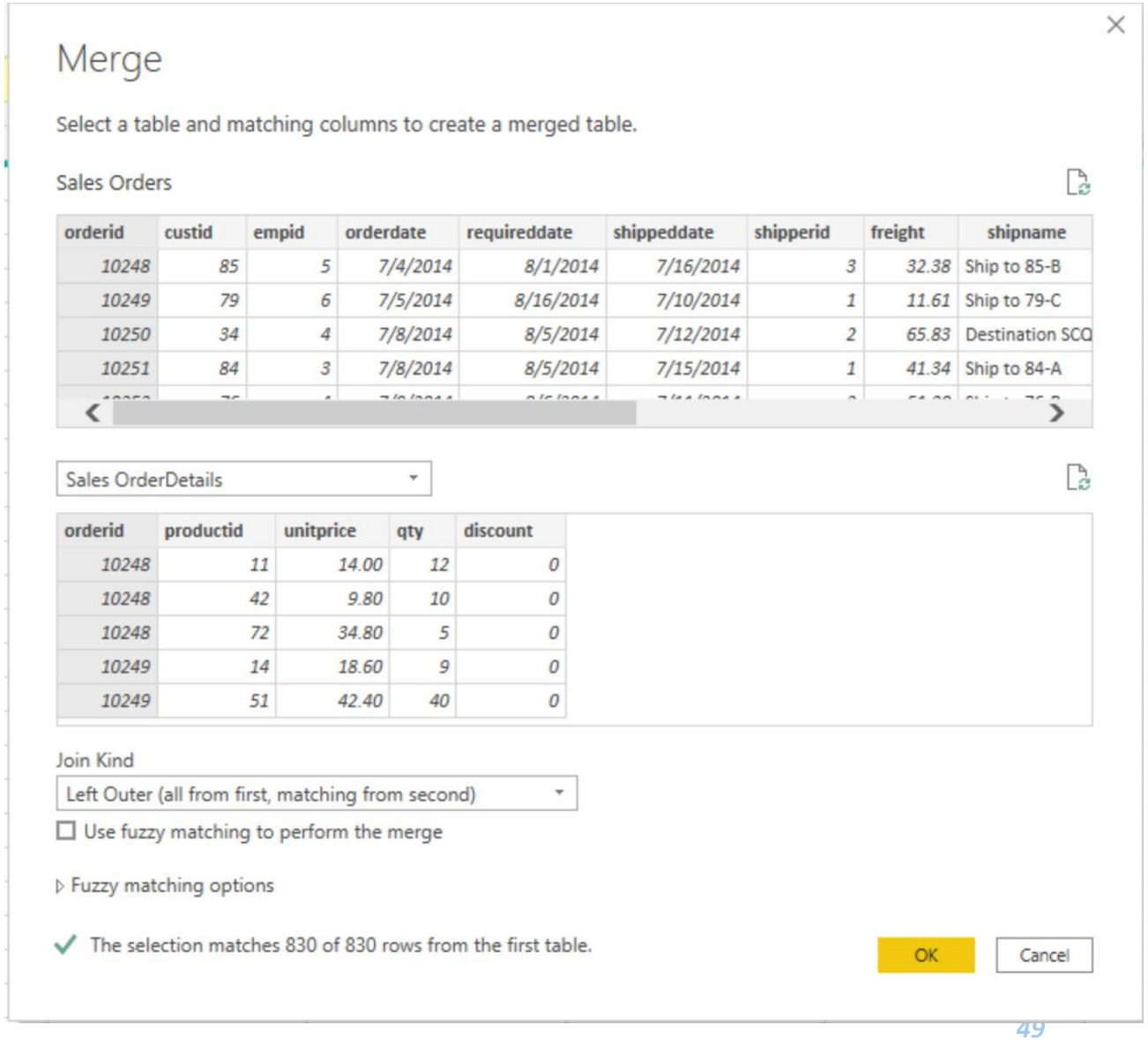
Merge Queries



Merge

Example

- Merge the two tables, **Orders** and **OrderDetails**
- The column that is shared between these two tables is **OrderID**.



Merge

Example

- Merge the two tables, **Orders** and **OrderDetails**
- The column that is shared between these two tables is **OrderID**.

	orderid	orderdate	shipperid	OrderDetails.productid	OrderDetails.qty	OrderDetails.unitprice
1	1	4/23/2018	12	124	12	14
2	2	4/25/2018	24	134	55	11.2
3	3	6/12/2018	19	641	57	45
4	4	6/13/2018	13	98	5	112.5
5	5	7/23/2018	11	312	23	11.1
6	6	7/25/2018	33	124	78	11.2
7	7	8/1/2019	77	137	11	572.1
8	8	8/10/2019	11	124	36	1331.9
9	9	8/11/2019	81	789	85	898.1

Knowledge Check

Merge, Group, Transpose or Append

You import two Microsoft Excel tables named Customer and Address into Power Query.
Customer contains the following columns:

- Customer ID
- Customer Name
- Phone
- Email Address
- Address ID

Address contains the following columns:

- Address ID
- Address Line 1
- Address Line 2
- City
- State/Region
- Country
- Postal Code

Each Customer ID represents a unique customer in the **Customer** table.
Each Address ID represents a unique address in the **Address** table.
You need to create a query that has one row per customer.
Each row must contain City, State/Region, and Country for each customer.
What should you do?

- A. Merge the Customer and Address tables.
- B. Group the Customer and Address tables by the Address ID column.
- C. Transpose the Customer and Address tables.
- D. Append the Customer and Address tables.