Data Transformation Introduction

The First step in Data Analysis is to have Clean Data as Dirty Data produces inaccurate results.

You Need to perform:

1. Data Quality checks
2. Transformations

Data Quality Checks:

1. Completeness
   - Remove null or missing data
   - Fill missing value with placeholder
2. Uniqueness
   - Remove Duplicate records
3. Timeliness
   - Appropriate date range
4. Accuracy
   - Remove inaccurate data

Transformations:

- Data shape transformations
- Change Data Type
- Combine queries
- Filter
- Apply User Friendly value replacements

Clean data advantages:

- Measures and columns produce more accurate results when they are performing aggregations and calculations.
- Tables are organized, where users can find the data in an intuitive manner.
- Duplicates are removed, making data navigation simpler. It will also produce columns that can be used in slicers and filters.
- A complicated column can be split into two, simpler columns.
- Multiple columns can be combined into one column for readability.
- Codes and integers can be replaced with human readable values.

Shaping Data:

- Identity Column headers and names
- Shaping table structure
  - Promote header
  - Rename Columns
  - Remove Top Rows
  - Remove Columns

- Pivot and UnPivot

Data Profiling and Examining the structure:

- Column Quality
- Column Distribution
- Column Profile

Enhance Data Structure

- Rename Query
- Replace Values
- Replace Null Values
- Remove Duplicates
- Evaluate and change Datatype
- Combine multiple Tables into Single table (Append, Merge)
- Use Advanced editor to modify M code

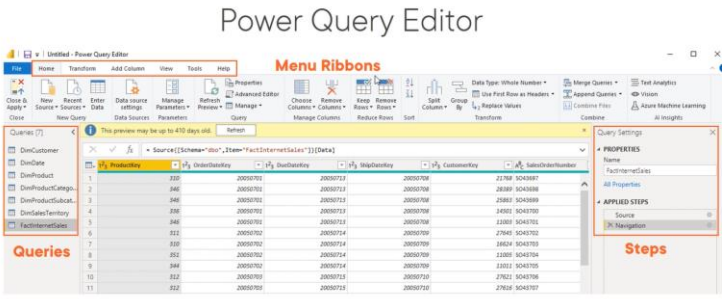
What is Power Query?

Power Query:It writes M code which helps to translate data transformation and cleansing into PowerBI Model

Introduction to Query Editor

What is Power Query?

- It is a common tool in Excel, PowerBI and Power Query Online, Synapse, Data Factory.
- It is used to perform data transformation and cleansing into PowerBI Model.



- Accessing the Query Editor:
  1. Get data→Edit option
  2. Power BI Desktop Home→Queries→Transform Data
- Power BI Query Editor, you can load data from a wide number of data sources, and apply transformations, including adding new columns and measures
- It has the four ribbons: Home, Transform, Add Column, and View
- There are four different panes

1. List of Queries; Left hand side pane will show list of all queries in this solution or file
2. Data Set Pane: In this area the result set will be displayed as preview with limited number of rows
3. Query Settings : Applied steps to the current query is visible in this pane.
4. Transformations Menu; Power Query has many transformations options in GUI that are available through the menu in top section

- It also has formula bar, which shows the code (written in the Power BI "M" language) that performs the selected transformation step.
- It has status bar (at the bottom of the window) that indicates useful information, such as the number of rows and columns in a query table, and the date when the dataset was downloaded
- Based on type of operations transformations can be divided into different categories like
  1. Data Shaping
  2. Data Cleansing
  3. Data Mashup
  4. Filter operations

Applied Step:

- Data transformation is by its very nature a sequential process.
- The various elements that make up a data transformation process are listed in the Applied Steps list of the Query Settings pane in the Query Editor
- You can see all the changes applied to data set step by step.
- You can click on a step and the main pane will show you the data at that step.
- You can undo the changes by removing the step. You can rename the step.
- You can reorder steps if there are no dependencies.
- You can delete step but be aware of dependencies.

Note:

Providing sensible names for the steps in your queries helps if you return to the data after a long time and have forgotten exactly what transformations were applied.

Power BI Desktop Query Editor Context Menus:

Power BI Desktop Query Editor makes full use of context (or "right click") menus as an alternative to using the ribbons. When transforming datasets, there are three main context menus

1. Table menu : This menu appears when you right-click the top corner of the grid containing the data
2. Column menu : This menu appears when you right-click a column title
3. Cell menu : This menu appears when you right-click a data cell

Data Profiling

Profiling data is about studying the nuances of the data.

It helps in

- Determining anomalies
- Examining and developing the underlying data structures
- Querying data statistics such as row counts, value distributions, minimum and maximum values, averages

Data Profiling in PowerBI:

Navigation:

Launch PowerQuery Editor→ Home→Transform Data

View Tab→Data Preview→Observe Column Distribution, Column Profile, Column Quality

Column quality:

- Shows you the percentages of data that is valid, in error, and empty. In an ideal situation, you want 100 percent of the data to be valid.

Column distribution:

- Shows you the distribution of the data within the column and the counts of distinct and unique values

Column profile:

- Gives you a more in-depth look into the statistics within the column.
- It provides basic statistics like Min, Max, Rowcount
- Value distribution graph tells you about the counts for each unique value in that specific column.
- **Column Statistics**, on numeric column, will also include how many zeroes and null values exist, along with the average value in the column, the standard deviation of the values in the column, and how many even and odd values are in the column.

Lab1:

Clean and Transform data of Worldwide movies using data from **http://www.boxofficemojo.com/alltime/world/ ,** which gives worldwide gross sales information of movies.

Clean and Transform data of Top 250 movies ranked by people in **IMDB website** (http://www.imdb.com/chart/top/)

Open following File

D:\PowerBI-Learn\ClassLabs\TransformData\Moviesdata-starter.pbix

Save it as Moviesdata-final.pbix

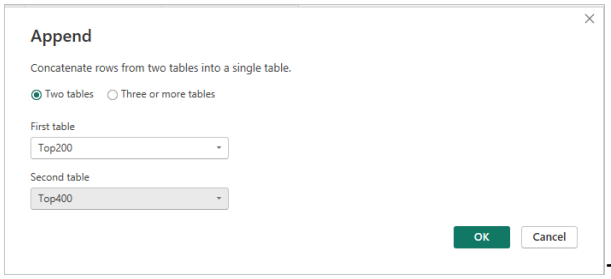
Launch PowerQuery Editor: Home→Transform Data

1. **Rename Query:**

   Queries Pane→Select Required Query (Table1)→Right Click→Rename→IMDB

2. **Combine two or more queries in single Query(Append Queries):**

Click Top200→Home Ribbon→Combine section→Append Queries→Click on Drop down Append Queries as New



→Ok

Note:Ignore and Data Privacy warning

3. Rename **Append1** Query as MoviesSales

Note: For append to works best queries have to be in the same structure (number of columns, order of columns, data type of columns….).It is like Union ALL in SQL Server

4. **Change the Datatype of Year Column:**

Select MoviesSales →Select year column→Right click→Change Type→Text

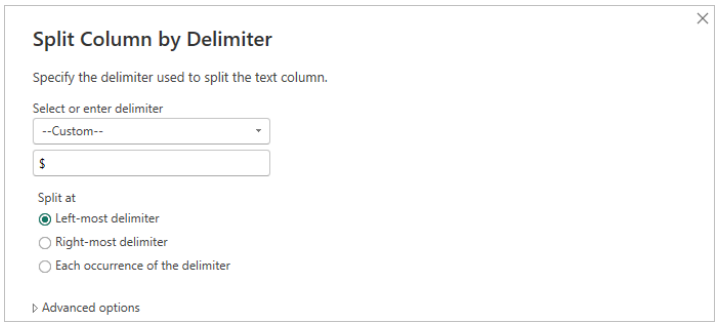
OR

You can just click on Small (1 2 3) icon in year column and change to Text

Note: *Detecting data types: Most of the columns will have the correct data type applied by using this following option* ***TRANSFORM Ribbon →Detect data types*** *The column names are prefixed with letters (ABC), numbers (123), currency symbols ($), or a calendar for datetime columns to represent the data type of the column*

5. Use **Split Column** Lifetime Gross to remove $ with numeric values

Select **Lifetime Gross column** →Right Click→Split Column→**By Delimiter**→

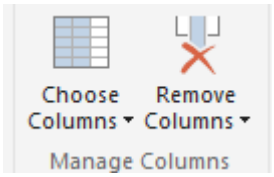


→OK

Note: You can only split columns if they are text data. The Split Column button remains grayed out if your intention is to try to split a date or numeric column.You can Split Columns by Number of Characters OR SplitColumn by a Delimiter

6. **Remove unwanted columns**

   Home→Mange columns→Choose Columns→Choose Columns→uncheck **Lifetime Gross.1→OK**

   Select Column→Home Tab→ 

   Note:

   Choose column allows you to go to particular column where you want to do modification or you can choose columns to keep.

   Remove column can remove selected columns or remove columns other than selected

   You can select multiple columns holding ctrl and select columns

7. **Rename Column Lifetime Gross.2**

   Select Lifetime Gross.2→Right Click→Rename→ Lifetime Gross

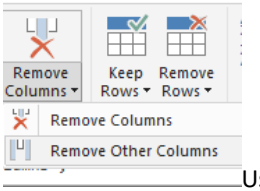
8. **Remove multiple unwanted columns(columns3,column5,column6,column7) in IMDB Query**

   Remove Multiple Columns:

   Select Ctrl-Click Title of Columns to be deleted→Home Tab→Remove Columns

   OR

   If you want to remove many columns and keep few columns

   Select Ctrl-Click Title of Columns to keep→ 

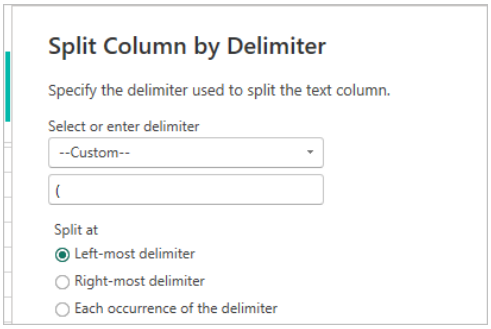
Use Remove Other Columns

9. **Split Column1 separating Id and Title**

   Select Column1→Right Click→Split column→By Delimiter→choose delimiter as "." And split at Left-most delimiter

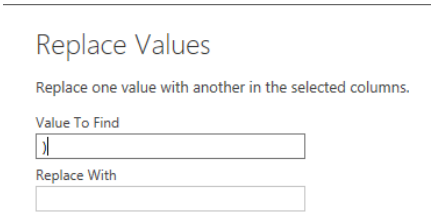
10. **Split Column2 to just get rating**

    Select Column2→Split



11. Replace Values

Select Column2.2 which has values like 2.8M)→Right Click→Replace values



→OK

Note:You can also use Transform section in HOME TAB

12. Rename Columns as Follows

Column1.1→Id

Column1.2→Title

Column2.1→Rating

Column2.2→Views

Column4→Year

13. Remove Extra spaces Using TRIM(Data cleansing)

Select **Title** Column→Right Click→Transform→Trim OR

Transfor Tab→Format→Trim

Note: *Merge Queries is equivalent to **Join** in SQL or database terminology. You can select from different joint Kind.* *Select the matching column based on which you want to merge.*

MERGING/APPENDING:

- When you have one or more columns that you would like to add to another query, you **merge** the queries.
- When you have, additional rows of data that you would like to add to an existing query, you **append** the query.

Other Transformations:

1. **Reorder Columns:**

   Drag the column left or right to its new position

2. **Merge Columns:**

   The order in which you select the columns affects the way that the data is merged.

   So, always begin by selecting the column whose data must appear at the left of the merged column, then the column whose data should be next, and so forth

   Merge Column option can be selected from Transform /ADD Column ribbon or Right click selected column.

   If it is selected from Add Column ribbon, a new column is added with given merge specification, leaving original columns as it is.

   If you use it from Transform ribbon or Context sensitive menu, all selected columns are replaced with single column with given merge specification.

3. **Removing Records**

   You can reduce the size of data set by keeping or removing/keeping some number or range of rows.

   Example: Keep first 50 records

   Home Tab→Reduce Rows→Observe the options

4. **Sorting Data**

   You can sort the data only to get a clearer idea of the data that you are dealing with.

   You can use option from HOME Ribbon/Context menu of selected column

Note: Observe Power BI Desktop Query Editor adds a tiny 1, 2, 3, and so on to the right of the column title to indicate the sort sequence.

5. **Reverse Row Order**

   If you want to arrange the records upside-down you can use this option

   Transform ribbon→ click the Reverse Rows

Filtering Data

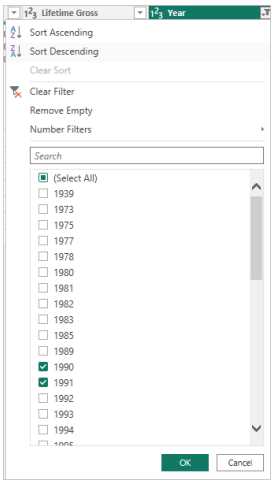
There are two approaches to filter data

- Select one or more specific values from the unique list of elements in the chosen column.
- Define a range of data to include or exclude.
- You have different range filters based on datatype of a column as follows
  - Text Filter
  - Number Filter
  - Date Filter
- Each of these range filters have several options for filter.

- You can combine up to two elements in a filter. These can be mutually inclusive (an AND filter) or they can be an alternative (an OR filter).
- You should not apply any formatting when entering numbers.
- Any text that you filter on is not case-sensitive.

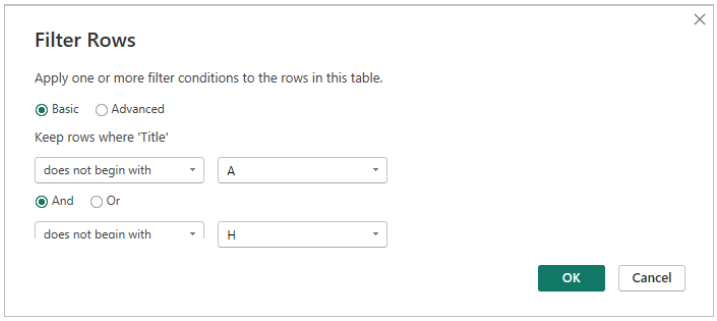
Example: Show data only for year 1990 and 1991 for MoviesSales Query

Click small arrow besides year Column→



Example: Show data only for Movies do not begin with A or H

Click small arrow besides Title→Text Filter→Does not begin With



Note: Enter a letter or a few letters in the Search box and Finding required Elements in the Filter List faster.

You can Filter text, numbers and date and time ranges. Several options are provided for each one of them.

Adding Custom Column

- You can create Custom Columns, also called as derived columns or calculated columns.
- Although they can do many things, their essential role is to
  - Concatenate (or join, if you prefer) existing columns.
  - Add calculations to the data table.

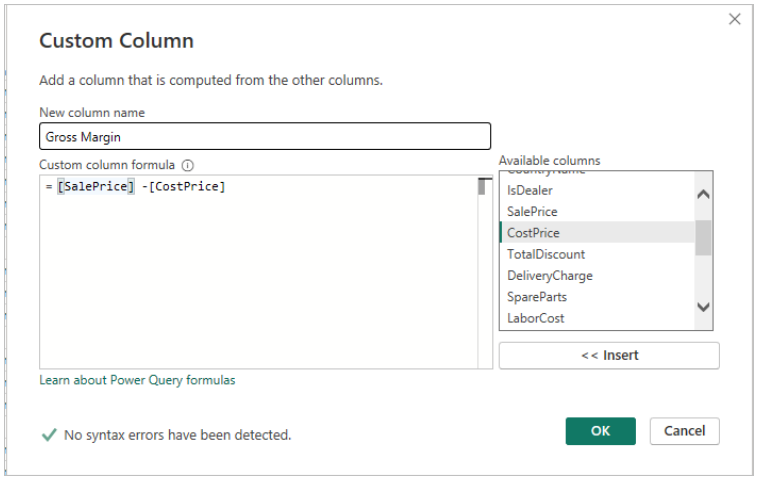
- Extract a specific part of a column.
- Add flags to the table based on existing data.

Lab3: Add Custom Column

1. Create new File →Save it as D:\PowerBI-Learn\ClassLabs\TransformData\TransformDemo2.pbix
2. Use D:\PowerBI-Learn\ClassLabs\TransformData\Salesdata.xlsx Load two queries sales and clients
3. Open Power Query Editor
4. **Add Custom colum in Sales as with formula as**

   Gross Margin = [SalePrice]-[CostPrice]

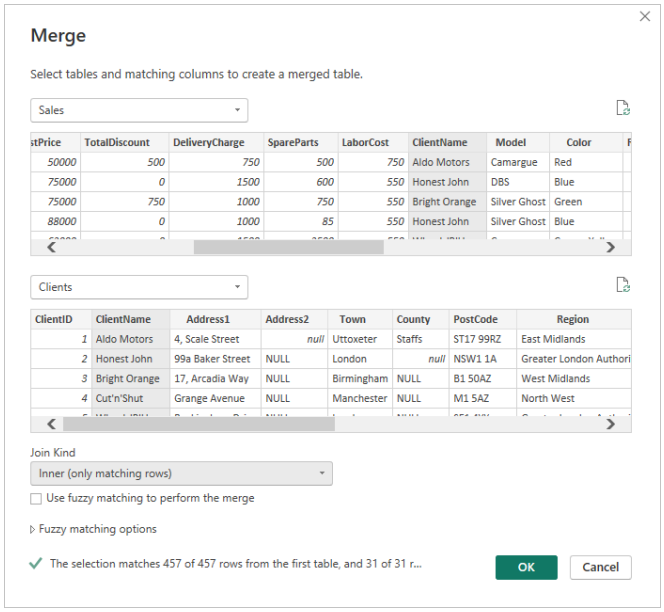
Select Sales Query →Add Column Ribbon->Custom Column→



Merging Data

Lab2: Use TransformDemo2.pbix file and Merge sales and clients Data

1. In Power Query Editor→Select Sales Query→HOME Ribbon→Merge Queries as New→



→OK

Note:

Here ClientName is used as join column.You can also join queries on multiple columns.

new column is added to the right of the existing data table.

In New Column and every row contains the word Table which can be expanded or aggregated.

*When **Merge query as New** is used a **new query created** .The queries participating in merge can be disabled for data load.*

Understanding Expand/Aggregate:

- You have two options: Expand, Aggregate
- Expand:

  When you select expand radio button,The selected columns from the linked table are merged into the main table, and the link to the reference table (New Column) is removed.

  You have useful options here like Use the Original Column Name As the Prefix, Search Columns to Expand
- Aggregate:

  When you select Aggregate radio button you can see columns with aggregated values.(SUM Sales price)
- You can choose the type of aggregation that you wish to apply (before clicking OK),
- To do this, place the cursor over the column that you want to aggregate and click the pop-up menu at the right of the field name and select the required aggregation.

How to prepare datasets for join?

- Before joining real-world datasets you might need to do some preparatory work
- Ensure that the columns used in join have similar data types.
- Ensure that data is clean in the columns that are used for joins before attempting to merge queries.

Example:

Remove trailing and leading spaces in text-based columns.

Isolate part of the column used in join

Verify appropriate datatypes used in join columns

UNPIVOT/PIVOT

- You can convert row values into column values using PIVOT.
- You can pivot columns and create a table that contains aggregated values for each unique value in a column.
- You can convert column values to row using UNPIVOT.
- Unpivot turns multiple column headers into a single column but in rows and store their values in another column

Lab4: Explore Pivot/Unpivot using excel file (unpivot-using-power-query.xlsx)

1. Create a new file TransformDemo3.pbix
2. Get Data from D:\PowerBI-Learn\ClassLabs\TransformData\unpivot-using-power-query.xlsx
3. →Select **Sales** Table → Click on transform→Power Query editor opens
4. Set first row as Row Header

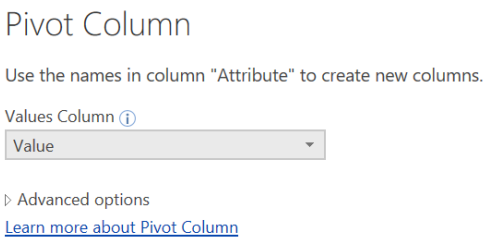
   Transform Ribbon→Click on Use First Row as Headers
5. **UnPivot**

- Select Product Column→Transform Ribbon→Unpivot Columns→Unpivot Other Columns
- It converts all the column headings for the "other columns" into data values for a new column. Since it doesn't know what to name the new column, it names it "Attribute",the other column name is "Values" Containing numeric values.

Note: *You can click on Columns that you want to unpivot, and then select Unpivot columns* *OR* *You can do reverse, select pass through columns, and select unpivot other columns*

6. **Pivot:**

- Select Sales Query→Right Click→Duplicate
- Use this query for PIVOT
- Select **Attribute** Column→Transform→Pivot Column



Managing Queries

- You can manage multiple queries in PowerBI Desktop.
- There are different options for
  - Organizing queries
  - Grouping queries
  - Duplicating queries
  - Referencing queries
  - Adding a column as a new query
  - Enabling data load
  - Enabling data refresh

Organizing queries:

- The Default order of queries in Query pane is that the most recently added data source appears at the bottom of the list.
- You can Move query Up/Down
- Select Query→Right Click→Move Up/Down

Grouping queries:

- You can create a group and add queries to that group for better organization.
- Select Query→Right Click→Move To Group→New Group→Give Name for the group
- All other queries will be now under *Other Queries Group.*
- For moving query from one group to another you can Just drag queries from one group to another OR take the same step as above

Duplicating queries:

- If you have done a lot of work transforming data, you could want to keep a copy of the original query before trying out any potentially risky alterations to your work.
- If you are looking to copy an entire query with all of its steps so that it can be applied to similar other set of data.

Referencing queries:

- You can create reference query. It is similar to Source query and any changes in source will be reflected in it.
- When you reference a query, the new query will have only one step: sourcing from the original query. A referenced query, will not have the applied steps of the original query.
- Reference is a good choice, when you want to branch a query into different pathes. One path that follows a number of steps, and another that follows a different steps, and both are sharing some steps in the original query.
- Example: If you keep only 100 rows in source. The reference query will also show 100 rows only.

Adding a column as a new query:

- You can create a separate query from existing column.
- Select Column→Right Click→Add as New Query

Enabling data load:

- You can enable/disable data Load to powerBI. If it is enabled, Query becomes part of PowerBI desktop data model.
- Some queries that contain only lookup data that is added to another table but not needed in the data model, for instance, you can disable data load and unclutter data model.

Enabling data refresh:

- By enabling data refresh, you gather the very latest data from the source into thePower BI Desktop Query Editor and then into the data model.

What are pending changes?

- After making changes in Query Editor when you switch to the Power BI Desktop View, normally, you then want to load the full data from the source into the data model.
- As data load can be time, consuming you can just close the query editor.
- Whenever convenient you can Apply Changes i.e. reload the modified source data into the data model.