**Final Course Project: Transforming multiple data sources**

[Order2022](https://d3c33hcgiwev3.cloudfront.net/nc1fW78PS22NF8moeFiTZA_5950e1bc92374abc8798250328b155e1_Order2022.xlsx?Expires=1700006400&Signature=E5RBhM9yNV9kJQYK5DJ6bA~J4Z9t8buIb9mgsjCDbUjznptEBiljxAZUnOgxUWq0RmKQr8x5teI0nS-hCA85zoiBarTbyNY-6P52oy~e1UQtMnZKg9G2y1jkH2~T-v3OHfdRoNDSTVFJhbaZg3Bq9aOutyvrBTuOwiLvG-NcBrE_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/nc1fW78PS22NF8moeFiTZA_5950e1bc92374abc8798250328b155e1_Order2022.xlsx?Expires=1700006400&Signature=E5RBhM9yNV9kJQYK5DJ6bA~J4Z9t8buIb9mgsjCDbUjznptEBiljxAZUnOgxUWq0RmKQr8x5teI0nS-hCA85zoiBarTbyNY-6P52oy~e1UQtMnZKg9G2y1jkH2~T-v3OHfdRoNDSTVFJhbaZg3Bq9aOutyvrBTuOwiLvG-NcBrE_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[Order2023](https://d3c33hcgiwev3.cloudfront.net/4qAUEXfvTK6mEswlvamRNg_efb7b7de223443cc8164016f8d9787e1_Order2023.xlsx?Expires=1700006400&Signature=FbkKgPt11TwlXPfp0fFzx6d8qhXiAackWht-dhkKAD40tzEOIHlwyoZcAeyQFHM5AXwJZ74c3qsFBFWp017~f28276cuxzd-kj6KbLjSGHaGYmOqSQfnFq2Y28Qe-Sc1T39d8V45qp6nvfQo3lXTiaVIhzRsl2kGK6SWzW8AskA_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/4qAUEXfvTK6mEswlvamRNg_efb7b7de223443cc8164016f8d9787e1_Order2023.xlsx?Expires=1700006400&Signature=FbkKgPt11TwlXPfp0fFzx6d8qhXiAackWht-dhkKAD40tzEOIHlwyoZcAeyQFHM5AXwJZ74c3qsFBFWp017~f28276cuxzd-kj6KbLjSGHaGYmOqSQfnFq2Y28Qe-Sc1T39d8V45qp6nvfQo3lXTiaVIhzRsl2kGK6SWzW8AskA_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[OrderDetails](https://d3c33hcgiwev3.cloudfront.net/kGQGkdsLTkyzrxjbpmS0_g_da0ea76d222348d0aa044b7fd3802fe1_OrderDetails.xlsx?Expires=1700006400&Signature=TR1NmqnowwPcd-4mMtfdH1e5Ql4c8EqVyvmdUSqZZudhoCjf-TGJv0f0pvGXbZ-uYz3YkZgv~zg3MS-YQKpgx408iggAx6RVsdenOQ5pZjd7SgHZLmT5YW1Jvgdf~J6D6s4VwNvJok3~RX9lc-D45-n~m926MDZ2WIXeVwVup4Q_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/kGQGkdsLTkyzrxjbpmS0_g_da0ea76d222348d0aa044b7fd3802fe1_OrderDetails.xlsx?Expires=1700006400&Signature=TR1NmqnowwPcd-4mMtfdH1e5Ql4c8EqVyvmdUSqZZudhoCjf-TGJv0f0pvGXbZ-uYz3YkZgv~zg3MS-YQKpgx408iggAx6RVsdenOQ5pZjd7SgHZLmT5YW1Jvgdf~J6D6s4VwNvJok3~RX9lc-D45-n~m926MDZ2WIXeVwVup4Q_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

**Introduction**

In this course, using Power BI, you covered the data extract, transform, and load process (ETL). You should now have a good understanding of:

* Connecting to data sources
* Column data types
* Common data errors
* Combining tables with merge and append
* Advanced data transformation techniques
* Loading and staging data
* Profiling data by using column quality
* Column distribution and column profile, and its practical application in identifying data anomalies.

In this exercise, you will apply your knowledge in an end-to-end scenario by using Power Query to clean and transform multiple data sources and join and merge them. You’ll also examine the valid, error, empty, min, max, unique, and distinct values in the rows. This will allow you to identify the anomalies in the data. Finally, you will remove the data sources with anomalies. This exercise will help you understand how to clean, transform, join, and merge data sources in Power Query, and identify potential data anomalies by using data profiling tools.

**Case study**

You are working as a data analyst at Adventure Works. Sales data is contained in two main tables, Order and OrderDetails.

**Data Sources**

The **Order** data table includes general information about the sales such as:

| **Order** |  |  |  |
| --- | --- | --- | --- |
| OrderDate | TotalDue | TerritoryID | SalesPersonID |

The **OrderDetails** data contains related details of each main sales record such as:

| **OrderDetails** |  |  |  |
| --- | --- | --- | --- |
| ProductID | OrderQty | UnitPrice | UnitPrice Discount |

**Files**

Adventure Works trades internationally and generates a large volume of sales data. To manage file sizes, the active Order table only includes data for the year 2023. Older data is stored in separate files for each year with the same fields and table structure.

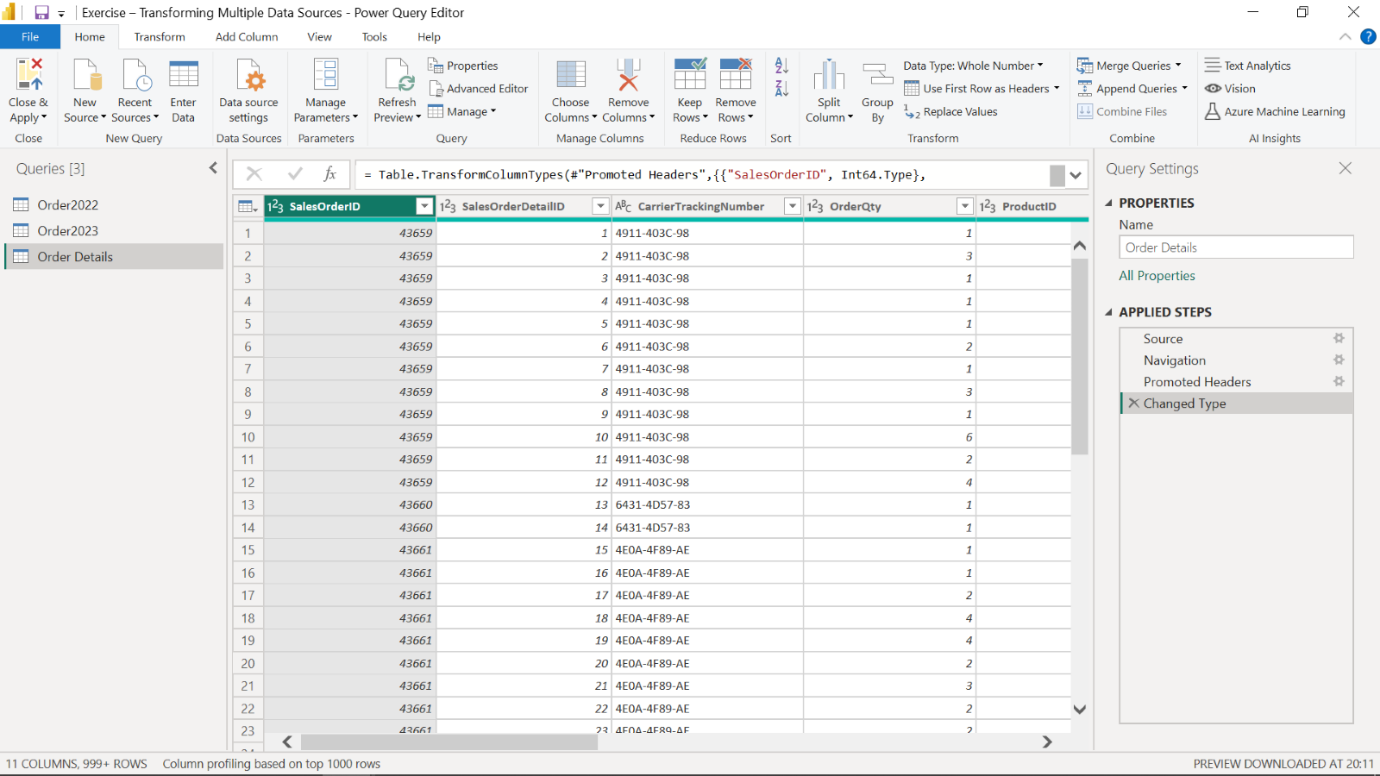
**The task**

Your manager, Adio Quinn, asks you to conduct a detailed analysis of store sales. In the detail table, **OrderDetails**, there are multiple fields, but you only need **ProductID**, the quantity sold (which is in the field **OrderQty**), and the **UnitPrice**. Therefore, you are expected to remove unnecessary fields, and also eliminate empty rows, and identify any anomalies to remove those rows if necessary. After performing these tasks, you will append the two separate sales data sources together and then merge that with **OrderDetails.** Follow the steps below to complete the exercise.

**Instructions**

**Step 1: Set up the project**

1. Create a new Power BI projectcalled *Exercise – Transforming Multiple Data Sources.*
2. Download the *Order2022.xlsx*, *Order2023.xlsx* and *OrderDetails.xlsx* files, which you will use in this exercise.



**Step 2: Open the Power Query Editor**

1. Use the **Get Data** feature in Power BI.
2. Select **Transform** to open the Power Query editor.
3. Import your .xlsx datasets, *Order2022*, *Order2023* and *OrderDetails.*

**Step 3: Choose columns from Order Details**

1. Open the **Order Details** query by selecting it from the Queries pane.
2. Keep **SalesOrderID**, **ProductID**, **OrderQty**, **UnitPrice** columns.
3. Remove the other columns by simply right clicking and selecting the **Remove** option in the shortcut-menu.

**Step 4: Profile data in Order Details**

1. To profile data, select **Order Details**.
2. On the View tab, in the Data Preview group, check Column Distribution, Column Quality and Column Profile checkboxes.
3. Note the amount of **distinct** values and **unique** values in **Column Distribution**.
4. Also check the **valid**, **error** and **empty** values of all columns in **Column Quality**.

**Step 5: Detect potential anomalies in the price**

To detect potential anomalies and assess column distribution for the price, in this step you will assess the **UnitPrice** column in **Order Details** list and find out the min, max, mean values and the distribution of the values.

1. In the **View** tab, in the **Data Preview** group, check **Column Profile** while keeping **Column Distribution** checkbox as checked.
2. Note the **min**, **max**, and **mean** values for the **UnitPrice** column and also assess the **Column Distribution**.
3. You should find that three rows are outliers in the **UnitPrice** column.
4. Consider these three rows as data anomalies (they are most probably mistypes when data was entered) . Remove them by **filtering** and **unchecking** these values to avoid confusion and incorrect calculations.

**Step 6: Append queries**

1. Append **Order2022** and **Order2023** queries in a new master table.
2. Check the newly created query, its **column names**, **row number**, and the **values** appended.
3. Make sure that the operation has been completed successfully.
4. Rename it as **Orders***.*

**Step 7: Merge queries**

1. Select the **Order Details** data in the Queries pane and choose **Merge Queries**.
2. In the opened window, the Order Details table will be automatically shown in the upper part.
3. You will be choosing Order as the next table for merging. **SalesOrderID** is the common column between the tables. To begin establishing a connection select the **SalesOrderID** column in each table.
4. For the **Join Kind** dropdown, choose the join type **Inner Join**, which selects the matching records from the left table and the right table.
5. Choose **Expand** near the newly added Orders table column and choose only the **OrderDate** column from **Orders** table in the opened window.
6. Rename the **Orders.OrderDate** column to **OrderDate** by simply double-clicking on it.

**Conclusion**

You have now successfully completed an end-to-end scenario by cleaning and transforming multiple data sources, joining and merging them, and identifying potential anomalies in the data using Power Query.