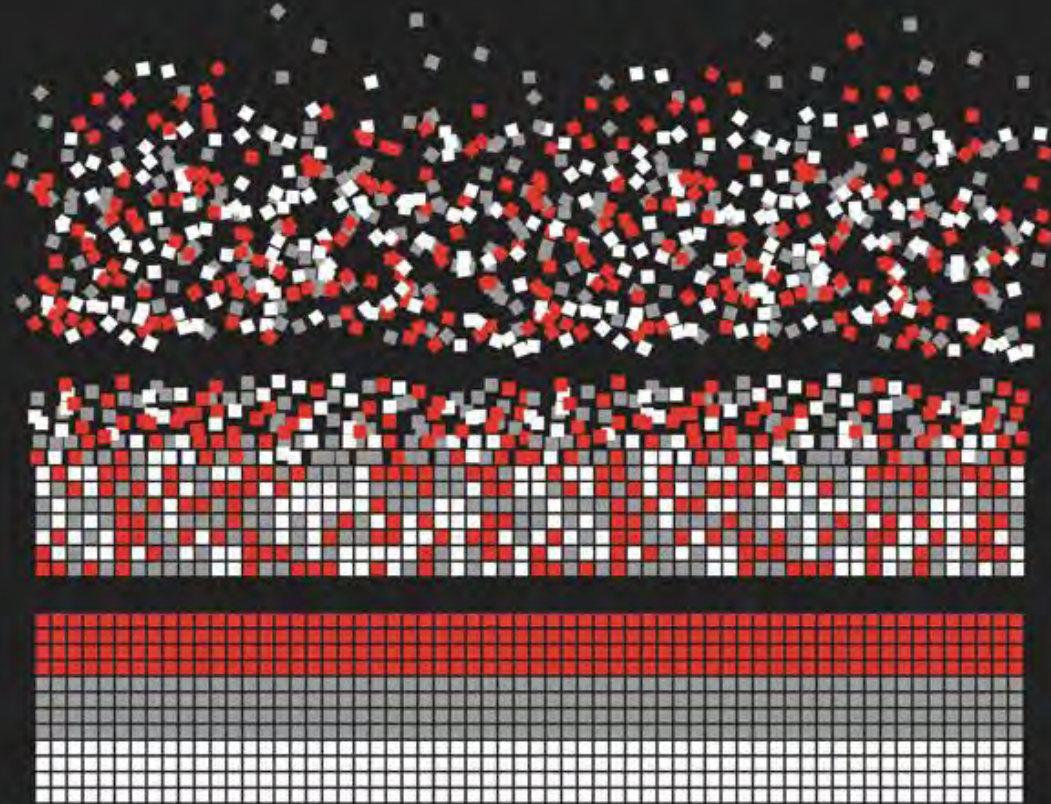


BIG DATA



Assessor Guide

BSBXBD402

Test big data samples

Assessment 3 of 4

Project



Assessment Instructions

Task overview

This assessment task is divided into three parts having eight (8) demonstration activities. Read each question carefully before typing your response in the space provided.

To complete this assessment, you will need the following:

Information and telecommunications equipment

- A computer installed with the Windows operating system.
- Microsoft PowerBI Desktop App - Download and install the free **PowerBI Desktop** App from Microsoft Store: [Downloads | Microsoft Power BI](https://powerbi.microsoft.com/en-au/downloads/) [Long URL: <https://powerbi.microsoft.com/en-au/downloads/>]
- Latest version of DAX Studio – An external tool that can be used for running queries and test scripts for PowerBI – Download and install the free **DAX Studio** App from [Downloads \[DAX Studio.org\]](https://DAX Studio.org/downloads/) [Long URL: <https://DAX Studio.org/downloads/>]

Additional resources and supporting documents

Assessment supporting documents [zipped folder] - This folder contains the following sub-folders, documents and templates required for reference and use when performing the tasks in this assessment.

- AUS Retail_Raw datasets [folder]
 - AUS Retail_Products [.csv]
 - AUS Retail_Sales 2018-2021 [.xlsx]
- AUS Retail_ Data flow and dataset schemas.pdf
- AUS Retail_Big data sample testing policy.pdf
- AUS Retail_Reporting requirements.pdf
- AUS Retail_STM&TestCase_template.xlsx

Assessment Information

Submission

You are entitled to three [3] attempts to complete this assessment satisfactorily. Incomplete assessments will not be marked and will count as one of your three attempts.

All questions must be responded to correctly to be assessed as satisfactory for this assessment.

Answers must be typed into the space provided and submitted electronically via the LMS. Hand-written assessments will not be accepted unless previously arranged with your assessor.

Reasonable adjustment

Students may request a reasonable adjustment for assessment tasks.

Reasonable adjustment usually involves varying:

- the processes for conducting the assessment [e.g. allowing additional time]
- the evidence gathering techniques [e.g. oral rather than written questioning, use of a scribe, modifications to equipment]

However, the evidence collected must allow the student to demonstrate all requirements of the unit.

Refer to the Student Handbook or contact your Trainer for further information.



Please consider the environment before printing this assessment.

Part A: Project scenario

All tasks in this assessment are conducted in a simulated environment where conditions are typical of a work environment that uses big data as it relates to a fictitious retail business organisation called 'AUS Retail'.

Read the project scenario carefully before doing the demonstration tasks in Part B.

A1. Company Background

- **AUS Retail** started off as a single retail store based in Sydney NSW. They now have retail store locations across several other states and territories in Australia and continue to grow with the goal of eventually setting up stores across all states in Australia. As the business is growing rapidly, the management requires a more accurate and efficient way to gain insights into their retail sales.
- The management requires an interactive business intelligence reporting system to be set up and tested so that every month the current retail turnover data reflects the most current information with the ability to see comparisons between the turnover of each month and seasonal patterns of sales each year.
- To achieve this goal, the company had set up a separate team to analyse the organisation's sales data. The team will be led by the **Chief Data Officer (CDO), Mia Gonzales**. Mia had prepared a policy and procedure document called **AUS Retail_Big data sample testing policy.pdf** with up-to-date information on the process that should be followed by the team when performing any data testing and validation tasks.

A2. Your role

- You have recently joined AUS Retail as a trainee and have been given the opportunity to work on the new big data sample testing project. Your supervisor is Mia Gonzales (CDO).
- You must comply with any legislative requirements and follow any standard operating procedures as outlined in the *AUS Retail_Big data sample testing policy.pdf* document when carrying out big data sample testing tasks.
- In your role in this new project, you are required to:
 - test captured transactional and non-transactional big data samples prior to using them in the organisation
 - ensure that the captured transactional and non-transactional big data samples are feasible and accurate so that the data can be used more broadly within the organisation.

A3. Standards, legislative requirements and procedures

You are provided with the following organisational documents and data files related to the fictitious organisation AUS Retail to assist with the big data sample testing process.

- **AUS Retail_ Data flow and dataset schemas.pdf** – This contains the details of internal organisational systems from which various types of data flows, their relationships and the recommended dataset schemas to be used for reporting purposes.
- **AUS Retail_Big data sample testing policy.pdf** – This includes organisational procedures, legislative requirements and industry standards.
- **AUS Retail_Reporting requirements.pdf** – This outlines the requirements for reporting as relevant for the sales and production departments of AUS Retail.
- **AUS Retail_STM&TestCase_template.xlsx** – This is an Excel worksheet template to be used when recording source to target mapping details of datasets, designing test case scenarios and recording test case implementation results.

A4. Sample of raw big datasets

The data analyst team is provided with access to the **AUS Retail_Raw datasets** folder which contains a sample of retail sales data and product data, extracted from the organisation's internal systems.

The **AUS Retail - Raw datasets** folder contains the following datasets (one transactional and one non-transactional) that need to be tested and validated.

- AUS Retail_Sales 2018-2021.xlsx
- AUS Retail_Products.csv

Refer to the **AUS Retail_ Data flow and dataset schemas.pdf** to understand the types of data that flow within each part of the organisation and the general business logic.

Part B: Validate assembled or obtained big data sample

To complete this part of the assessment, you are required to:

- carefully read the scenario details and follow any guidelines and procedures outlined in the *AUS Retail Big data sample testing policy.pdf* document
- access the Microsoft PowerBI technology platform to perform the demonstrations from task B2 onwards.

B1. Establish a sampling strategy and identify a representative sample for big data testing.

Read the scenario details carefully before doing the following task.

Scenario continued:

Your supervisor had asked you to examine the transactional dataset and non-transactional datasets provided to you in the **AUS Retail - Raw data** folder and think of a sampling strategy to identify a representative sample from each big dataset for testing according to the requirements below:

- Dataset 1 [Transactional] - AUS Retail_Sales 2018-2021.xlsx – The representative sample should include product sales from each distinct month in each distinct year.
- Dataset 2 [Non-transactional] AUS Retail_Products.csv – Randomly selected products representative of each distinct category and distinct sub-category.

Refer to the guidelines outlined in section *4.1 Guidelines for establishing a sampling strategy* of the *AUS Retail Big data sample testing policy.pdf* document.

Task:

Determine a sampling strategy that should be used to identify a representative sample from each dataset.

Specify the details for each sampling strategy criteria for both transactional and non-transactional datasets in the answer tables given below.

Answer tables:

Assessor instructions: Students must demonstrate understanding of an appropriate sampling strategy to be used and should fill-in the **Details** column for all criteria. The sampling strategy should be outlined for both transactional and non-transactional datasets.

A sample answer is provided below.

Table 1 - Sampling strategy for Dataset 1 [Transactional]

Transactional dataset criterion:	Sampling strategy details
Dataset source details: <i>[Filename, extension]</i>	AUS Retail_Sales 2018-2021.xlsx
Population: <i>[Size of the raw dataset in terms of rows/records]</i>	9994
Identifying the frame and units [Ensure that distinct groups/categories/segments are captured]:	Are there any distinct groups/categories/segments that the sample needs to be representative of? <input checked="" type="checkbox"/> Yes <input type="checkbox"/> No If Yes , provide details of the relevant groups/categories/segments. Product sales from each distinct month in each distinct calendar year.

Transactional dataset criterion:	Sampling strategy details
Sampling method to be used:	Stratified sampling
Details of the sampling strategy: [Confidence Level, Proportion, Error rate]	Confidence Level = 95% [or can be 99%] Proportion = 0.4 [a value ranging between 0 – 1] Confidence interval = 0.12 [can vary depending on the student's calculation]
Sample size: [Outline your calculations for the suggested sample size]	480 [The student may choose a different value, depending on their strategy. See example calculations below.] Sample size calculation [Example 1]: The strategy was to allow for 5 sales records from each month of each year. 5 [sales records] X 12 [month] X 4 [years from 2018-2021] = 240 Sample size calculation [Example 2]: The strategy was to allow for 10 sales records from each month of each year. 10 [sales records] X 12 [months] x 4 [years from 2018-2021] = 480

Table 2 - Sampling strategy for Dataset 2 (Non-transactional)

Non-transactional dataset criterion:	Sampling strategy details
Dataset source details: <i>[Filename, extension]</i>	AUS Retail_Products.csv
Population: <i>[Size of the raw dataset in terms of rows/records]</i>	9593
Identifying the frame and units [Ensure that distinct groups/categories/segments are captured]:	Are there any distinct groups/categories/segments that the sample needs to be representative of? <input checked="" type="checkbox"/> Yes <input type="checkbox"/> No If Yes , provide details of the relevant groups/categories/segments. Products from each distinct category and each distinct sub-category.
Sampling method to be used:	Stratified sampling
Details of the sampling strategy: [Confidence Level, Proportion, Error rate]	Confidence Level = 95% [or can be 99%] Proportion = 0.4 [a value ranging between 0 – 1] Confidence interval = 0.12 [can vary depending on the student's calculation]
Sample size: [Include calculations where relevant]	340 [The student may choose a different value, depending on their strategy. See example calculations below.] Sample size calculation [Example 1]: The strategy was to allow for 10 product records from distinct subcategory that belongs to each category. 10 [product records] X 17 [sub-categories] = 170 Sample size calculation [Example 2]: The strategy was to allow for 20 product records from distinct subcategory that belongs to each category. 20 [product records] X 17 [sub-categories] = 340

B2. Assemble and obtain a sample of raw big data

Read the instructions carefully and do the following task.

Instructions:

As preparation for this task, do the following first.

- Create a new folder in your local computer called 'BSBXBD402_Firstname_Lastname'. – All the files you will be working on in this assessment should be saved in this folder location.
- Within the *BSBXBD402_Firstname_Lastname* folder, create the following sub-folders
 - Phase 1 – Data validation
 - Phase 2 – MapReduce validation
- Copy the raw data files from the **AUS Retail – Raw datasets** folder into the 'Phase 1 – Data validation' folder.
- Open the *PowerBI Desktop* application
- Save a blank PowerBI file in the 'Phase 1 – Data validation' folder as 'Phase1_Data validation_YourNameInitials_DDMMYYYY'.
E.g. A file saved on the 12th April 2022 by John Smith should have the name: Phase1_Sample data validation_JS_12042022'

Refer to the following sections in the *AUS Retail_ Big data sample testing policy.pdf* document and adhere to any legislative requirements when performing the following tasks listed in the table.

- **Section 4.2 Procedure to assemble raw data into PowerBI**
- **Section 4.3 Procedure to obtain a representative sample from raw data**

For each of the following tasks, provide evidence in the form of screenshots in the answer table below. The screenshots should clearly show:

- an expanded view of the *Fields* column in *PowerBI Desktop* having all data fields/columns loaded from the source dataset
- the *Formula* section displaying any DAX functions used – if relevant to the task
- the number of rows displayed at the bottom of the PowerBI window for each table
- the name of the PowerBI file displayed on the title bar as 'Phase1_Data validation_YourNameInitials_DDMMYYYY'.

Tasks:

B2.1 Assemble Dataset 1 [transactional] from the raw data source onto the PowerBI platform according to the relevant organisational procedures and legislative requirements [see section 4.1 of the *AUS Retail_ Big data sample testing policy.pdf*].

B2.2 Obtain the representative sample from Dataset 1 [transactional] by following the relevant organisational procedure [see section 4.2 of the *AUS Retail_ Big data sample testing policy.pdf*].

B2.3 Assemble Dataset 2 [non-transactional] from the raw data source onto the PowerBI platform according to the relevant organisational procedures and legislative requirements. [See section 4.1 of the *AUS Retail_ Big data sample testing policy.pdf*].

B2.4 Obtain the representative sample from Dataset 2 [non-transactional] by following the relevant organisational procedure. (see section 4.2 of the *AUS Retail_ Big data sample testing policy.pdf*).

Evidence of performing the task:

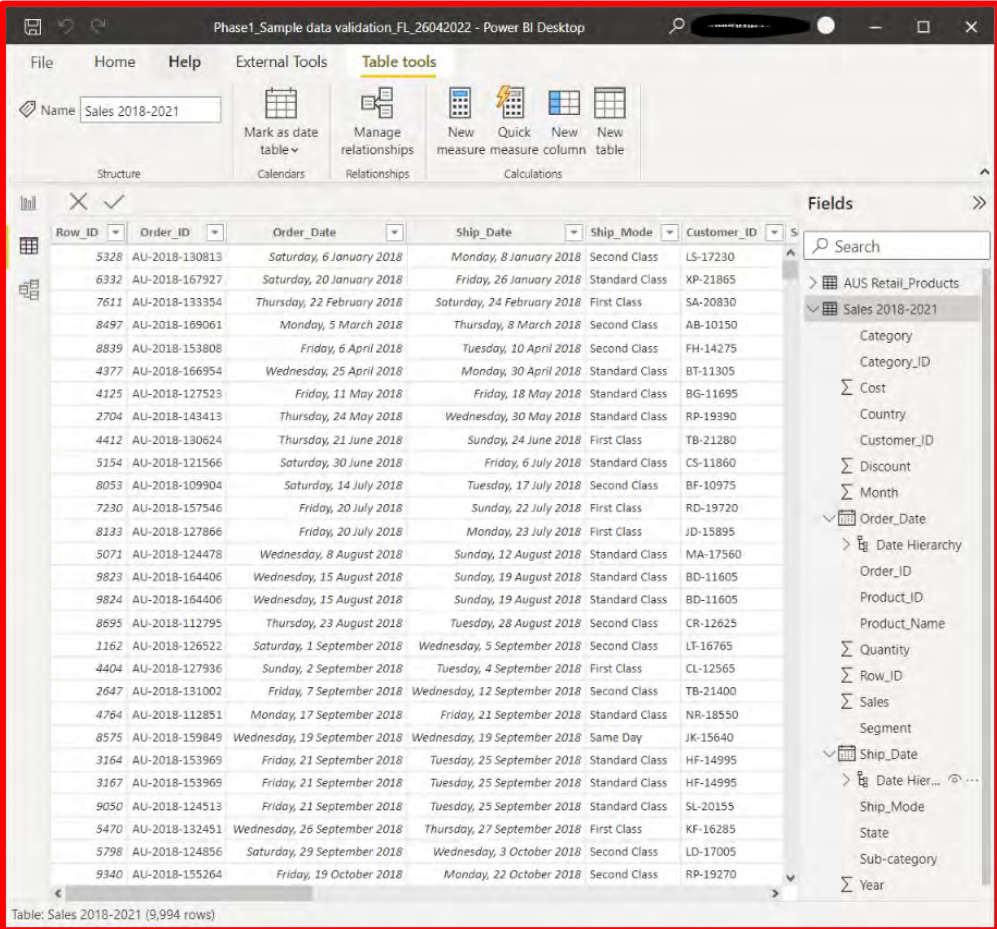
In addition to the screenshots you will include in **Table 3** given below, your assessment submission must include the following documents in the 'Phase 1 – Data validation' sub-folder

- A copy of the raw data files
 - AUS Retail_Sales 2018-2021
 - AUS Retail_Products
- PowerBI work file
 - Phase1_Data validation_YourNameInitials_DDDMMYYYY

Assessor guidelines:

- Students should provide screenshots to show transactional data (e.g. Sales) and non-transactional data (Products) have been loaded to PowerBI correctly.
- The number of records loaded for each data type should tally with the representative sample information outlined in task B1.
- Top row of each table should be correctly identified as the header row.
- Refer to the contents in the **BSBXBD402_AG_03_Project_Exemplar (student submission folder)** sample work files.

Table 3 - Evidence of performing the demonstration task B2.

Demonstration tasks:	Evidence (Screenshot):
<p>B2.1 Assemble Dataset 1 (transactional) from the raw data source onto the PowerBI platform according to the relevant organisational procedures and legislative requirements (see section 4.1 of the <i>AUS Retail_ Big data sample testing policy.pdf</i>).</p> <p>Assessor guidelines: To demonstrate that legislative requirements have been followed, the students <u>should not</u> be loading any customer name details onto the PowerBI platform. However, Customer ID should be loaded.</p>	 <p>Figure 1 - Screenshot for task B2.1 using PowerBI Desktop © Microsoft</p>

Demonstration tasks:

B2.2 Obtain the representative sample from **Dataset 1** (transactional) by following the relevant organisational procedure (see section 4.2 of the *AUS Retail_ Big data sample testing policy.pdf*).

Assessor guidelines:
The number of records loaded for each data type should tally with the representative sample information outlined in Task B1.

Evidence (Screenshot):

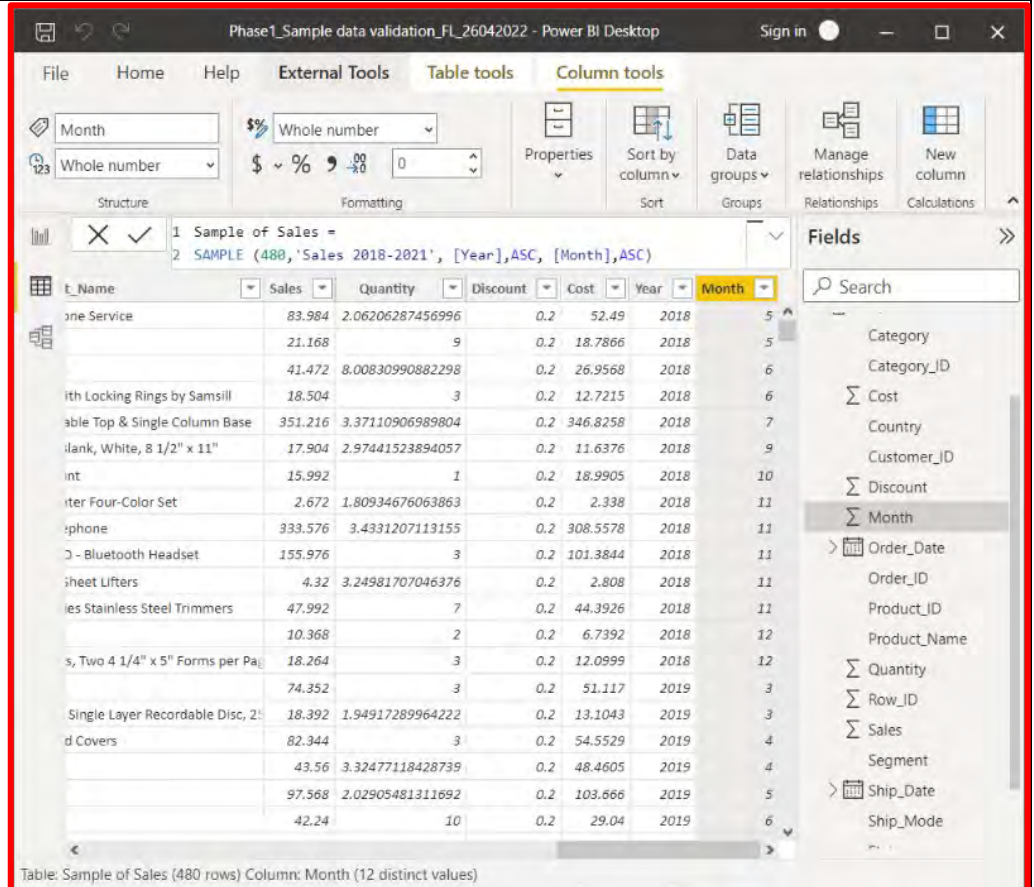


Figure 2 - Screenshot for task B2.2 using PowerBI Desktop © Microsoft

B2.3 Assemble **Dataset 2** (non-transactional) from the raw data source onto the PowerBI platform according to the relevant organisational procedures and legislative requirements. (See section 4.1 of the *AUS Retail_ Big data sample testing policy.pdf*).

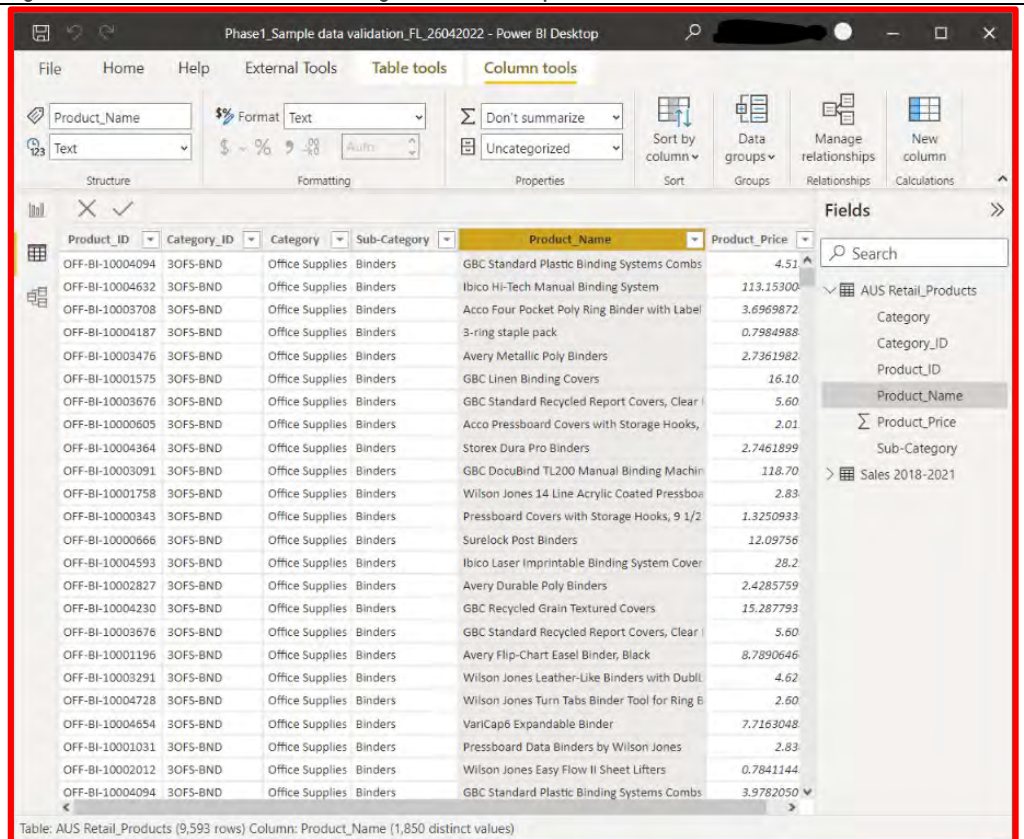


Figure 3 - Screenshot for task B2.3 using PowerBI Desktop © Microsoft

Demonstration tasks:

Evidence (Screenshot):

B2.4 Obtain the representative sample from **Dataset 2** (non-transactional) by following the relevant organisational procedure (see section 4.2 of the *AUS Retail_ Big data sample testing policy.pdf*).

Assessor guidelines:
The number of records loaded for each data type should tally with the representative sample information outlined in Task B1.

The screenshot shows the Power BI Desktop interface with a table named 'Sample of Products' displayed in a grid view. The table has the following columns: Product_ID, Category_ID, Category, Sub-Category, Product_Name, and Product_Price. The data is filtered to show 340 rows. The table is titled 'Table: Sample of Products (340 rows)'. The interface also shows the 'Fields' pane on the right with a search bar and a list of fields including AUS Retail_Products, Sales 2018-2021, Sample of Products, Category, Category_ID, Product_ID, Product_Name, Product_Price, and Sub-Category.

Figure 4 – Screenshot for task B2.4 using PowerBI Desktop © Microsoft

B3. Validate the accuracy of the big data samples

In this task, you are required to validate the obtained big data samples of Dataset 1 (transactional) and Dataset 2 (non-transactional) against the source dataset to ensure that the big data samples contain accurate data.

Task:

Use *DAX Studio* to validate the sample dataset against the source data by using a customised validation script according to the specifications outlined in *AUS Retail_ Big data sample testing policy.pdf* > 4.4 Big data sample validation procedure.

- record evidence of performing this task and the validation outcomes in Table 4 and Table 5 including:
 - screenshots of the query used and the validation results (for both transactional and non-transactional datasets)
 - copies of the validation results saved as DAX Query files for each dataset:
 - Dataset1-Validation Results (Sales)
 - Dataset2-Validation Results (Products)
 - written comments on the accuracy of the data in the big data samples based on the data validation reports. (Word count: 25 – 45 words per comment for each data sample).
- copy the sample data table contents from *PowerBI* table view and paste it onto a new Excel worksheet and save it in the 'Phase 2 – MapReduce validation' folder using the following format for each dataset sample.
 - For dataset 1 (transactional) name the validated sample dataset as:
 - *AUS Retail Sales_sample*

- For dataset 2 [non-transactional] name the validated sample dataset as:
 - *AUS Retail Products_sample*

Evidence of performing the task:

In addition to the screenshots you will include in **Table 4** and **Table 5** given below, your assessment submission must include the following documents in the 'Phase 1 – Data validation' sub-folder

- *Validated sample datafiles:*
 - *AUS Retail Sales_sample*
 - *AUS Retail Products_sample*
- *Validation results – evidence (DAX Query files)*
 - *Dataset1-Validation Results (Sales)*
 - Dataset2-Validation Results [Products]

Assessor instructions: Refer to the contents in the **BSBXBD402_AG_03_Project_Exemplar (student submission folder)** sample work files.

Table 4 – Evidence of validating Dataset 1 [Transactional]

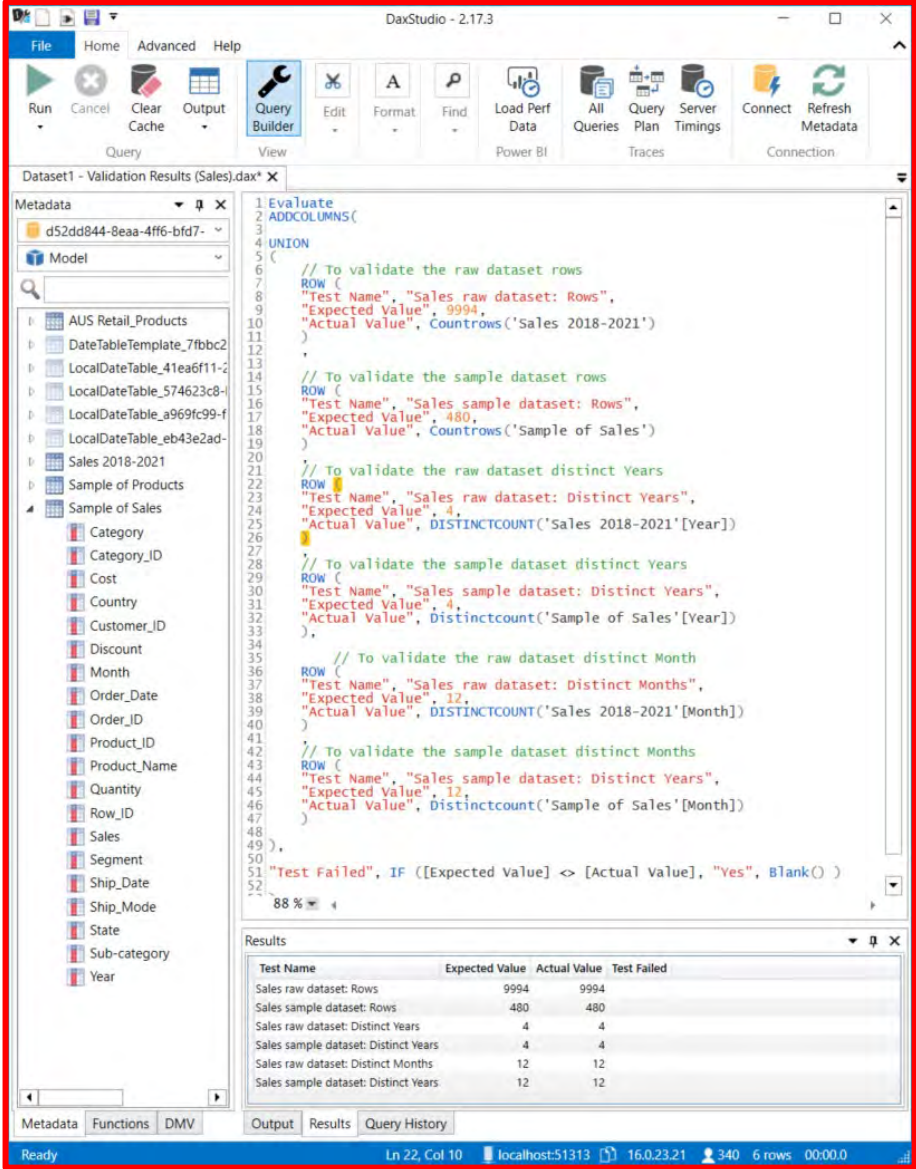
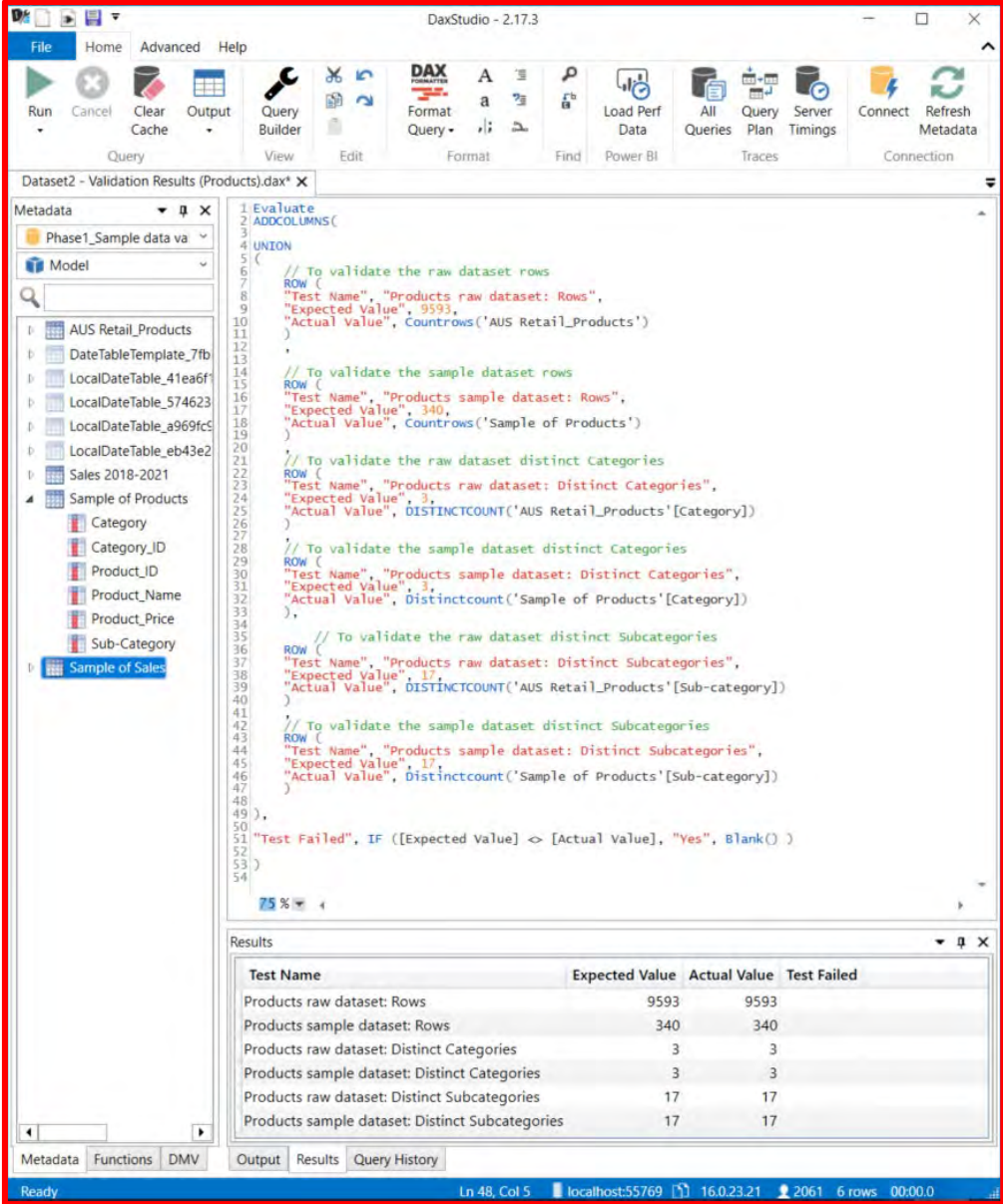
Criterion	Sample validation test script and outcome of results
<p>Validation report <i>[Screenshot]:</i></p>	 <p>Figure 5 – Screenshot for task B3 Dataset 1 using DAX Studio © DAX Studio</p>
<p>Comment on the accuracy of the data in the big data sample. <i>[Word count: 25 – 45 words]</i></p>	<p>Distinct values for both the raw dataset and sample data set are the same as shown in the report. Therefore can ensure accurate categories of data had been captured in the sample dataset.</p>
<p>Validated big data sample filename saved as:</p>	<p><i>AUS Retail Sales_sample.xlsx</i></p>

Table 5 - Evidence of validating Dataset 2 (Non-transactional)

Criterion	Sample validation test script and outcome of results																												
<p>Validation report <i>(Screenshot):</i></p>	 <p>The screenshot shows the DAX Studio interface with a query script and its results. The script is as follows:</p> <pre> 1 Evaluate 2 ADDCOLUMNS(3 4 UNION 5 (6 // To validate the raw dataset rows 7 ROW (8 "Test Name", "Products raw dataset: Rows", 9 "Expected Value", 9593, 10 "Actual Value", Countrows('AUS Retail_Products') 11) 12 13 // To validate the sample dataset rows 14 ROW (15 "Test Name", "Products sample dataset: Rows", 16 "Expected Value", 340, 17 "Actual Value", Countrows('Sample of Products') 18) 19 20 // To validate the raw dataset distinct Categories 21 ROW (22 "Test Name", "Products raw dataset: Distinct Categories", 23 "Expected Value", 3, 24 "Actual Value", DISTINCTCOUNT('AUS Retail_Products'[Category]) 25) 26 27 // To validate the sample dataset distinct Categories 28 ROW (29 "Test Name", "Products sample dataset: Distinct Categories", 30 "Expected Value", 3, 31 "Actual Value", Distinctcount('Sample of Products'[Category]) 32) 33 34 // To validate the raw dataset distinct Subcategories 35 ROW (36 "Test Name", "Products raw dataset: Distinct Subcategories", 37 "Expected Value", 17, 38 "Actual Value", DISTINCTCOUNT('AUS Retail_Products'[Sub-category]) 39) 40 41 // To validate the sample dataset distinct Subcategories 42 ROW (43 "Test Name", "Products sample dataset: Distinct Subcategories", 44 "Expected Value", 17, 45 "Actual Value", Distinctcount('Sample of Products'[Sub-category]) 46) 47 48), 49 50 "Test Failed", IF ([Expected Value] <> [Actual Value], "Yes", Blank()) 51) 52) 53) 54 </pre> <p>The Results table below the script shows the following data:</p> <table border="1" data-bbox="646 1120 1412 1332"> <thead> <tr> <th>Test Name</th> <th>Expected Value</th> <th>Actual Value</th> <th>Test Failed</th> </tr> </thead> <tbody> <tr> <td>Products raw dataset: Rows</td> <td>9593</td> <td>9593</td> <td></td> </tr> <tr> <td>Products sample dataset: Rows</td> <td>340</td> <td>340</td> <td></td> </tr> <tr> <td>Products raw dataset: Distinct Categories</td> <td>3</td> <td>3</td> <td></td> </tr> <tr> <td>Products sample dataset: Distinct Categories</td> <td>3</td> <td>3</td> <td></td> </tr> <tr> <td>Products raw dataset: Distinct Subcategories</td> <td>17</td> <td>17</td> <td></td> </tr> <tr> <td>Products sample dataset: Distinct Subcategories</td> <td>17</td> <td>17</td> <td></td> </tr> </tbody> </table>	Test Name	Expected Value	Actual Value	Test Failed	Products raw dataset: Rows	9593	9593		Products sample dataset: Rows	340	340		Products raw dataset: Distinct Categories	3	3		Products sample dataset: Distinct Categories	3	3		Products raw dataset: Distinct Subcategories	17	17		Products sample dataset: Distinct Subcategories	17	17	
Test Name	Expected Value	Actual Value	Test Failed																										
Products raw dataset: Rows	9593	9593																											
Products sample dataset: Rows	340	340																											
Products raw dataset: Distinct Categories	3	3																											
Products sample dataset: Distinct Categories	3	3																											
Products raw dataset: Distinct Subcategories	17	17																											
Products sample dataset: Distinct Subcategories	17	17																											
<p>Comment on the accuracy of the data in the big data sample. <i>(Word count: 25 – 45 words)</i></p>	<p>Distinct values for both the raw dataset and sample data set is the same as shown in the report. Therefore can ensure accurate categories of data had been captured in the sample dataset.</p>																												
<p>Validated big data sample filename</p>	<p><i>AUS Retail_Products_sample.xlsx</i></p>																												

Part C: Validate big data sample process and business logic

As preparation for this task, do the following first.

1. Save a copy of the *AUS Retail_STM&TestCase_template.xlsx* in the 'Phase 2 – MapReduce validation' folder for each dataset and rename the files as follows:
 - *AUS Retail_STM&TestCase_Dataset1(Sales)_YourNameInitials_ddmmyyyy.xlsx*

- *AUS Retail_STM&TestCase_Dataset2[Products]_YourNameInitials_ddmmyyyy.xlsx*
E.g. A file saved on the 20th April 2022 by John Smith should have a filename as follows:

- *'AUS Retail_STM&TestCase_Dataset1[Sales]_JS_20042022.xlsx'*

2. Place a copy of the previously validated sample dataset files from the 'Phase 1 – Data validation' folder into the 'Phase 2 – MapReduce validation' folder as you will be performing process validation tasks on these sample datasets in this part of the assessment.

- *AUS Retail Sales_sample*
- *AUS Retail Products_sample*

C1. Align datasets to relevant parts of the organisation

In this task, you are required to further evaluate the contents of each sample dataset and align the source fields (column names) to specific entities and relevant parts of the organisation.

Instructions:

Refer to the following documents, specifications, and advice from your supervisor to understand AUS Retail's business logic and reporting requirements.

- *AUS Retail_ Data flow and schemas.pdf* – outlines the operational data types, sources, flows and recommended schemas to be implemented.
- *AUS Retail_Reporting requirements.pdf* – outlines the requirements for reporting as relevant for the sales and production departments.
- *AUS Retail_ Big data sample testing policy.pdf* > section **5.1 Target table field/column alignment with source systems** – outlines specific requirements to consider when aligning source system dataset fields/columns with the target output table fields/columns.
- Advice received from your supervisor as shown below:

"Please consider the following additional requirements for sales related target output fields that need to be captured in the source to target mapping table.

- *A new 'Profit' column to calculate profit from each sales order. (Profit = Revenue – Cost)*
- *A new 'Location' column that combines the County, State details to indicate the location from where each order is placed*
- *The Sales column should be renamed as Revenue.*
- *The Management had also informed that any shipping related data is not required to be included in the sales reports."*

Task:

Complete the *AUS Retail_STM&TestCase_template.xlsx* > **Source to Target Mapping** tab for each dataset by:

- using the *AUS Retail_STM&TestCase_template.xlsx* documents to record source to mapping details separately for each dataset
- identifying and recording the following details for each dataset in the *Source to Target Mapping* tab
 - source system details such as table name and source field names - record the sample dataset table name and its associated column names for each dataset
 - target output table details such as the table names and target field names - align each Source Field [Column Name] recorded for each dataset to the relevant entities of the organisation by filling the Target Output [Table Name] and Target Fields [Column Name] columns
 - transformation logic – this includes details of any queries, functions, expressions, filters or calculation formulas etc. that can be used to generate the required result to create the target fields/columns.
 - use the comments column to:

- record any discrepancies between the source field data and the target field data considering the target output requirements for PowerBI reporting
- make notes of any columns that are not required for reporting.
- including any new target output fields that are required to get the desired output according to the reporting requirements and advice from your supervisor.

Evidence of performing the task:

Your assessment submission must include the following documents in the 'Phase 2 – MapReduce validation' sub-folder. The *Source to Target Mapping* tab should be completed with the required information.

- *AUS Retail_STM&TestCase_Dataset1(Sales)_YourNameInitials_ddmmyyyy.xlsx*
- *AUS Retail_STM&TestCase_Dataset2(Products)_YourNameInitials_ddmmyyyy.xlsx*

Assessor instructions: Samples of the completed Source to Mapping information for both datasets are given below. Also refer to the contents in the **BSBXBD402_AG_03_Project_Exemplar (student submission folder)** sample work files.

AUS Retail_STM&TestCase_Dataset1(Sales)_YourNameInitials_ddmmyyyy.xlsx

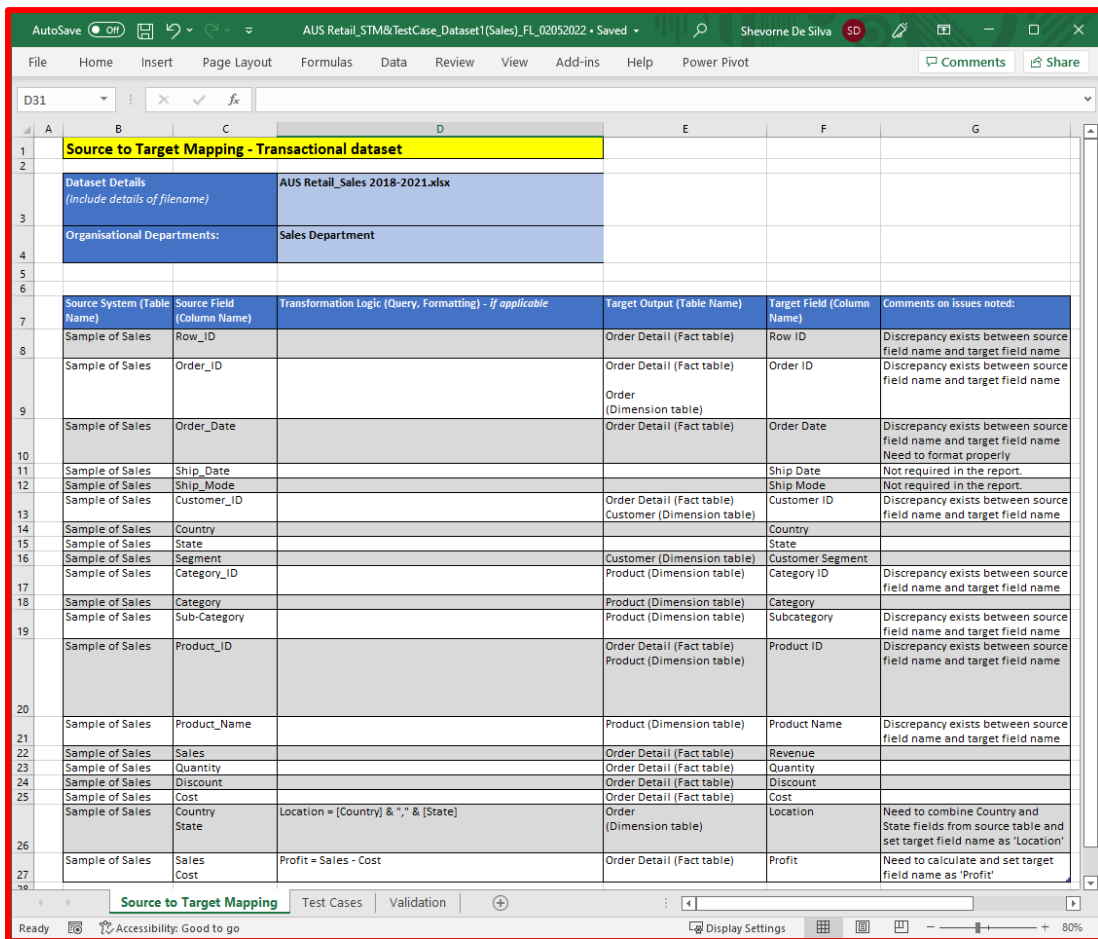


Figure 7 - Screenshot for task C1 Dataset 1 using Microsoft Excel © Microsoft

AUS Retail_STM&TestCase_Dataset2(Products)_YourNameInitials_ddmmyyyy.xlsx

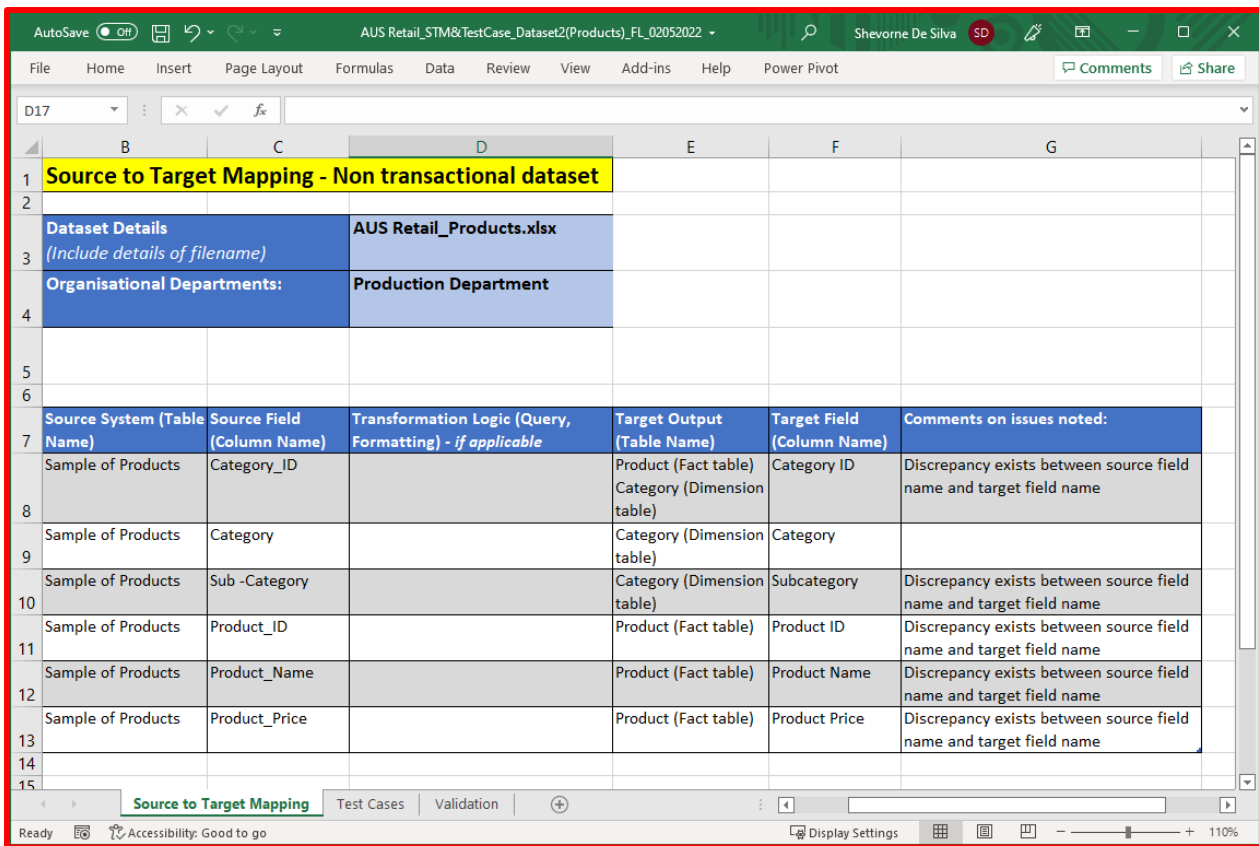


Figure 8 - Screenshot for task C1 Dataset 2 using Microsoft Excel © Microsoft

C2. Implement data segregation rules

In this task, you are required to implement data segregation rules to create the required target output tables from the sample datasets loaded in *PowerBI* according to the **Source to Target Mapping** table (included in *AUS Retail_STM&TestCase_template.xlsx*) created in the previous task.

Instructions:

Open the *PowerBI Desktop* application and save a blank *PowerBI* file in the 'Phase 2 – MapReduce validation' folder for each dataset as:

- 'Dataset1_MapReduce validation_YourNameInitials_DDMMYYYY'
- 'Dataset2_MapReduce validation_YourNameInitials_DDMMYYYY'.

E.g. A file saved on the 12th April 2022 by John Smith should have the name: 'Dataset1_MapReduce validation_JS_12042022'

When providing screenshots, ensure that they clearly show the *Report* view tabs that are named appropriately to indicate which type of data is displayed in the report.

Tasks:

C2.1 Load the validated sample datasets (*AUS Retail Sales_sample.xlsx* and *AUS Retail Products_sample.xlsx*) into the associated *PowerBI* files.

- Rename the loaded sample datasets tables in *PowerBI Desktop* accordingly. (e.g. Sample of Sales, Sample of Products)

C2.2 Create new tables to segregate the sample datasets into separate target tables using DAX queries according to the *Source to Target Mapping* you've completed in task C1. In doing so, ensure that you:

- rename each new table with the relevant target output table name
 - use the correct DAX function to select required columns from the validated sample dataset whilst renaming and creating new target field/column names as required
- Important note: If you notice any anomalies or inconsistencies in the output data in the tables, do not try to fix them at this stage. You will be reporting on and fixing these issues and anomalies at a later task.**
- Once all the target tables have been created, select the option to **Hide** the sample dataset table (Sample of Sales, Sample of Product) from Report View.
 - for each new table created, provide screenshots separately for each dataset using the answer tables,
 - Table 6: Target output tables for Dataset1 (Transactional)
 - Table 7: Target output tables for Dataset2 (Non-transactional)
 - The screenshots should clearly show:
 - an expanded view of the *Fields* column in *PowerBI Desktop* showing the new tables created
 - the name of the PowerBI file in the “BSBXBD402 – Firstname_Lastname_DDMMYYYY” format on the title bar of PowerBI window.

C2.3 In the *model* view in *PowerBI* for each dataset,

- create a new tab and rename it to reflect the correct department name of the dataset
 - drag and drop the tables relevant for each department within each data model view tab
 - create the appropriate relationships between the tables
- Important: Select the recommended relationship type in PowerBI at this stage. If you notice any anomalies make a note of them as these will need to be addressed at a later stage.**
- provide a screenshot of each new data model view tab created in **Table 8: New data model views for each department**. Your screenshots should clearly show:
 - the *Data model* view tabs for each department with the department name.
 - the name of the PowerBI file in the “BSBXBD402 – Firstname_Lastname_DDMMYYYY” format on the title bar of PowerBI window.
 - the name of the PowerBI file displayed on the title bar as ‘Dataset#_MapReduce validation_YourNameInitials_DDMMYYYY’.

Evidence of performing the tasks:

In addition to the screenshots you will include in **Table 6**, **Table 7** and **Table 8** given below, your assessment submission must include the following documents in the ‘Phase 2 – MapReduce validation’ sub-folder. The PowerBI work files should contain evidence of implementing data segregation rules.

- ‘Dataset1_MapReduce validation_YourNameInitials_DDMMYYYY’
- ‘Dataset2_MapReduce validation_YourNameInitials_DDMMYYYY’.

Assessor instructions: Refer to the contents in the BSBXBD402_AG_03_Project_Exemplar (student submission folder) sample work files.

Table 6 - Target output tables for Dataset 1 (Transactional)

New table name:

Evidence of performing the task:
(Screenshots showing DAX functions/queries used to create the tables from the source table)

Order Detail

```

1 Order Detail = SELECTCOLUMNS('Sample of Sales',
2   "Row ID", [Row_ID],
3   "Order ID", [Order_ID],
4   "Order Date", [Order_Date],
5   "Customer ID", [Customer_ID],
6   "Product ID", [Product_ID],
7   "Quantity", [Quantity],
8   "Cost", [Cost],
9   "Discount", [Discount],
10  "Revenue", [Sales],
11  "Profit", [Sales]-[Cost]
12 )
    
```

Order ID	Customer ID	Product ID	Cost	Revenue	Profit
AU-2018-147627	HL-15040	FUR-FU-10003194	27.02	38.6	11.58
AU-2018-143637	MS-17710	FUR-FU-10002813	25.9072	40.48	14.5728
AU-2018-101462	BP-11230	FUR-FU-10000409	32.3568	59.92	27.5632
AU-2018-103429	LW-16825	TEC-PH-10003505	329.44	464	134.56
AU-2018-141796	JG-15160	TEC-PH-10001578	862.5435	1214.85	352.3065
AU-2018-124478	MA-17560	TEC-PH-10001128	215.9856	299.98	83.9944
AU-2018-131002	TB-21400	FUR-FU-10004665	608.1912	821.88	213.6888
AU-2018-109456	LS-17245	TEC-AC-10003610	93.5844	179.97	86.3856
AU-2018-106439	GG-14650	TEC-AC-10004568	204.0471	251.91	47.8629
AU-2018-151005	OH-18715	FUR-FU-10002948	27.0207	60.73	33.7093

Table: Order Detail (480 rows)

Figure 9 - Screenshot for Dataset1 Order Detail table using PowerBI Desktop © Microsoft

Order

```

1 Order = SELECTCOLUMNS('Sample of Sales',
2   "Order ID", [Order_ID],
3   "Location", [Country]&" "&[State]
4 )
    
```

Order ID	Location
AU-2018-156349	Australia,NSW
AU-2018-105767	Australia,NSW
AU-2018-140858	Australia,NSW
AU-2018-115812	Australia,NSW
AU-2018-165862	Australia,NSW
AU-2018-146283	Australia,NSW
AU-2018-138240	Australia,NSW
AU-2018-159618	Australia,NSW
AU-2018-126340	Australia,NSW

Table: Order (480 rows)

Figure 10 - Screenshot for Dataset1 Order table using PowerBI Desktop © Microsoft

New table name:

Evidence of performing the task:
(Screenshots showing DAX functions/queries used to create the tables from the source table)

Product

The screenshot shows the Power BI Desktop interface with the 'Table tools' ribbon active. The 'Name' field is set to 'Product'. The DAX query in the formula bar is:

```
1 Product = SELECTCOLUMNS('Sample of Sales',  
2     "Product ID", [Product_ID],  
3     "Product Name", [Product_Name]  
4 )
```

The data preview below the query shows a table with two columns: Product ID and Product Name. The table contains 360 rows of data, including items like 'Dixon Ticonderoga Core-Lock Colored Pencils' and 'Kensington 6 Outlet MasterPiece HOMEOFFICE Power Control Center'. The 'Data' pane on the right shows the 'Product' table selected.

Figure 11 - Screenshot for Dataset1 Product table using PowerBI Desktop © Microsoft

Customer

The screenshot shows the Power BI Desktop interface with the 'Table tools' ribbon active. The 'Name' field is set to 'Customer'. The DAX query in the formula bar is:

```
1 Customer = SELECTCOLUMNS('Sample of Sales',  
2     "Customer ID", [Customer_ID],  
3     "Customer Segment", [Segment]  
4 )
```

The data preview below the query shows a table with two columns: Customer ID and Customer Segment. The table contains 480 rows of data, all of which are 'Consumer'. The 'Fields' pane on the right shows the 'Customer' table selected.

Figure 12 - Screenshot for Dataset1 Customer table using PowerBI Desktop © Microsoft

Table 7 - Target output tables for Dataset 2 (Non-transactional)

New table name:

Evidence of performing the task:
(Screenshots showing DAX functions/queries used to create the tables from the source table)

Product

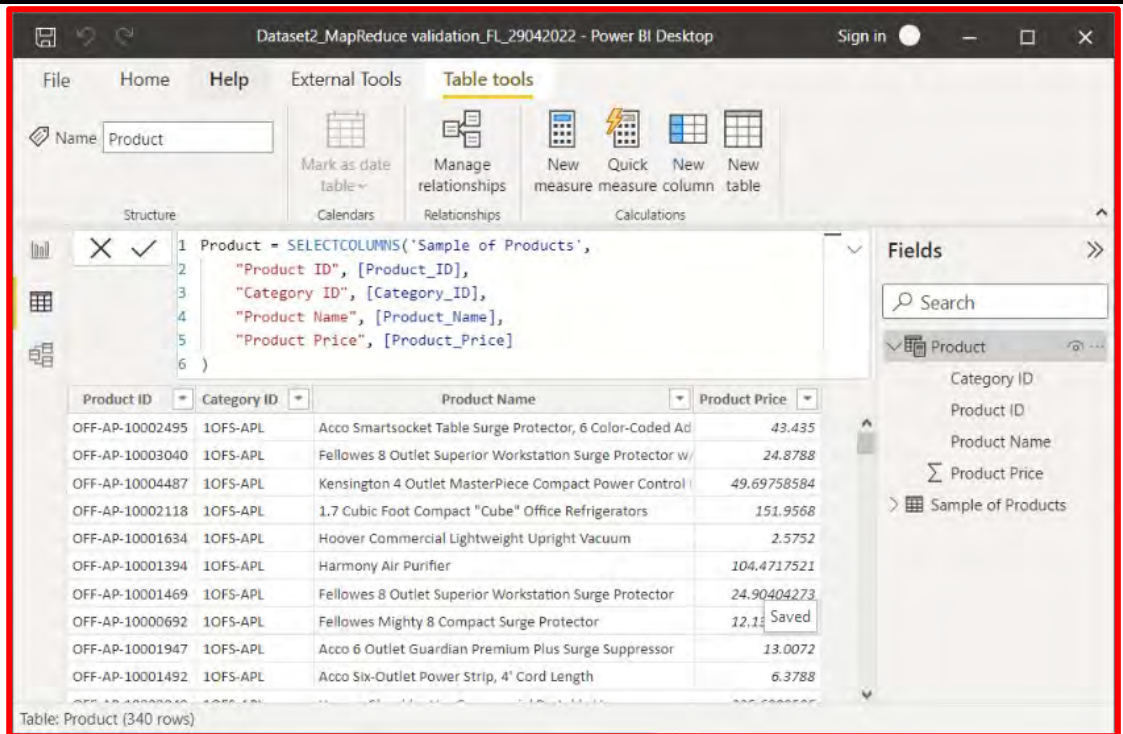


Figure 13 - Screenshot for Dataset2 Product table using PowerBI Desktop © Microsoft

Category

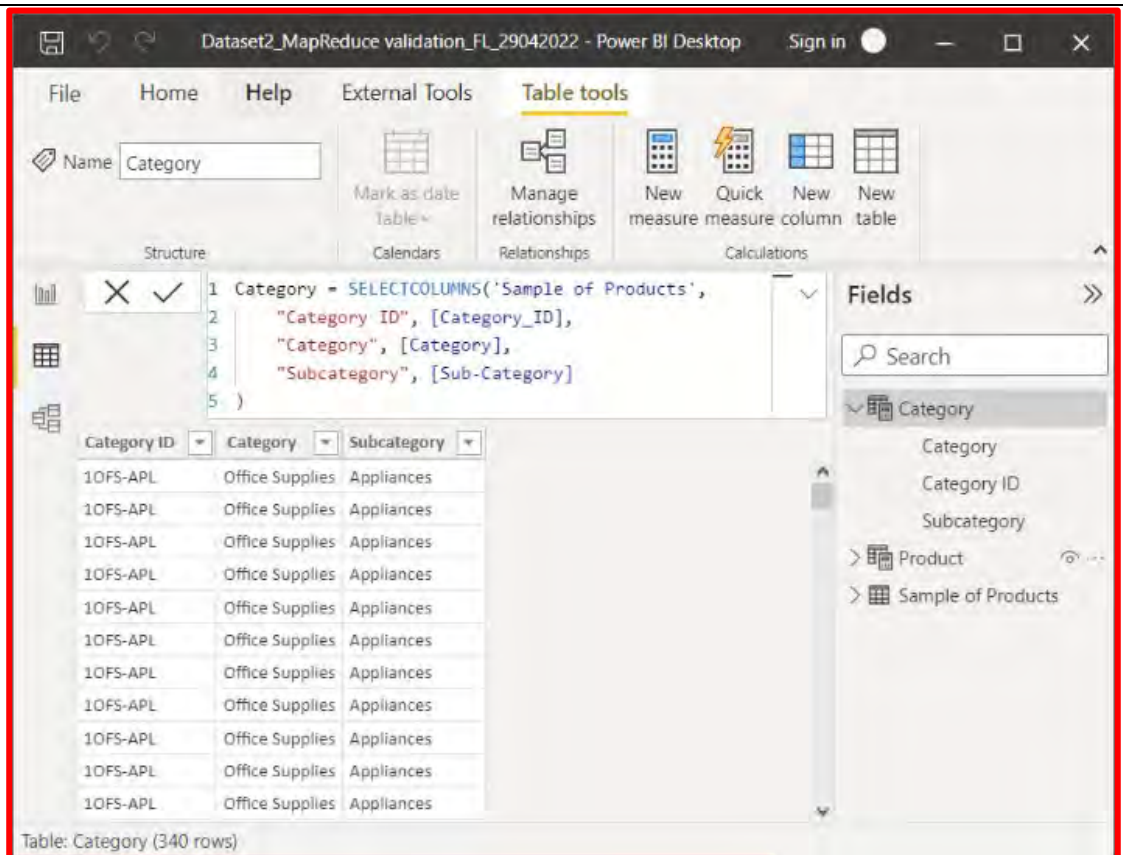


Figure 14 - Screenshot for Dataset2 Category table using PowerBI Desktop © Microsoft

Table 8 - New data model views for each department

New data model view name:

Evidence of performing the task:
(Screenshots)

Sales Department

Assessor guidelines:
The transactional dataset *Sample of Sales* should be aligned with the data types in the Sales Department

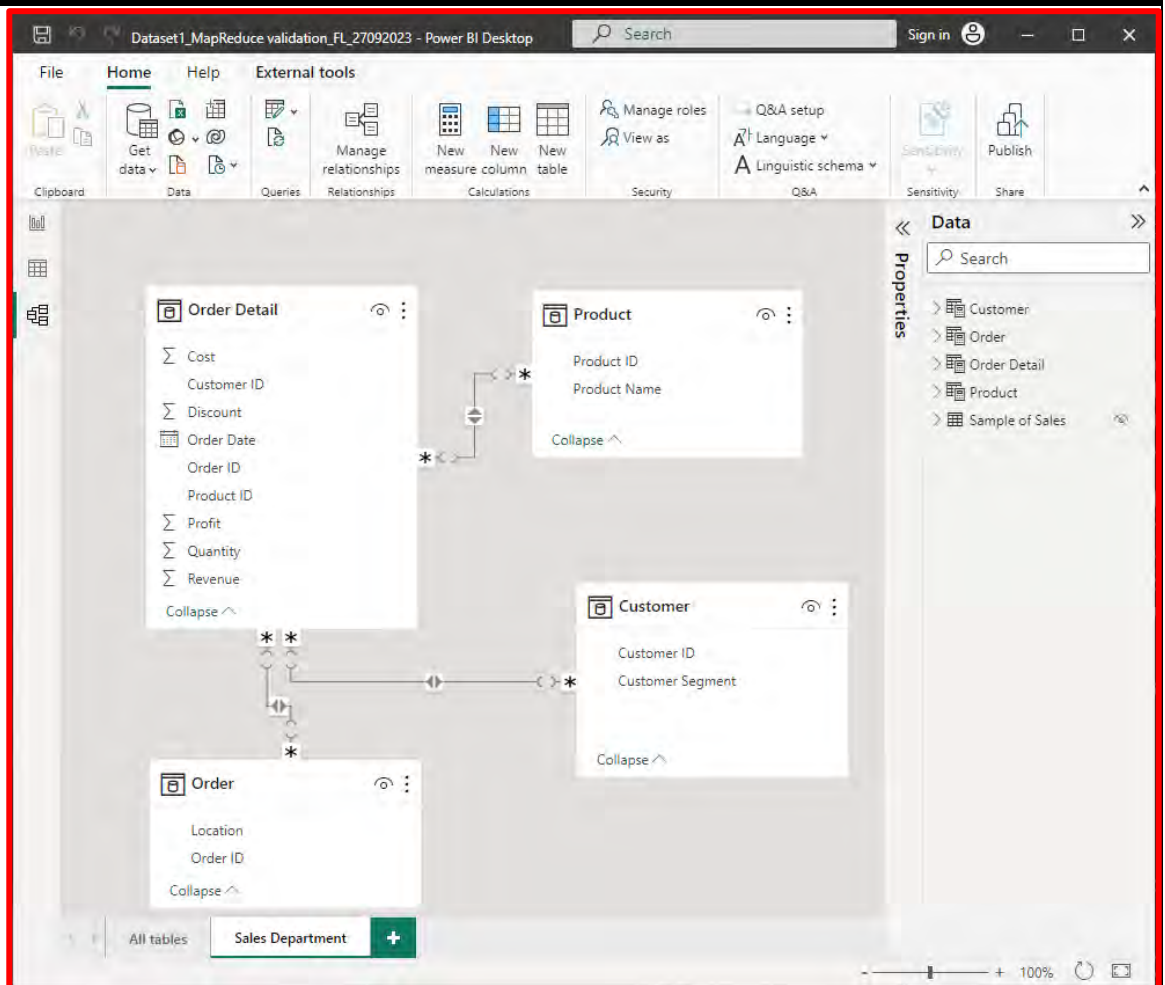


Figure 15 - Screenshot for Dataset1 Data model view using PowerBI Desktop © Microsoft

New data model view name:

Evidence of performing the task:
[Screenshots]

Production Department

Assessor guidelines:
Non-transactional dataset *Sample of Products* aligned with the data types in the Production Department

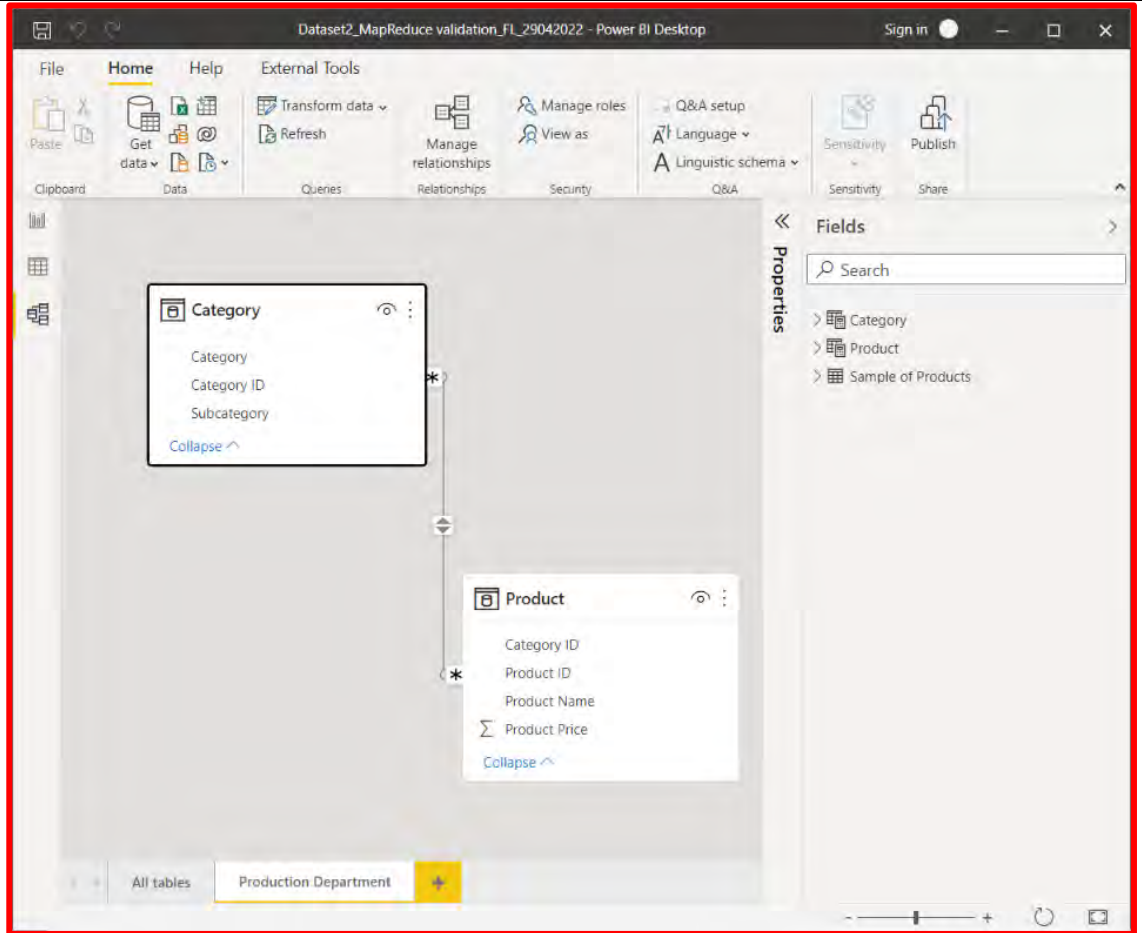


Figure 16 - Screenshot for Dataset2 Data model view using PowerBI Desktop © Microsoft

C3. Implement data aggregation rules

In this task, you are required to implement data aggregation and segregation rules on the small set of sample data and datasets with appropriate visualisations to display the required data.

Instructions:

Do this task using the PowerBI files created in task C1.

Refer to the **reporting requirements** outlined for *Sales* and *Production* departments in the **AUS Retail_Reporting requirements.pdf** document.

The screenshots you provide as evidence should clearly show the *Report* view tabs that are named appropriately to indicate which type of data that is displayed in the report.

Tasks:

C3.1 Do the following using the Dataset 1 PowerBI file:

- rename *Page 1* of the report view tab as 'Sales Report 1'
- use the correct visualisations and measures to implement the data aggregation rules for the *Sales* department according to the reporting requirements provided.

Assessor instructions: Students must do the following to implement data aggregation and segregation rules on the transactional dataset.

- Add a **Card** visual to display: *Cost, Revenue* and *Profit*.
- Add a **Clustered bar chart** visual to display: *Cost, Revenue* and *Profit*.
- Add a **Slicer** visual to filter report data based on yearly and quarterly.
- Add a **Map** visual to display *Revenue* details by *Location*.
- Add a **Donut chart** to display cost, revenue and profit for each product category.
- Add a **Stacked area chart** displaying total revenue and profit for each customer segment.

C3.2 Do the following using the Dataset 2 PowerBI file:

- rename *Page 1* of the report view tab as 'Product Report 1'
- use the correct visualisations and measures to implement the data aggregation rules for the *Production* department according to the reporting requirements provided.

Assessor instructions: Students must do the following to implement data aggregation and segregation rules on the non-transactional dataset.

- Add a **Scorecard** visual to display the total number of distinct products.
- Add a **Pie chart** visual to display [the percentage of distinct products in each sub-category].
- Add a **Matrix** visual to list the total number of distinct products in each category and Sub-category.
- Add a **Slicer** visual to filter report data based on the product category.

The student should create a measure called **Distinct Products – to get the distinct no. of products in each category**.

Evidence of performing the tasks:

In addition to the screenshots you will include in **Table 9** given below, your assessment submission must include the following documents in the 'Phase 2 – MapReduce validation' sub-folder. The PowerBI work files should contain evidence of implementing data aggregation rules.

- 'Dataset1_MapReduce validation_YourNameInitials_DDMMYYYY'
- 'Dataset2_MapReduce validation_YourNameInitials_DDMMYYYY'.

Assessor instructions: The student should provide two screenshots showing aggregated and segregated data in each department. Sample screenshots are given below.

Refer to the contents in the **BSBXBD402_AG_03_Project_Exemplar (student submission folder)** sample work files.

Table 9 - Evidence of performing demonstration task C3

Tasks: Evidence of performing the tasks:
(Screenshots)

Sales Report 1 report:
(The screenshot should show the required visualisations as per the reporting requirements provided)

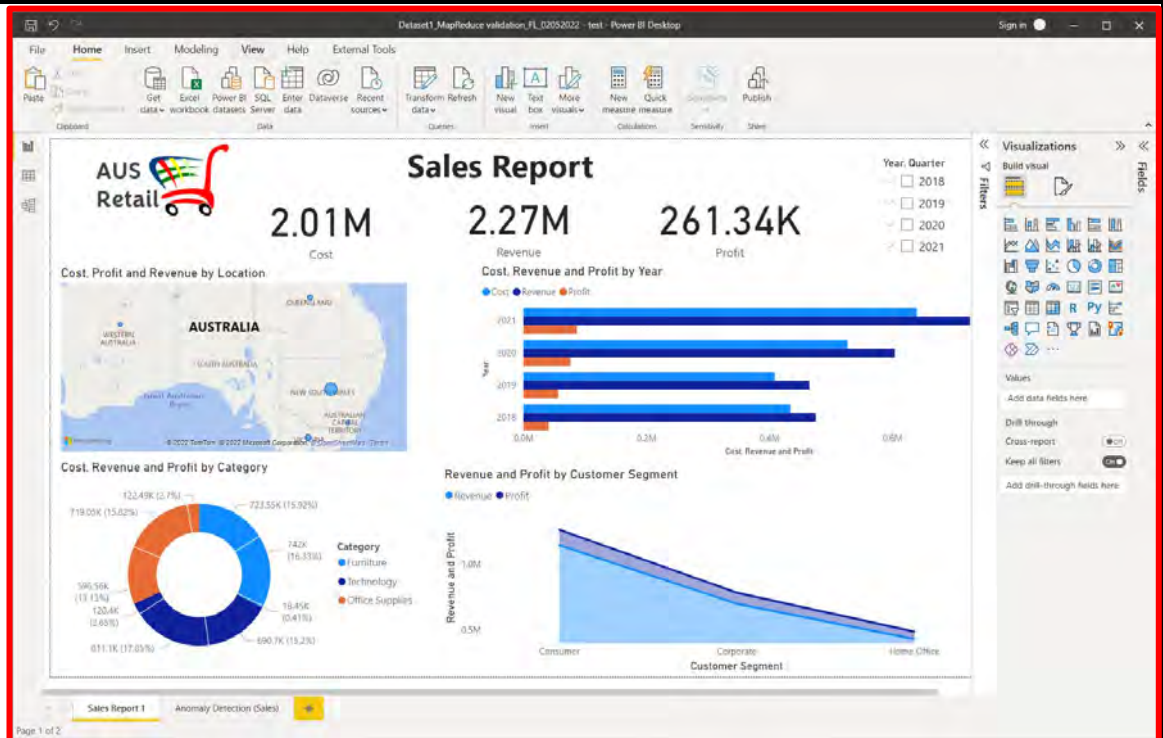


Figure 17 - Screenshot for Dataset1 Sales report view using PowerBI Desktop © Microsoft

Product Report 1 report:
(The screenshot should show the required visualisations as per the reporting requirements provided)

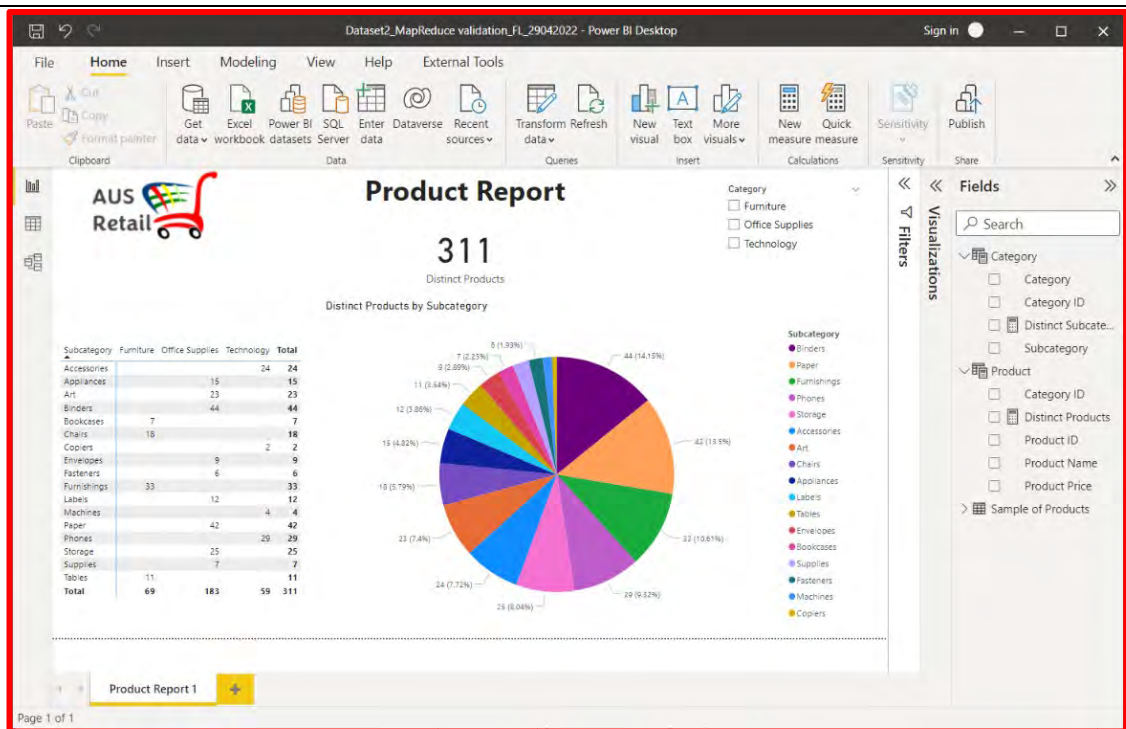


Figure 18 - Screenshot for Dataset2 Product report view using PowerBI Desktop © Microsoft

C4. Identify anomalies

In this task, you are required to further evaluate Dataset 1 (Sales) and Dataset 2 (Products) to identify anomalies in the aggregated data.

Tasks:

C4.1 Check for anomalies in Dataset 1 [Sales] by doing the following.

- a. Create a new tab in PowerBI Report mode called 'Anomaly Detection [Sales]'
- b. Create line charts to display each of the following time series data from Dataset1 [Sales]
 - **Cost** by Year, Quarter and Month
 - **Revenue** by Year, Quarter and Month
 - **Profit** by Year, Quarter and Month
- c. Use the *Find Anomalies* feature in PowerBI to detect any anomalies in the sales data.
- d. Provide a screenshot of the 'Anomaly Detection [Sales]' tab showing the detected anomalies for each line chart visualisation for **Cost**, **Revenue** and **Profit** in the answer table given below.

C4.2. Check for anomalies in Dataset 2 [Products] by doing the following.

1. Create a new tab in PowerBI Report mode called 'Anomaly Detection [Products]'.
2. Add Matrix visualisation to list the Product Name details by the number of Distinct Products. Ensure the Product IDs' and Product Price values are grouped within the Product Name lists so that any anomalies can be identified with specific product IDs and their prices.
3. If there is a value greater than 1 displayed for *Distinct Products*, that indicates an anomaly in the data.
Business logic: *One product ID should have one distinct product name with one standard price.*
4. Provide screenshot(s) of the detected anomalies in the Matrix visualisation for in the answer table given below.
Note: Expand the items that have an anomaly in the Matrix visual to obtain further details of the Product IDs and its price.

Evidence of performing the tasks:

In addition to the screenshots you will include in **Table 10** given below, your assessment submission must include the following documents in the 'Phase 2 – MapReduce validation' sub-folder. The PowerBI work files should contain evidence of identifying anomalies in each dataset.

- 'Dataset1_MapReduce validation_YourNameInitials_DDMMYYYY'
- 'Dataset2_MapReduce validation_YourNameInitials_DDMMYYYY'.

Assessor instructions: In addition to the sample screenshots provided in the answer table below, refer to the contents in the **BSBXBD402_AG_03_Project_Exemplar (student submission folder)** sample work files.

Table 10 - Evidence of performing demonstration task C4

Report view:

Evidence of detected anomalies in the data.
[Screenshots]

Anomaly Detection (Sales)

Assessor guidelines: The anomalies displayed may have a variation due to the sample dataset chosen by the student.

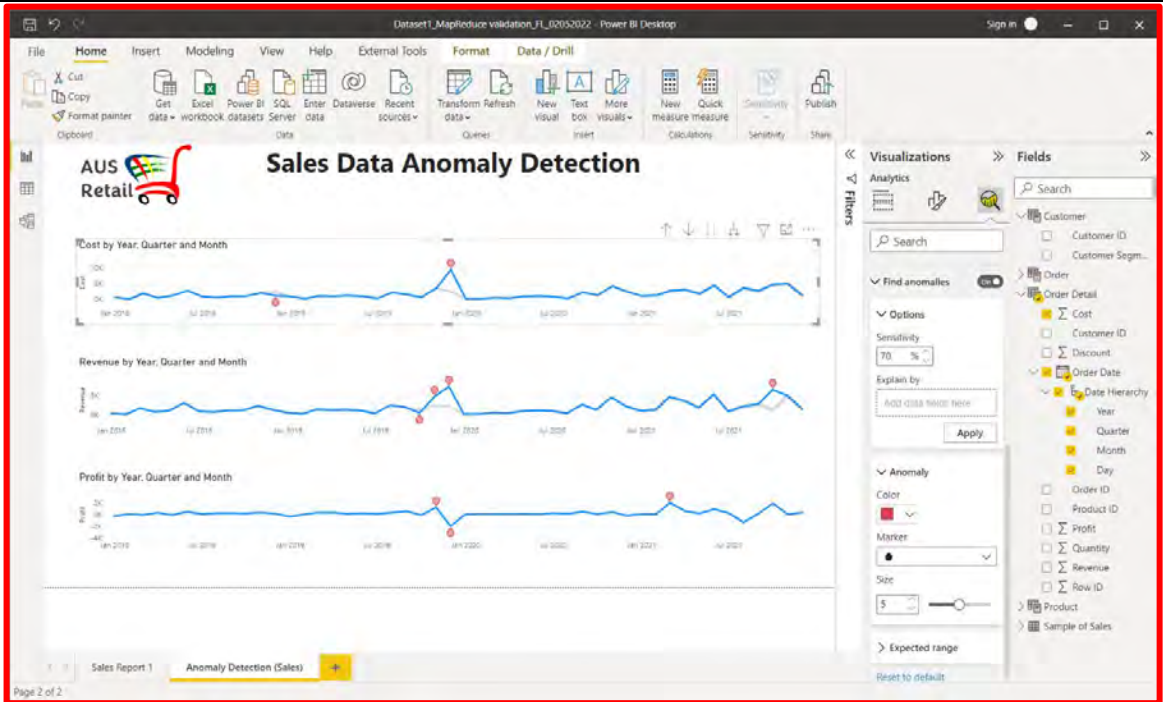


Figure 19 - Screenshot for Dataset1 Anomaly detection using PowerBI Desktop © Microsoft

Anomaly Detection (Product)

Assessor guidelines: The anomalies displayed may have a variation due to the sample dataset chosen by the student.

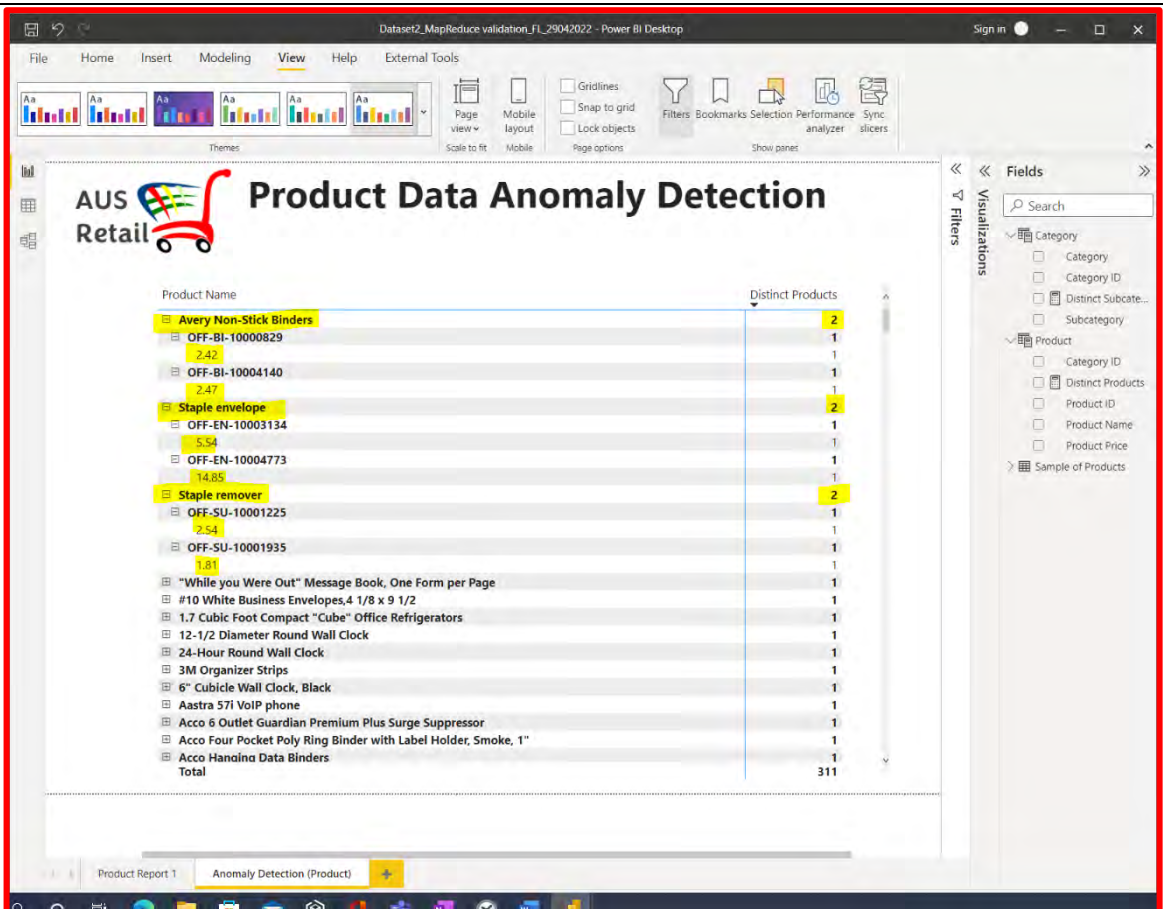
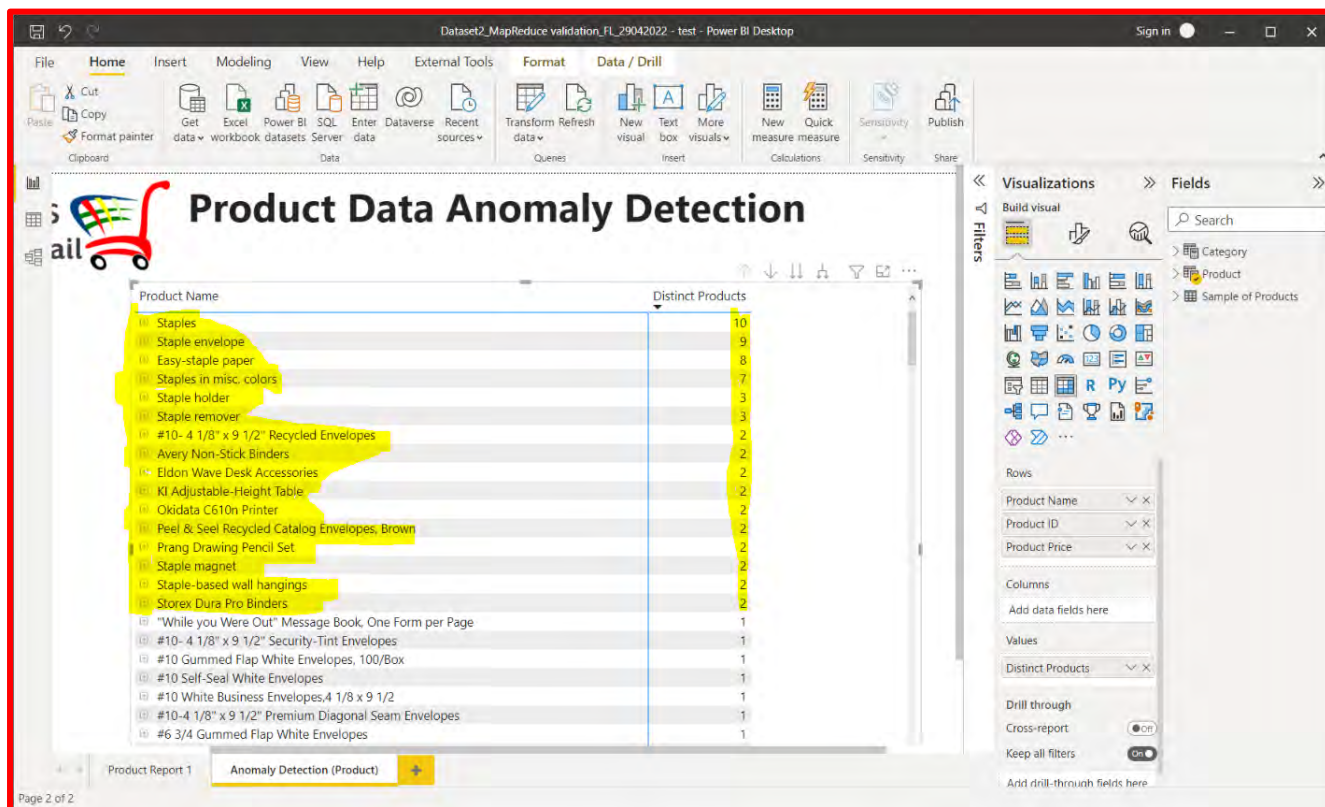


Figure 20 - Figure 17 - Screenshot for Dataset2 Anomaly detection using PowerBI Desktop © Microsoft

Assessor guidelines: The following variety of anomalies may be detected by the student. The screenshot below shows all possible anomalies in product-related data.



C5. Consult supervisor to clarify and resolve identified anomalies

In this task, you are required to consult your supervisor to clarify and receive advice on resolving the identified anomalies in the previous task C3.

Task:

Write a draft email addressed to your supervisor for the purpose of clarifying and obtaining advice on how to resolve the identified anomalies. When drafting the email, you must:

- briefly outline the details of the anomalies detected in the sales data and product data
- include relevant screenshots to clearly indicate the detected anomalies (highlight, circle, to draw attention to specific issues you've identified)
- use clear, specific and industry-related terminology when presenting your validation test results in your email
- use the email template given below.

Answer: Drafted email to Supervisor

Lastname, Firstname

From: Lastname, Firstname

Sent: Monday, 14 February 2022 10:44 AM

To: Lastname, Firstname

Subject: Sample Email Template

Dear [Name]

Email body goes here.

List Bullet

List Bullet

Kind regards

Firstname Lastname

Your role

Firstname.Lastname@ausretail.com.au



Before printing this email please consider the environment.

This message may contain privileged information or confidential information or both and is intended for the recipient named. If you are not the intended addressee, please delete it and notify the sender.

Assessor instructions: The email drafted by the student should indicate that:

- the email was addressed to the supervisor, **Chief Data Officer (CDO), Mia Gonzales**
- it contains all necessary information regarding the data anomalies identified in the previous task for the purpose of consultation to clarify and resolve issues.

Please note that the anomalies detected by the student may have differences depending on the representative sample they've chosen to conduct testing.

A sample answer is provided below.

Lastname, Firstname

From: Lastname, Firstname

Sent: Monday, 05 May 2022 12:53 PM

To: Lastname, Firstname

Subject: Sample Email Template

Hi Mia,

I have conducted some validation checks on the sales and product related datasets and have found the following anomalies in the data.

Refer to the circled areas in the screenshot given below, where it indicates that the total profit has an unexpectedly low figure (minus value) on the 25th of November 2020. This is resulting from the unexpectedly high costs on the same day and very low sales figures on the same day. A similar issue occurred on the 02nd of October 2021 as well.

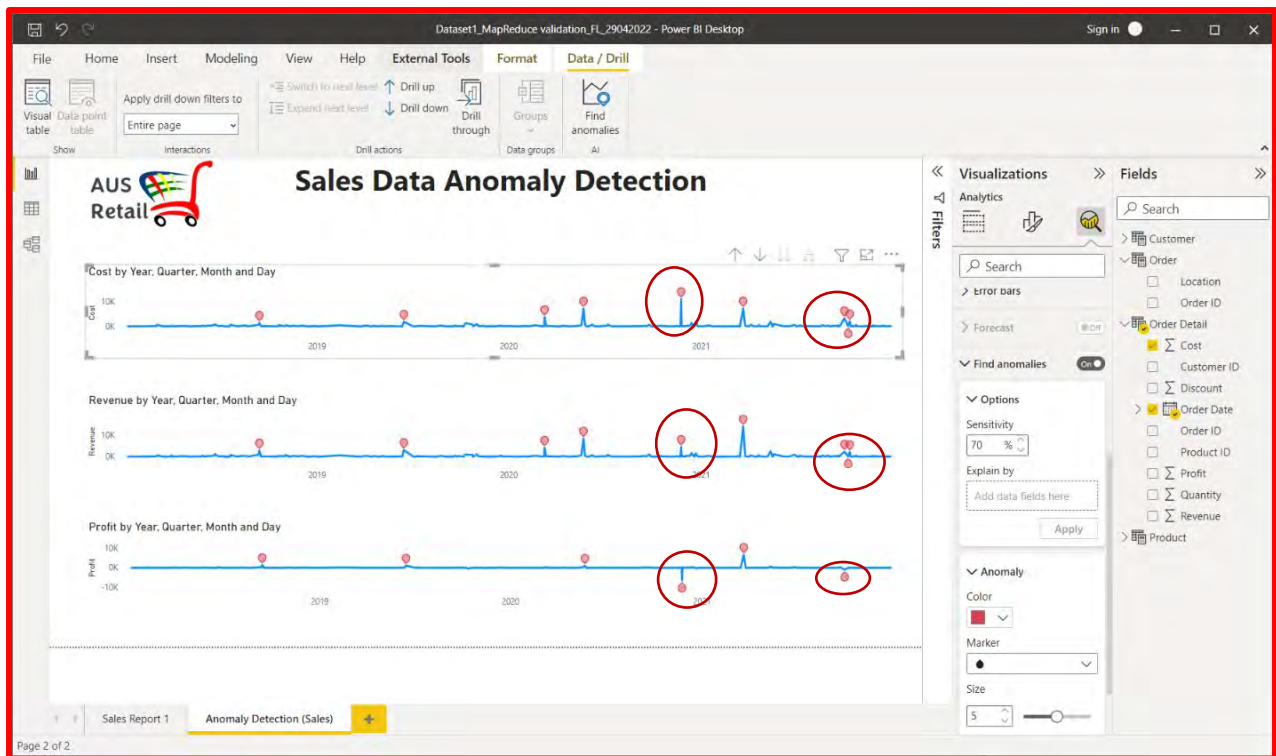


Figure 21 - Anomaly detection in Sales Data using PowerBI Desktop © Microsoft

Concerning the product related data, there are multiple products exist that have the same exact name but different Product IDs and Prices.

The product names that show this anomaly are as follows.:

- Avery Non-Stick Binders
- Staple envelope
- Staple remover

The anomalies found in the product dataset are highlighted in the screenshot given below.

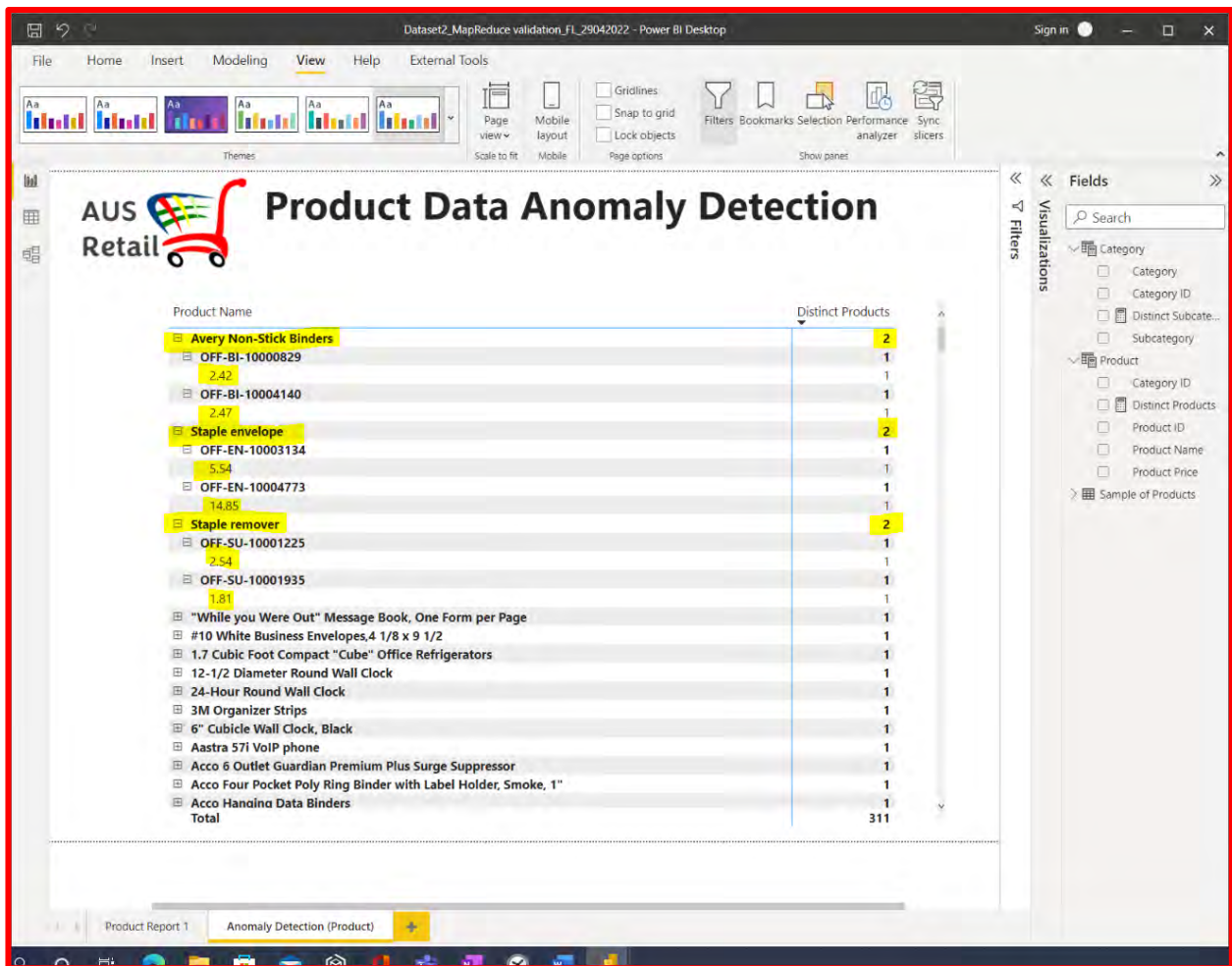


Figure 22 - Anomaly detection in Product Data using PowerBI Desktop © Microsoft

Please do let me know how these anomalies can be resolved before moving forward in the data validation process.

Thanks and kind regards

Firstname Lastname

Trainee analyst

Firstname.Lastname@ausretail.com.au



Before printing this email please consider the environment.

This message may contain privileged information or confidential information or both and is intended for the recipient named. If you are not the intended addressee, please delete it and notify the sender.

Assessment checklist:

Students must have completed all activities within this assessment before submitting. This includes:

Part B: Validate assembled or obtained big data sample		
B1	Table 1- Sampling strategy for Dataset 1 (Transactional) Table 2 – Sampling strategy for Dataset 2 (Non-transactional)	<input type="checkbox"/>
B2	Table 3 – Evidence of performing demonstration task B2	<input type="checkbox"/>
B3	Table 4 – Evidence if validating Dataset 1 (Transactional) Table 5 – Evidence if validating Dataset 2 (Non-transactional)	<input type="checkbox"/>
Part C: Validate big data sample process and business logic		
C1	Excel templates, <i>Source to Target Mapping</i> tab: <ul style="list-style-type: none">• <i>AUS Retail_STM&TestCase_Dataset1(Sales)_YourNameInitials_ddmmyyyy.xlsx</i>• <i>AUS Retail_STM&TestCase_Dataset2(Products)_YourNameInitials_ddmmyyyy.xlsx</i>	<input type="checkbox"/>
C2	Table 6 – Target output for Dataset 1 (Transactional) Table 7 – Target output for Dataset 1 (Non-transactional) Table 8 – New data model views for each department (Sales and Production)	<input type="checkbox"/>
C3	Table 9 – Evidence of performing demonstration task C3	<input type="checkbox"/>
C4	Table 10 – Evidence of performing demonstration task C4	<input type="checkbox"/>
C5	Email to Supervisor – email draft for clarification and resolution advice.	<input type="checkbox"/>



Congratulations you have reached the end of Assessment [3]!

© UP Education Online Pty Ltd 2022

Except as permitted by the copyright law applicable to you, you may not reproduce or communicate any of the content on this website, including files downloadable from this website, without the permission of the copyright owner.

References:

Learning Container. 2020. *Sample sales data excel xls*. [online] Available at: <https://www.learningcontainer.com/download/sample-sales-data-excel-xls/> [Accessed 04 April 2022].