Activity 2 – Data quality dimensions investigation challenge

Learner guide

CONTENTS

How to Use This Workbook

Activity

Alongside this icon you'll find details of the group/individual activity or a point for everyone to discuss.

Important Idea or Concept

Generally, this icon is used to draw your attention to ideas that you need to understand by this point in the course. Let your trainer know if you do not understand or see the relevance of this idea or concept.

Key Point

This icon is used to indicate something that practitioners in this field should know. It's likely to be one of the major things to remember from the course, so check you do understand these key points.

Definition

Where a word with a very specific definition (or one that could be described as jargon) is introduced this will highlight that a definition is provided. (These words will also be found in the Glossary at the back of the workbook.)

Useful Tool

This icon indicates a technique that will help you put what you learn into practice.

Helpful Hint

This icon guides you to tips or hints that will help you avoid the standard pitfalls that await the unwary practitioner or to show you how you might increase your effectiveness or efficiency in practising what you have learnt.

Reference Material

When we have only touched briefly on a topic this icon highlights where to look for additional information on the subject. It may also be used to draw your attention to International or National Standards or Web addresses that have interesting collections of information.

Warning

This icon is used to point out important information that may affect you and your use of the product or service in question.

Introduction

In this activity, your goal is to analyse a dataset and find as many issues as you can regarding the **data quality dimensions** previously discussed.

Prerequisite knowledge

Familiarity with Microsoft Excel pivot tables, creating charts, and Excel functions, such as SUM, SUMIF, and IF functions. Ensure you have completed all activities in Module 1 and Module 2 regarding these topics.

Resources

You will need the following files to complete this activity:

- Data Quality Dimensions Investigation Challenge.pdf (this document)
- QAdbury Shop.xslx
- Data Quality Dimension definitions.pdf

Scenario

A fictional chocolate shop named QAdbury began trading on **11 July 2019.**

Since opening, QAdbury has maintained a record of sales, which runs from **11 July 2019 to 2 May 2020.** An analyst was hired to analyse QAdbury's sales data and to generate a report. The analyst submitted their report on **12 July 2020.** It was based on the entire Sales table.

The shop **opens at 11:00** each morning. The closing time is not known. However, 95% of the time, the last sale of the day occurs before **17:20.** The shop is usually **closed on Tuesdays.**

QAdbury dataset

The Microsoft Excel spreadsheet **"QAdbury Shop".xlsx** will used to complete this activity, including the following tabs:

- Sales
- Report
- Prices

The **Sales** table records the date and time of each sale, the product that was sold, the quantity sold, and the customer's name.

The **Prices** table records the current and historical price of each product. Both the **Sales** table and the **Prices** table may exhibit data quality issues.

The **Sales** table is based on real sales data from a publicly available data set. The labels have been changed, and some data issues have been added manually. The **Prices** table is fabricated.

The analyst's **report** may also contain data quality issues, as well as mistakes made by the analyst themselves.

Analyse the data any way you like. This might include manual inspection, or the creation of charts, formulas, and pivot tables. Some of the data quality issues are in plain sight, while others will require some detective work to find.

It is recommended that you work on a copy of the Excel spreadsheet.

Task 1

Group activity

SUMIF and SUM functions or Pivot table

- Briefly browse the **Sales table** on the "Sales" tab. There is a noticeable data quality issue with the sale value of one particular product. Find a way to fix this data quality issue.

- Verify that the sum of sales for that product is now **£3,531.67**.

Helpful Hint

One option is to use the SUMIF() function or create a pivot table.

- Verify that the total sum of sales for all products is now £10,279.83.

- **Ensure that you have verified these sums before you move on to Task 2.**

Helpful Hint

Refresh the pivot tables on the "Report" tab by going to > **Data Ribbon** > **Queries & Connections** group > **Refresh All.** Alternatively, press **Ctrl + Alt + F9**. (On some laptops, **Ctrl + Alt + Fn + F9.**)

Task 2

Group activity

IF, SUMIF, and SUM functions OR Pivot table

- Look carefully at the Sales table. There is still a data quality issue with the sale value of the product you identified in Task 1.
- Find and fix this data quality issue. You may wish to create a new column.

Helpful Hint

Look at the “sale value” and the quantity (“Qty Sold”) of that particular product.

- Verify that the sum of sales for that product is now **£5,779.91.**
- Verify that the total sum of sales for all products is now **£12,528.07.**
- **Ensure that you have verified these sums before you move on to Task 3.**

Helpful Hint

Refresh the pivot tables on the "Report" tab by going to > **Data Ribbon** > **Queries & Connections** group > **Refresh All.** Alternatively, press **Ctrl + Alt + F9**. (On some laptops, **Ctrl + Alt + Fn + F9.**)

Before you move on to Task 3, ensure that you have thoroughly read the ‘Scenario’ and ‘QAdbury Dataset’ sections. Use the information to guide you in completing Task 3.

Task 3

Group activity

Data quality dimensions

View the analyst's report on the 'Report' tab. The report is divided into the following six sections, with each section consisting of a chart, analyst's comments, and the source data for the chart. The source data is a table or pivot table drawn from the **Sales table** in the 'Sales' tab.

- Sales by month
- Sales by day of week
- Sales by customer
- Sales by time of day
- Total sales by product
- Quantity sold by product

- For each of the six sections, note any instances of bad/incorrect analysis or misleading data visualisation that you discover.

- Note any source data that you find that fail to meet one of the **data quality dimensions.**

Reference material

Please refer to the 'Data Quality Dimension definitions' pdf to support you.

Task 4

Group Activity

Data quality issues

- Browse the ‘Prices’ tab and note any data quality issues that you find, including discrepancies with the Sales table. In the event of a discrepancy, you may treat the Sales table as ‘ground truth’.

Helpful Hint

Note: if you changed any data in the sales table, you will need to update the pivot tables by going to > **Data Ribbon** > **Queries & Connections** group > **Refresh All.** Alternatively, press **Ctrl + Alt + F9**. (On some laptops, **Ctrl + Alt + Fn + F9.**)