# Dataset Improvement Report

2024-08-07

## Introduction

This report highlights key issues within the dataset and provides recommendations for improving its usability, for automation and analysis.

## Issues Identified and Recommendations

The dataset must be more consistent across sheets, as varying structures currently complicate data processing. To address this, the sheets could be consolidated into a single sheet with a date column and standardised formats, making the data easier to read and reducing errors when carrying out calculations.

Merged cells, especially in the "Unit price" column, disrupt uniform data processing, making calculations difficult. Avoiding merged cells and ensuring each row contains complete data entries would allow more accurate analyses. Additionally, inconsistencies in category naming and invalid entries like "0" and "nan" complicate data aggregation. Standardising category names and removing invalid entries is needed to reduce the time spent fixing inconsistencies.

The lack of a year column is an issue for time-series analysis, leading to potential confusion. Adding a year column would provide the necessary context for accurate temporal analysis. Some values, such as the "Total in May" and "Price" rows, are unclear and require better documentation or restructuring, possibly in a summary tab, to improve clarity.

Reliance on manual data entry and Excel-based calculations increases the risk of errors. Automating these processes using scripts and transitioning to an API-based approach would enhance accuracy and efficiency, addressing the limitations of reliance on Microsoft Cloud services.

Lastly, the current dataset structure may not scale well with future data additions. Implementing a more robust database system or standardised templates will ensure the dataset remains manageable as more data is collected.