# Dataset Improvement Report

2024-08-07

## Introduction

In this report, I analyse a dataset comprised of Data from Fareshare that tracks food donations across various store locations for the Harper Adams University community fridge initiative. The dataset, composed of multiple sheets with varying structures, records food items received and distributed, including categories, quantity, weight, and cost. However, the current state of the dataset presents several challenges for efficient data processing and analysis. I aim to identify these challenges and provide recommendations to enhance the dataset by standardising the data structure, improving data accuracy, and attempting to enable automation for more time-efficient data handling. Addressing these issues can transform the dataset into a more reliable tool when monitoring food donations.

## Issues Identified

The dataset contains inconsistent data structures across various sheets, complicating data processing and analysis. For instance, the Waitrose data is structured differently than the other sheets, making it challenging to standardise and aggregate the data.

Merged cells further disrupt data processing, leaving every other row empty except for the "Unit price" column, which is filled with "GBP." This inconsistency complicates automated calculations and data extraction, especially as some rows contain more merged cells than others.

Inconsistent category naming and invalid entries such as "0" and "nan" present additional challenges. Different sheets use varying names for the same categories, with spelling mistakes and other errors within the category column further complicating data analysis.

The absence of a year column hinders accurate time-series analysis, particularly when dealing with data spanning multiple years. Also, certain entries, such as the "Total in May" and "Price" rows, are challenging to interpret due to unclear value calculations and inadequate documentation. These ambiguities can increase the risk of data analysis errors and confuse those reviewing the data.

The dataset also heavily relies on manual data entry and Excel-based calculations, which heightens the risk of errors and can be more time-consuming than using an API.

Additionally, with inconsistencies and manual processes, the current dataset structure may not scale well with future data additions, posing long-term usability challenges. As more data is collected, the lack of standardisation will make it increasingly challenging to manage and analyse the dataset effectively, highlighting the benefits of transitioning to an API-based approach.

## Steps to create a clean dataset:

I began by restructuring the dataset in Excel, removing empty columns and rows to ensure that the column names were positioned in the top-left corner, making the data easier to process programmatically. Next, I standardised column names across the dataset to ensure consistency to allow data processing and analysis.

I filled in dates where necessary to address missing values and removed any empty rows. Merged cells that complicated data processing were unmerged, and the data was reorganised to ensure that each row contained complete and independent entries.

Data from different sheets were consolidated into a single sheet, where an added Store column is used to keep track of where the data originated from. Columns were renamed for consistency, and missing values, such as total weight, were calculated when necessary.

The Unit column was standardised by converting any entry marked as "Kg" to "Item" to represent individual items while retaining "Crate" for larger groupings. Inconsistent category names and invalid entries were corrected or removed, and the categories were refined to ensure they fell into the main five categories.

Finally, I ensured that numerical data columns were correctly formatted as numbers rather than strings. The cleaned dataset was sorted by date to ensure chronological order and saved for further analysis and reporting.