# Dataset Improvement Report

2024-08-07

## Introduction

This report identifies issues and proposes recommendations to enhance usability, especially for automation and analysis purposes.

## Issues Identified and Recommendations

This dataset lacks consistency across sheets. While some sheets have the same structure, mainly the Market Draton sheets, each sheet within the spreadsheet still differs, leading to inconsistencies and making systematic data processing and analysis difficult. To address this, consolidating all monthly sheets into a single sheet with a date column, where possible, would streamline the data, making it more manageable and reducing the chances of errors. Furthermore, if all the data were consolidated, an additional column could be used to identify the location of the food. In cases where the data comes from separate locations with differing structures, it is crucial to standardize the format to minimize the need for extra manipulation.

Another significant issue is the presence of merged cells, particularly in the "Unit price" column. Merged cells complicate data extraction and manipulation programmatically, disrupting the uniformity needed for scripts to process the data efficiently. For instance, the "Unit price" values are not uniformly placed within the cells, making accurate calculations and analyses difficult. Avoiding merged cells and ensuring each row has complete and independent data entries would facilitate easier data processing.

Additionally, the dataset lacks a year indication, crucial for time-series analysis and maintaining historical records. The absence of this information can lead to confusion and misinterpretation, especially when dealing with multiple years. Including a year column alongside the existing date information will provide the context for accurate temporal analysis.

Moreover, some values within the cells are challenging to interpret. For instance, the "Total in May" and "Price" rows are not immediately clear in their context or basis for calculation. Better documentation or a more transparent calculation method is needed to ensure these rows are correctly understood and utilized. Adding comments or additional columns to explain these entries' purpose and calculation method will help clarify their context. They could also be moved to a summary tab, and creating a data dictionary could be helpful for clarification.

Another issue is reliance on calculations performed directly in Excel. This practice can introduce human error and inconsistencies. Performing these calculations programmatically using scripts would ensure greater accuracy and reliability. Furthermore, due to accessibility problems with OneDrive, relying on the Microsoft Cloud for automation is not feasible. An API-based approach would be more reliable and efficient for handling data automation.