Data Intake Report

Name: **Retail Forecasting**
Report date: **1/19/2024**
Internship Batch: **LISUM28**
Version:**1.0**
Data intake by: **Rana Fakeeh**
Data intake reviewer:
Data storage location:
https://docs.google.com/spreadsheets/d/1sOTsmkY4ZeNzww_yDGePGYt1iXtZjNHb/edit?usp=sharing&ouid=110600711982317630177&rtpof=true&sd=true

Tabular data details:

Table 1 - retail.csv

Total number of observations	1,218
Total number of files	
Total number of features	12
Base format of the file	.xlsx (converted into .csv)
Size of the data	50.3 KB

Proposed Approach:

- Dedup validation (identification):
    1. For each product in the dataset, the ***date*** column is the primary key.
    2. The uniqueness of the primary key is validated by invoking ***is_unique*** property on the ***date*** column.

- Mention your assumptions (if you assume any other thing for data quality analysis).
    1. The data lacks information about the company's super-markets therefore the sales for each product is assumed to be the aggregated sales.
    2. The multi-product time-series dataset is split into six separate single-product time-series.
    3. The ***Price Discount (%)*** feature is converted into float type.
    4. The ***date*** feature is converted into datetime type, spans from **2/5/2017** to **12/27/2020**, and set as index.
    5. Time-series data is resampled to week dates and week starts on ***Sunday***.
    6. All six single-product time-series are confirmed to have equal time frequency except for one product named ***SKU6*** which misses the last six weeks as compared to all other products. Missing week dates are appended to ***SKU6*** time-series.
    7. The data has no null values found in any feature except for ***SKU6*** time-series which have null values in all features at the last six week dates newly appended.