Instructions

I1. Refer link at https://archive.ics.uci.edu/ml/datasets/Online+Retail+II. When you click data folder it will take you to https://archive.ics.uci.edu/ml/machine-learning-databases/00502/ and there is an excel file named "online_retail_II.xlsx" for the data. It has two sheets with each one having a dataset.

I2. Use any programming language (preferably Python) for coding, any BI tool or programming language to generate visualizations/dashboards (you may use PowerBI Desktop which is free to create reports).

I3. It may be easier to load the dataset to a SQL table (although it is not required) in a database like MySQL, PostgreSQL, Redshift, Snowflake, SQLite etc.) in order to do some data manipulations below. In those cases, mention what database technology was used and provide the SQL scripts.

Assignment

A1. Append the data sets together from the two the sheets. You should have ~1 million records.

A2. Create a "profile" of the data set (A1). That is, generate some stats like min, max, mean, missing counts, number of unique values etc.

A3. Create a StockCode, Invoice Year, Invoice Month level data set. That is, one row will represent unique StockCode, Invoice Year, Invoice Month combination. E.g. If a StockCode had invoices for 3 months for a particular year and 4 months for another year, then there should be only 7 records for that StockCode in the summarized data set.

A4. When creating the above dataset (in A3), the other attributes need to be summarized. In addition to the summarized columns you create, below columns needs to be created (for the StockCode-InvoiceYear-InvoiceMonth combination).

   a. Number of unique customers
   b. Number of unique prices
   c. Customer ID who purchased the highest quantity. If there are multiple customers qualified, choose the minimum Customer ID.
   d. Number of price changes
   e. Weighted average of the price (using quantity)
   f. Amount (quantity*price) below weighted average price (which calculated in e)
   g. Amount (quantity*price) above weighted average price (which calculated in e)

A5. Give 5 insights which are interesting to you from the above dataset (A4).

A6. Referring to the main dataset in A1, create below grouping using the description column. Output should be a dataset with unique group column and corresponding description column and the corresponding StockCode.

   a. Group based on the core item. For example, below items are all pencils.

Group	Description	StockCode
Pencil	12 PENCILS SMALL TUBE SKULL	20974
Pencil	12 PENCILS TALL TUBE POSY	20984
Pencil	36 PENCILS TUBE POSY	20980

   b. Group this column based on some other criteria you think which will be useful when thinking about summarizing data across the grouping.

Deliverable definitions:

Code: Form of text file or Jupyter Notebook

Visualization: Excel file or Jupyter Notebook or pdf or image generated by a BI tool. Visualization could be a table with numbers, graph, dashboard etc.

Dataset: csv file with pipe delimited

Deliverables:

A1: Code. Mention total row count

A2: Code and Visualizations

A3,A4: Code (including SQL scripts if used) and Dataset

A5: Visualizations, bullet points

A6: Code and dataset. Not looking for 100% accurate grouping here. The method used will be assessed.