Start from the cvs data file

Extract the data from the source, and store in a dataframe object

Transform

1. Create column “product name”

2. based on the new column “product name”, regenerate the data of column “unit price

3.recalculate the "Total" & "Tax 5%" column's data

4.regenerate “gross margin percentage” column’s data based on its product line

5.we recalculate the value of column (Cost of goods sold ) & gross income based on the new value of gross margin percentage

6. Check the data type of each column

7.combime "Date" & "Time" column and transform it to datetime type and create a new row

8.Check for null values after merge

9.check for types' uniqueness

10.Delete repeat record with the same product name (but has different other attributes), and reset index

11.round all float64 type to 2 decimal precision

12. data discretization

13.add a new column "Brand" which can provide a new aspect for analysis

14.Create a column called "Weekday" to indicate what day of the week it is

15.Create 2 column called "Is_holiday" & "Holiday name" to indicate what holiday it is

16.format all the column names and let the first letter of each word in the column name to be lowercase #and replace the " " (space) with "_" (underscore)

LOAD

17.create sub dataframes for each dimension table

18.establish the connection to the PostgreSQL database

19. load the data for each dimension table

20.create sub dataframe for fact table, and load the data

END

Data Quality Issues Encountered and Handled:
Incomplete Data:

Issue: The original dataset lacked detailed product names and corresponding prices and gross margin percentages.
Resolution: Generated product names using GPT-4 and estimated prices and gross margin percentages for each product to enrich the dataset.
Data Transformation and Standardization:

Issue: The dataset required transformation to match the expected schema for analysis, including formatting dates and creating a composite 'datetime' column from separate date and time columns.
Resolution: Combined the 'Date' and 'Time' columns to create a new 'datetime' column, ensuring the data type aligns with the database schema.
Missing Category Labels:

Issue: Categorical data for product ratings and sales categories were missing.
Resolution: Implemented binning strategies to categorize continuous variables such as sales and ratings into discrete categories.
Column Naming Consistency:

Issue: The naming convention for columns was inconsistent, with mixed cases and underscores.
Resolution: Formatted all the column names to have the first letter of each word in lowercase and replaced the underscores with spaces to maintain a consistent naming convention.
Data Integration:

Issue: Needed to incorporate data from GPT-4 generated attributes into the original dataset, ensuring seamless integration.
Resolution: Integrated the new attributes by appending them to the original dataset and ensuring that the merged data maintains the integrity of the original data structure.
Noisy or Erroneous Data:

Issue: The potential presence of noise or errors in the newly generated data (e.g., prices and gross margins) due to estimations.
Resolution: Applied data validation checks to ensure that generated data falls within reasonable ranges and follows expected patterns. Adjustments were made as necessary.

Throughout the process, the goal was to handle these issues in a way that would make the dataset usable for loading into a DBMS for further analysis.