Review the unstructured csv files and answer the following questions with code that supports your conclusions:

- **Are there any data quality issues present?**

Data issues:

1. There are missing values in all three csvs.
2. The date columns are not stored in datetime format.
3. There are less than 1% duplicate entries in both products and transactions tables.
4. FINAL_QUANTITY appears to be a numeric column but there are some string values (zero)
5. FINAL_SALE appears to be numeric column but there are a few empty strings.
6. Even after removing duplicates there were multiple entries for a single RECEIPT_ID. There a few entries where FINAL_QUANTITY is zero but FINAL_SALE is non zero. Similarly there are few entries where FINAL_SALE is zero but FINAL_QUANTITY is non zero. Upon looking deeper into the tables, it appears that rows with either FINAL_SALE or FINAL_QUANTITY as 0 are redundant and not adding any value. If we remove such entries we end up filtering out more than 50% of the rows.

- **Are there any fields that are challenging to understand?**

FINAL_SALE AND FINAL_QUANTITY fields have numeric values but we are unable to determine the metric for these columns.

Interesting Trend:

1. It seems FL is the state where most receipts are being scanned
2. It appears that most of our users are aged 30 plus.
3. Females are the most loyal customer base, 82% of users are females.
4. 95% of our users have English as their language.
5. "Snacks", "Health & Wellness" and "Beverages" are the three most popular categories, "Snacks" being the most popular and by a large margin.
6. Pepsi co is most popular manufacturer.
7. Most of the users are scanning their receipts within 7 days of purchases.

8. **Walmart is the most popular store in all regards be it sales value, sales quantity or sales volume. However, if we look at sale value Costco is the store with highest sale but it doesn't even rank 10th in total quantities**

Request for additional information:

What are the metrics that we are trying to record and evaluate.

How is the data stored in databases?

I would like to look at the data cleaning or data pipelines and the data dictionaries if available.