**Project Notes**

**<u>Data Analysis Outline:</u>**

1. Explore & Clean Data
2. Create Aggregated Data for K-Mean Clustering in SQL
3. Feature Engineering (Performing K-Mean Cluster)

a. Exclude outliers.

b. Scale Data

c. Determine # of Clusters.

d. Cluster

4. Use Segments for Datasets for the Metrics mentioned above.

**<u>EDA Process:</u>** (Detailed Notes in the Markdown) (Python)

Noteworthy concerns:

- CustomerID contains Nulls
- Both Quantity and UnitPrice contain negatives.
    - Checked the values of negative Quantity, InvoiceNo contains a C prefix. The Dataset specified that these values represent canceled orders.
    - Investigate if there are any other unique InvoiceNo values. Which A prefix also exists.
        - Searched for InvoiceNo's that start with A. It looks like "Adjust bad debt?"
- Multiple "Stockcode" values were different from the 5-digit nominal code. Some include letters.
    - The Dataset notes did not specify any meaning for these unique StockCodes.
    - I manually searched through the miscellaneous stockcodes, but most appeared invalid. I decided not to include the potentially valid Stockcodes as well due to an insignificant amount.

**<u>Data Cleaning:</u>** (Detailed Notes in the Markdown) (Python)

1. Dropped Nulls from CustomerID
2. Dropped Invoice Numbers that did not have 6 digits.

3. Dropped StockCodes that aren't 5 digits or 5 digits followed by an object
4. Dropped Unit Prices that are less than 0

- 27% of the dataset was removed as a result of cleaning.

**<u>Aggregating Values:</u>** (SQL)

- CTEs used for Basket Size calculation and DateDiff for the Recency.
    - Basket Size formula:

Basket Size = Sum of Items Purchased Across all Invoices (CTE) /Number of Invoices

- Recency formula:

Recency = Overall Latest Invoice Date (CTE max) - Max Customer Date

- The rest of the Dataset is self-explanatory and a general overview is in the Readme.

**<u>Feature Engineering:</u>** (Detailed Notes in the Markdown) (Python)

1. Clustering

- Identify outliers in the RFM columns before Kmean Clustering.
- Isolate outliers into a different Dataset.
    - There is a singular outlier on the high end. Creating a new Dataset for a singular customer isn't necessary. But is good practice regardless.
- Scale/Normalize Data.
- Determine the number of Clusters.
- Remove noise from Cluster.
- Run the Kmean Cluster function.
- Determine Segment/Cluster characteristics based on Kmean Centers.

2. Feature Creation:

- Combine new Clusters into the Dataset.
- Churn Rate per Segment.
- Basket Size per Segment.

- Churn Rate Over Time.
- Basket Size Over Time.