Dominic B. Lee 213367 Ad Hoc Analysis

ITMGT 25.03-G Long Test 1

Before approaching the problems, I first checked the provided transaction data file. Seeing that it is in a JSON file, I had to convert and transform it into a DataFrame for easier access.

There are five problems:

1. Quantity of Items Sold per Month

First, I created a copy of the DataFrame that only contains transaction_items, transaction_date, and transaction_value. Then, I added new columns for year, month, and day by splitting the transaction_date. Since all items are in a single transaction row, I also distributed each item into a separate row and added a new column for the corresponding quantity of each item. After this preprocessing, I used the *groupby* function to group the months and transaction_items together and compute the total quantity of each item. To display the values properly, I used the *pandas pivot table* function.

2. Total Item Sales per Month

Since the unit prices were not provided, I first had to explore the provided data and compute for the unit prices of each item. After getting the unit prices, I created a function *computeSales* which will be applied to the whole DataFrame from #1. This function gets the transaction_item of each row, looks for the corresponding unit price of that item, and multiplies the unit price by the quantity. This resulted in some columns with NaN, so I had to do additional preprocessing to remove the NaN values. Then, I displayed it using *pandas pivot table*.

3. Repeater Customers

To get the repeater customers, I had to compare this month's and last month's customers. First, I created separate DataFrames for each month. I got the number of repeater customers by merging the two DataFrames using an inner merge. This would get the intersection of the two months.

4. Inactive Customers

To get the inactive customers, I had to compare this month’s and previous months’ customers and count those that are not customers for the current month. First, I created separate DataFrames for each set of months: Jan, Jan-Feb, Jan-Mar, Jan-Apr, and Jan-May. Then, I dropped the duplicates. I got the number of inactive customers by merging the two DataFrames

using an outer left-only merge. This would get the customers present in previous months but not in the current month.

5. Engaged Customers

To get the engaged customers, I had to compare this month's and previous months' customers and count those that are still customers for the current month. For this one, I reused the DataFrames I created for #4: Jan, Jan-Feb, Jan-Mar, Jan-Apr, and Jan-May. I got the number of engaged customers by merging the two DataFrames using an inner merge. This would get the customers present in past months up to the current month.

Lastly, to display the repeater, inactive and engaged customer counts in one table, I utilized the DataFrame.