After reading and analyzing the JSON file on JupyterNotebook, I prepared the dataset by transforming the transaction date to DateTime so that it would be easier to account for the transactions in the subsequent procedures. Then, to clean the data, I split the content of the values in column 'transaction_items' using the split function and transformed the rows into a granularity of one "line item" per row using the explode function. To also ensure that the values under 'transaction_value' and 'item_count' were numbers and not strings, I used the .to_numeric function to transform these. Lastly, since the data were already cleaned, I added a column 'is_single' to determine whether each row was from a transaction that had multiple items bought or not. This column was used to extract the unit price of each item only from the rows with single items to make the procedure easier.

After getting the unit price, I made a dictionary containing these so that it would be easier to access it in the following few procedures and added a new column with the unit prices so that the updated and final data frame after cleaning is complete. Subsequently, I used the groupby and .agg functions to get the quantity sold of each product per month. I then added a new column, 'UnitPrice,' to multiply it by the quantity sold to get the Total Amount of Sales.

Next, in getting the number of Repeaters, Inactive, and Engaged Customers, I first created a table for transaction history by using groupby and count to analyze the data and know what procedures to do next. Then, I thought of making it a matrix with only two values, 0 and 1, where 0 indicates a null value and 1 does not. Thus, it would be boolean values that would be very useful in determining whether a customer is a Repeater, Inactive, or Engaged. In getting the repeater, I multiplied 2 consecutive columns by each other, and if the product is more than 1, then it's a repeater; otherwise, if it's 0, then it's not and won't be included in the data. The same boolean values were also utilized in getting the Inactive and Engaged with different codes as there are different specifications in getting the number of Inactive and Engaged customers. The last items I put in the Notebook were the graphs, which were visuals to demonstrate further the data gathered.