First, I imported the json data into the Jupyter Notebook. Then I used the pandas slicer to isolate my needed columns into a new dataframe called items\_for\_cleanup. Next, I realized that I needed to separate the data by month so I could make a dataframe by month, so I separated it by looking at the date which had ‘2022/01’, ‘2022/02’ and so on. So I made an empty list, then made a for loop that iterates through the transaction items column of the dataframe in order to clean up the data by looking for this character ‘,(‘ since that precedes the number of items of each type the person bought. Finally, I nested a for loop inside the for loop in order to find the number inside the character containing the number of items bought by the person, which I converted into an integer, then I appended a dictionary containing item and count as keys with the item names and count as values into the empty list. The resulting list of dictionaries I made into a dataframe. I used the groupby and transform pandas functions to get the total number of items bought by customers, then I dropped duplicates of the item names so I can get the final dataframe. In order to get total values, I coded a scratch for loop which iterated through the transaction items column and looked for items that only had one item name inside with a brand name. I used simple math to get the prices, which I then used to code a function which I then vectorized to my final column with the count of total items bought so I can get total sales. I added that value to my final dataframe then I just repeated this for all the other months.

Next for the customer metrics I used the email and transaction date for my analysis, then made different dataframes for each month. So in making the repeaters per month, I got the length of the intersection of the current month and previous month using the .merge() function. Next, I got the number of engaged customers by getting the length of the set intersection of the current month, then all the months previous using the merge function. Finally, I got the inactive by getting the set difference of the union of the sets of customers in the preceding months and the set of customers in the current month. I did that by using .concat on the previous months and the data frame of the current month then dropping the duplicates, then getting the length of what I got to see inactive customers. I put all the values I got into lists per month so I can make a dataframe based on those lists along with a list of months to get the desired output

To end, I made simple bar graphs using matplotlib using my final dataframes of the sales and items sold per month, and I decided to put them separately because the differences between the values in the items sold and sales are just too large. Then, I summarized the customer metrics in a multicolumn graph so the comparisons between months are easily seen, and I found the lists useful since I was able to use them to make my multi column graph.