**Explanation**

**Exploring the dataset:**

Conducted basic analysis like finding the null values, improper values, count of columns, unique values etc.

Findings:

* The columns UserId, NumberOfItemsPurchased, ItemCode, CostPerItem had negative values.
* The column CostPerItem has ‘0’ values.
* The column ItemDescription has NaN values (df.isnull().values.any())
* 2,70,160 user ids had -1 value. (3710 unique)
* It was found that the user ids of a few transaction ids has -1 value.

**Data Cleaning:**

* All negative values in UserId were removed and replaced with by assigning values starting from 11 to the UserIds of those TransactionId where value was -1. Similarly, all negative values in ItemCode were removed and replaced with by assigning values starting from 10000000 to the ItemCode of those ItemDescription where value was -1.
* There were only 4 rows having negative values in CostPerItem and hence those were removed.
* The rows with 0 in CostPerItem were removed.
* The rows having null values in ItemDescription were removed.

**Output: Identification of Segments and corresponding UserId**

* A bar plot was plotted to find out the top 10 countries according to the no. of customers. United Kingdom was leading with a very high majority of 985958 out of 1078784.
* DataFrame uk\_data was made which contained values of the country: United Kingdom.
* Since there were negative values in NumberOfItemsPurchased, those data were considered in analysis where it was greater than 0.
* RFM Analysis was done to identify the different segments according to customer transactions. (Recency: number of days between present date and date of last purchase each customer, Frequency: number of orders for each customer, Monetary: sum of purchase price for each customer)
* This analysis is used to identify the potential customers or the most preferred customers. (Based on how recently, how often, and how much did a customer buy)
* The Recency, Frequency, Monetary values for each customer is calculated, segment bin values are added to RFM table using quartile and finally all scores are combined in a single column(RFM\_Score).
* 60 segments were identified and it can be filtered to find the top or best customers (Filtering it by descending order of monetary of any segment).