Q2

In my data cleaning process, I mainly followed these steps:

1. Reading the Data

I began by reading the dataset named ecommerce_dataset.csv using pandas.

2. Age (age) Noise Detection

To detect and handle outliers in the age field, I used two methods.

Method 1: Setting a Reasonable Range

I defined a reasonable age range, assuming that the age should be between 1 and 150 years. If any value fell outside this range, I considered it an outlier.

Method 2: IQR Method

In addition to the reasonable range, I also used the IQR (Interquartile Range) method to detect outliers. The process involved calculating the 25th and 75th percentiles of the data, and determining if any values exceeded the upper bound (i.e., Q3 + 1.5 * IQR). I also set a maximum age of 150, so any value above this threshold was treated as an outlier.

Finally, I merged the results from both methods to identify all the age outliers.

3. Purchase Amount (purchase_amount) Noise Detection

Next, I performed outlier detection on the purchase amount field using two methods:

Method 1: Setting a Reasonable Range

I set a maximum purchase amount of 999,999, and considered any values exceeding this threshold as outliers.

Method 2: Z-Score Method

I also calculated the Z-Score for each purchase amount. If the Z-Score of a value was greater than 3, indicating it was far from the mean, I considered it an outlier.

As with the age field, I combined the results from both methods to identify all purchase amount outliers.

4. Merging All Outlier Data

For convenience, I merged all the outlier data (both age and purchase amount outliers) into a single dataset and removed any duplicate records.

5. Cleaning the Data

I then removed all the identified outliers from the original dataset, leaving the cleaned data.

6. Age Grouping

To perform a more detailed analysis of age, I categorized the age field into five groups: Teen, Young Adult, Middle Aged, Older Adult, and Senior. Each age value was assigned to one of these groups.