Problem Statement

We have an ecommerce dataset with records of millions of customer interactions with online stores over a 7 month period from October 2019 to April 2020. The problem at hand is that we need to find a way to extract insights from the dataset such that we can better understand each customer's behavior as well as know what product recommendations we can provide them with.

Context

The dataset is composed of millions of records, each record containing the following nine columns:

event_type: (datetime) When the event happened.
event_type: (str) What occurred during the event.
product_id: (int) ID of the product involved.
category_id: (int) ID of the product category.
category_code: (str) Meaningful category name.
Brand: (str) Product brand name.
price: (float) Price of the product
user_id: (int) Permanent user ID.
User_session: (str) The user session ID of the event.

Most columns can take on a variable value within a certain format, such as datetime, float, and int while event_type takes on one value from four possible values (view, cart, remove_from_cart, purchase). Only two columns, category_code and brand, are partly composed of null values, while the other seven columns are all composed of valid values.

Proposal

1. Build customer profiles by grouping rows based on user_id. This can be done by sorting the values in the dataframe column by user_id value.
2. Refine the grouping within each user by event_type to see which sessions ended in views, carts, cart removals, and purchases.
3. Compare customer activity with that of customer's similar to them in activity, comparing for columns with similar even_type and category_code, for the purposes of recommending other products each user may have purchased.

Data Source

eCommerce behavior data from multi category store
https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store?select=2019-Oct.csv

Choice of technology

Python and associated python libraries such as pandas and numpy are the intended technology. The rationale behind this choice is because the dataset comes in a CSV format, making the pandas DataFrame the optimal data structure to process the data set.

Feedback:

- Do we have data related to customer activity and behaviour? If so, which attributes and what is our approach to understand / use those attributes?
- Dataset -> Attributes -> means so and so -> transform it (this could be any algos or any transformation technique) -> derive / understand customer profile or behaviour