Data Science Course: Capstone Project 1
eCommerce behavior in events

Capstone Mini-Project: Data Wrangling

Dataset: Collected two months (Oct & Nov 2019) e-commerce events dataset from a large multi-category online store.

The dataset was downloaded from Kaggle https://www.kaggle.com/mkechinov/ecommerce-events-history-in-cosmetics-shop

Data Wrangling

Data is imported and cleaned individually for each month - Oct & Nov since the dataset is huge - approx. 4 million records each month.
Following steps are taken for data cleaning & wrangling for each month's data:

Step1: Using right data types for each column of data
After importing the data from .csv files, inspected the default data type that was assigned to each column of data.
There were 2 columns - event_type and category_code that had a small set (4 to 10) of distinct categorical values. Therefore, converted these two columns' data type from 'object' to 'category'.
There was a event_time column which contained the date/time of the event. Converted this column to date-time data type so it could be used for easier analysis.

Step 2: Deleted duplicate data
Checked the imported data for presence of duplicates and deleted all duplicate rows. From Oct data, 213,155 duplicate rows were deleted. From Nov data, 246,693 duplicate rows were eliminated.

Step3: Checked null values in each column
Used 'assert' function to check null values in each column of Oct and Nov dataset. Category_code, brand and user_session had null values.
User_session did not have many null values - only 574 rows in Oct dataset and 755 rows in Nov dataset. Since the total number of rows in each dataset was around 4 million, I **decided to drop all the rows with null user_session data as this would not affect the overall large dataset in analysis. Also, null user_session data seems to be erroneous.**

Category_code and brand has a huge number of null values. There are around 1.5 million rows with null brands in Oct.

Missing 'brand' was filled with value 'missing' category type as there are around 1.5 million rows with missing brand values.

Every product_id has a brand. So, created a dictionary between product_id and brand. Used this dictionary to fill missing brand values. A total of 2769 rows of data were filled. Remaining null brands are filled with a new category called ‘missing’.

Missing category_code was filled with value ‘other’ category type as there are around 3.8 million rows with missing category_code values.

Step4: Handling float data type column

‘Price’ is a float type column. **Examined the dataset for negative price values**. There were **only 20 rows with negative ‘price’ values so deleted all those rows**.

Step5: Frequency distribution of ‘price’ column
Created ‘histogram’ frequency distribution plot for ‘price’ column. It is observed that 90% **of the data lies in the price range of 0-8$**

Step6: Used one-hot encoding to create dummy columns based on 4 event_type(s)
Following 4 columns were added - event_type_purchase, event_type_cart, event_type_view, event_type_view.
These new columns help in getting insights into data as per each event type.