**Data Analysis Project   
eCommerce behavior data from multi category store  
By Orli Kagan**

**1. Introduction and Data set**

This SQL project is based on real behavioral data from the 01 of October 2019 and 01 of November 2019, from a large multi-category online store.

The main reason for this analysis is the importunacy of user behavior and the motives to try and understand their choices and preferences by different tested parameters, from the results of our analysis, a business can make recommendations for improvements and goal achievements, based on understanding of their customer’s needs.

Each row in the file represents an event- there are 2,097,150 events by user’s sessions in this data.

Each event is like many-to-many relation between products and users.

All events are related to products and users -after count by distinct user id there is 332,902 users.

Data collected by Open CDP project.  
The dataset that can be found and downloaded here:  
<https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store>

**2. The database Description**

1.The Data set is consisting of 2 CSV. Files for each 01 day of the month’s Sep and Oct, 2019.  
  
2. There are some NULLs in the database in the Category code and Brand columns.  
  
3.The database consists of 67,501,979 + 42,448,764 rows and of 9 columns.

These are the columns and their description:

event\_time, Time when event happened at (in UTC).

event\_type, only one kind of event: purchase.

product\_id, ID of a product

category\_id, Product's category ID

category\_code, Product's category taxonomy (code name) if it was possible to make it.   
 Usually present for meaningful categories and skipped for different kinds of accessories.

brand, string of brand name. Can be missed.

price, price of a product. Present.

user\_id, Permanent user ID.

\*\* user\_session\*\*, Temporary user's session ID. Same for each user's session. Is changed every time user come back to online store from a long pause.

**Different types of events and flow:**

User (user\_id ) during session (user\_session) added to shopping cart (property event\_type is equal cart) product (product\_id) of brand (brand) of category (category\_code ), category code with price (price) at specific time and date (event\_time)

**Event types**

view - a user viewed a product

cart - a user added a product to shopping cart

remove\_from\_cart - a user removed a product from shopping cart

purchase - a user purchased a product

**3. Data preparation**After loading the data to SSMS, I created a new table with same columns that combined the two separated tables into new one.   
I changed the values types of some of the columns to new type, to be able to do calculations between them (varchar to big\_int), and also the date to datetime in order for me to separate the exact hours and days from different months.  
  
Later I created a another new table, only for the products, which was comprised out of existing columns and new calculated columns that I created, added and arranged:  
count\_views, count\_addcart, count\_purchased, count\_removecart,  
max price, minimum price, price at sale point, P\_purchased (% purchased),P\_abanden (% didn’t purchased eventually)  
  
All of this is done in order to estimate the correlation between users behavior and motivations for purchasing specific products- is it effected by price? Or by Brand? What is the most selling category of good being sold, and what could be possibly the reason for the different between different time periods.

**4. Analysis description**Buying behaviour always fascinated me, I wanted to better understand the numbers behind the decision making of individuals track the actions ant translate them for better decision making, why people buy more or less of particular products.

To better answer those questions, I have pointed few notions that I wanted to see prove and test, by writing SQL commands on the giving data set. From the results I received, I can understand more about customer preferences and give my recommendations ,that enable better decision making later ,while pricing new products or re-ordering more of specific category or brand.

**Questions (SQL files with queries attached) :**

1.Joining table October/November (Oct\_Nov\_new table)

* Creating one table for OCT\_NOV together

2.Cleaning the data (cleaning\_brand &update)

* Creating temporary table (products) to Identify products that have more than one category ID, category code or brand name (Null is counted as its own segment)
* View our products and their category/brand info
* Check for product\_id with two or more actual brand values and we will ignore updating them
* Exploring the result of the specific product
* Exclude those that have more than one actual brand
* Get the brand value for those products we want to update, see results and updating and verifying.

3.Using join function to create analysis (Joining table analysis)

* Total oct rev by product IN Oct, Nov.
* Total sum by month,
* Calculation of increase over time (amount of change between 2 months)
* Top selling product by rev
* Top selling category and brand by month
* The percentage of category out of total month revenue (2 steps)
* Top brand by category in each month

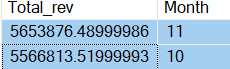
4. Creating product table (creating product table)

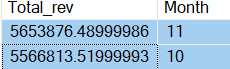
* Creating a table for the products and new columns calculations
* Insert data to new created table by using select statement from another table

5.Queries to understand the data (Investigation Product table)

* Finding the amount of purchased percentage
* Reviewing data anomaly/data integrity -proof that some users had more than 1 sessions
* People who purchased but we don’t know the brand only the category code/name
* Which category code(name) we had the most percentage of people abandoning the cart out of those who viewed
* Creating new calculated columns, and order by percentage of abending/purchased
* Verification of data NULLS between 2 columns: category\_code, category\_id
* Top 1 ranking bestselling by category Id
* Ranking the top 5 best product within a category using rank function
* Top bestselling products (also by month)
* Where views=0 , from purchased
* Proof that some users had more than 1 sessions

**Recommendations**

After seeing the queries results, I can understand that in both months, the top selling **brand** is Apple, Smart-phone devices. Therefore, we can understand that users in general prefer that company over Samsung.



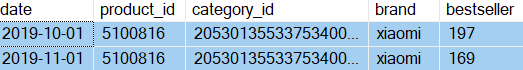
Furthermore, in October the revenue was higher (change of gap between was 1.56%), maybe due to lack of inventory of previous bestselling products and new stack of different brands on November that users aren’t familiar with.

After deeper drill to the data, we can see that this is probably due to Apple higher prices, that caused to higher revenue.

Nevertheless, the overall top number of people who made a purchase after first time viewed a Smart-phone were of Samsung brand.

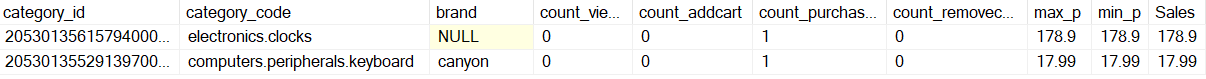
To support this and see the real user preferences we calculated by top 1 rank, the bestselling category in Oct and Nov, and sew that on category of “electronics.smartphone”: Samsung, is top first place.

Even that max count of people that purchased product, showed that “xiaomi” is the first bestselling product in both dates.



Those 3 brands have different strength’s and people prefer them over other due to different reasons , top design ,low price and more.

Later, we discovered an anomaly in the data, in each month there is one product that is purchased without previous view – this is not usual and can be only explained if user A view the product and passed the link to make a purchase directly to user B .



We checked on each month the category that had the biggest amount of people abandoning the cart after view, without making a purchase. Our indicator was from 100% abandoned rate to smallest- this gave us the understanding regarding which products from category we should order less next time.

Eventually we sew that some users had more than one session, this might point to a fact that, a user needs to take time, for making a decision to complete a purchase. In order for further investigation of how to lower this barrier and minimize the amount of time from the first view to first time purchase/action, more Data needs to be collected.

A big disadvantage to this analysis that this data set had only data of the first day of each month, it is not enough for making any firm conclusion, without ability to track changes or repeated patterns/connection across different days in each month.