Homework 2 - eCommerce analytics

E-commerce, also known as electronic commerce or internet commerce, refers to the buying and selling of goods or services using the internet, and the transfer of money and data to execute these transactions. The first e-commerce implementations date back to the 1990s and since then, millions of people every day visit some e-commerce sites to look for some product or service and, eventually, to purchase it.

You have been hired as a data scientist from a big multi-category online store. You and your team have been required to perform an analysis of the customer behavior in the store. Each row in the dataset represents an event, which catches different interactions (views, a product added/removed to/from the cart, purchases) of customers with your e-commerce. All events are related to products and users.

Your **goal** is to answer some research questions (RQs) that may help us discover and interpret meaningful patterns in data and eventually increase the number of sales.

Before starting

Among all numerous things and good practises a data scientist needs to do before running any analysis, there is one the is of uttermost importance: **get data and understand it**!

Here you find the list of tasks you need to perform before digging into the world of e-commerce.

- **Get your data!** Go to this website and download the files **2019-Oct** and **2019-Nov**.
- **Understand your data.** Read the legend of each column to understand what it refers to. Additional information about the labels can be found in the description of the data section on the web page. Please, be sure that you've understood the data before start coding.
- **Handling data.** The data are provided in two `.csv` files, with the same columns present in both files. For this reason, in order to answer the RQs, we kindly suggest you to import the `.csv` files as `pandas DataFrame` object and then, based on what you want to analyze, perform the necessary operations.

Remember, **Google is your best friend!**

VERY VERY IMPORTANT

1. **!!! Read the entire homework before coding anything!!!**
2. *My solution it's not better than yours and yours is not better than mine.* In any data analysis task, there **is not** a unique way to answer to RQs. For this reason it is crucial (**necessary and mandatory**) that you describe any single decision you take and all the steps you do.
3. Once performed any exercise, comments about the obtained results are **mandatory**. We are not always explicit where to focus your comments, but we will always want some brief sentences about your discoveries.

Research questions

Exploratory Data Analysis

1. [**RQ1**] A marketing funnel describes your *customer's journey* with your e-commerce. It may involve different stages, beginning when someone learns about your business, when he/she visits your website for the first time, to the purchasing stage, marketing funnels map routes to conversion and beyond. Suppose your funnel involves just three simple steps: 1) view, 2) cart, 3) purchase. Which is the rate of complete funnels?
    - What's the operation users repeat more on average within a session? Produce a plot that shows the average number of times users perform each operation (view/removefromchart etc etc).
    - How many times, on average, a user views a product before adding it to the cart?
    - What's the probability that products added once to the cart are effectively bought?
    - What's the average time an item stays in the cart before being removed?
    - How much time passes on average between the first view time and a purchase/addition to cart?
2. [**RQ2**] *What are the categories of the most trending products overall?* For each month visualize this information through a plot showing the number of sold products per category.
    - Plot the most visited subcategories.
    - What are the 10 most sold products per category?
3. [**RQ3**] For each category, what's the brand whose prices are higher on average?
    - Write a function that asks the user a category in input and returns a plot indicating the average price of the products sold by the brand.
    - Find, for each category, the brand with the highest average price. Return all the results in ascending order by price.
4. [**RQ4**] How much does each brand earn per month? Write a function that given the name of a brand in input returns, for each month, its profit. Is the average price of products of different brands significantly different?
    - Using the function you just created, find the top 3 brands that have suffered the biggest losses in earnings between one month and the next, specifing bothe the loss percentage and the 2 months (e.g., brand_1 lost 20% between march and april).
5. [**RQ5**] *In what part of the day is your store most visited?*Knowing which days of the week or even which hours of the day shoppers are likely to visit your online store and make a purchase may help you improve your strategies. Create a plot that for each day of the week show the hourly average of visitors your store has.
6. [**RQ6**] The conversion rate of a product is given by the purchase rate over the number of times the product has been visited. What's the conversion rate of your online store?
    - Find the overall conversion rate of your store.
    - Plot the purchase rate of each category and show the conversion rate of each category in decreasing order.
7. [**RQ7**] The Pareto principle states that for many outcomes roughly 80% of consequences come from 20% of the causes. Also known as 80/20 rule, in e-commerce simply means that most of your business, around 80%, likely comes from about 20% of your customers.
    - Prove that the pareto principle applies to your store.

Bonus points

For this homework, you are required to work with all data in the 2019-October and 2019-November range. An extension of the dataset is available at this link. It is not necessary to use the extension for this homework, however, if you decide to use it, we will take it into account in the final evaluation.