Module 2: EDA

Remember to download the data from s3 using python `boto3` or `AWS CLI` and save them locally to avoid having to download them over and over again every time you want to use them.

1. Understanding the problem space

We have provided you with sample tabular data in parquet format that you can find in `s3://zrive-ds-data/groceries/sampled-datasets/`.

The data is partitioned over multiple dataset and comes from a groceries e-commerce platform selling products directly to consumers (think of it as an online supermarket):

- orders.parquet: An order history of customers. Each row is an order and the item_ids for the order are stored as a list in the item_ids column
- regulars.parquet: Users are allowed to specify items that they wish to buy regularly. This data gives the items each user has asked to get regularly, along with when they input that information.
- abandoned_cart.parquet: If a user has added items to their basket but not bought them, we capture that information. Items that were abandoned are stored as a list in item_ids.
- inventory.parquet: Some information about each item_id
- users.parquet: Information about users.

Work to do

Pull the different datasets, manipulate them and join them as necessary to carry out an initial data analysis to understand the problem space we will be working on..

1. Do all the quick checks and find any potential issues with the data. Fix them as you see fit.
2. Understand the problem at hand through data interrogation and hypothesis testing.
3. Gather all the hypothesis and insights you have tested and found, in a document or clean jupyter notebook. What do we know about the real situation represented by the data?

REMEMBER: The most important part of an initial EDA is to think what you want to know from the data and systematically answer those questions.

2. Exploratory Data Analysis

We have also provided a ready to use dataset that have been prepared for you. You can find it `s3://zrive-ds-data/groceries/box_builder_dataset/sampled_box_builder_df.csv`.

In this dataset, every row representes an (order, product) pair where outcome indicates whether the product was bought or not on that specific order and every other feature has been computed only looking at information prior to that order in order to avoid information leakage.

Work to do

Carry out an exploratory data analysis of the dataset.

1. Do all the quick checks and find any potential issues with the data. Fix them as you see fit.
2. Do all he data integrity checks and understand the shape of your data.
3. Document the analysis in a document or clean jupyter notebook.