**Shopee Python-Pandas Test (45 mins)**

In this task, you'll be analysing listings data from our Shopee Platform.

You may use the Pycharm IDE installed, Sublime or other windows native text editors. Please save your python source code on the desktop. You may use the internet for help.

The dataset is stored in the Test\_Pandas.xlsx file. It contains listing information posted on Shopee. One single listing corresponds to one row in the dataset.

The dataset has 12 columns, and 464433 rows.

**Here are the brief descriptions of each column:**

Itemid - a unique ID of the product

Shopid - a unique ID of the shop

item\_name - product title

item\_description - detailed product description

item\_variation - stores variations of a product (e.g. different colours or sizes, in the format like {variation 1 name: variation 1 price, variation 2 name: variation 2 price})

price - how much does the item sold

stock - how many stocks left

category - which category does the product belongs to

cb\_option - 1 indicates the product is sold by a cross border shop

is\_preferred - 1 indicates the product is sold by a preferred shop

sold\_count - how many products have been sold

item\_creation\_date - when are the product uploaded by the seller

1. Use pandas function to read the Test\_Pandas.xlsx file in:
   1. Assign the result to a variable named “data”
   2. Assign all column names to a variable named “columns”
2. Use pandas function to find:
   1. How many unique shops are in the dataset?
   2. How many unique **preferred** and **cross border** shops are in the dataset?
   3. How many products have zero sold count?
   4. How many products were created in the year 2018?
3. Use pandas function to find:
   1. Top 3 **Preferred shops’ shopid** that have the largest number of **unique** products
   2. Top 3 **Categories** that have the largest number of **unique** **cross-border** products
4. Find Top 3 **shopid** with the highest revenue (Assumption: the product price has not been changed.)
5. Find number of products that have more than 3 variations (do not include products with 3 or fewer variations)
6. Use pandas function to identify duplicated listings within each shop (If listing A and B in shop S have the **exactly same** **product title, product detailed description**, and **price**, both listing A and B are considered as duplicated listings)
   1. Mark those duplicated listings with True otherwise False and store the marking result in a new column named “is\_duplicated”
   2. Find duplicate listings that has less than 2 sold count and store the result in a new excel file named “duplicated\_listings.xlsx”
   3. Find the preferred shop shopid that have the most number of duplicated listings