eyos

a)   Business Context

eyos collects granular transaction receipt data from a panel of independent grocery retailers. The collected data is parsed from printed paper receipts and is subjected to both data collection errors and store – level nuances.

In this assessment, applicants are invited to suggest approaches to support scaling of data analytics models, and develop analytical models to support property enrichment of data entities.

b)   Evaluation

There are two different questions below. Applicants are expected to complete both questions in code (Python/ R) and submitted in presentation format (i.e. HTML, PDF etc.) Applicants should summarise the steps that they undertook and describe their considerations taken.

All of these should be submitted by the deadline given.

c)   Dataset

Transactions from a set of 117 stores are recorded in a 1 month window. Accompanying product attributes are included in a secondary table.

fct_sales

receipt_id	product_id	store_id	sales_datetime	sales_turnover	sales_pieces

dim_products

product_id	product_category	product_pack	product_type

The csv files have been compressed. Use the following python command to read the files.
pd.read_csv(' *file name* ',compression='gzip',sep=';')

d) **Assessment Questions**

Section 1: Load Balancing

Internal models typically process each store independently of other stores, and computational resource requirements for each store increase based on the number of transactions recorded. We aim to parallelize these models by processing multiple stores concurrently i.e. multithreading.

Given a user defined parameter *k,* write an algorithm to partition the dataset into *k* chunks such that:

- Transactions for one store are in the same chunk
- Each chunk has approximately the same total number of transactions

Section 2: Shopper Analytics – Store Segmentation & Association Rules

1. Clients would like to understand the types of stores that are present in the panel. Arrange stores into mutually exclusive groups and describe the groups. Code libraries are allowed for this question.

2. These store owners are interested to identify cross-selling opportunities in their stores via 1 to 1 product associations. However, due to limited resources, we are only able to select two stores.
   a. Select two stores from the dataset in which you have identified as ready for a pilot study and provide your rationales.
   b. Identify opportunities via 1-1 product associations (if_bought_this_sales_item_id -> likely_to_buy_this_sales_item_id), evaluate and rank the opportunities based on relevant metrics. Code libraries used should only be limited to linear algebra and dataframe manipulation.
   c. Compare the differences in the opportunities between both stores and share your findings.