Background

Pernalonga, a leading supermarket chain of over 400 stores in Lunitunia, sells over 10 thousand products in over 400 categories.  Pernalonga regularly partners with suppliers to fund promotions and derives about 30% of its sales on promotions.  While a majority of its promotion activities are in-store promotions, it recently started partnering with select suppliers to experiment on personalized promotions.  In theory, personalized promotions are more efficient as offers are only made to targeted individuals who required an offer to purchase a product.  In contrast, most in-store promotions make temporary price reductions on a product available to all customers whether or not a customer needs the incentive to purchase the product.  The efficiency of personalized promotion comes from an additional analysis required on customer transaction data to determine which customers are most likely to purchase a product to be offered in order to maximize the opportunity for incremental sales and profits.

Problem

Your analytics consulting firm is being considered by Pernalonga (the client) to develop a marketing campaign to experiment on personalized promotions.  While the details of specific partnerships with suppliers to fund the experimental personalized promotions are still being negotiated, you have started to receive data from the client.  You have two weeks to analyze and understand the data and report back initial insights to the client.  In order to be selected as the sole-developer of the marketing campaign, your team needs to demonstrate that you know the data very well, i.e., you need to show the client that you know the profiles of their stores, products and customers better than they do and are ready to take on the task of developing the marketing campaign.

From the client’s point of view, they need to be confident that you know the answers to the following key questions:

* Who are the best customers in terms of revenues, profits, transactions/store visits, number of products, etc.?
* What are the products and product groups with the best volumes, revenues, profits, transactions, customers, etc.?
* Which stores rank the highest in volumes, revenues, profits, transactions, customers, etc.?
* Are there interesting groupings of customers, e.g., most valuable (buy everything at any price) or cherry-pickers (buy mostly on promotions), defined by certain categories (buy baby products or never buy milk), etc.?
* Other than product categories and sub-categories, are there other product groupings, e.g., Key Value Items (KVI) and Key Value Categories (KVC), traffic drivers, always promoted versus seldom/never promoted, etc.?
* Are there natural groupings of stores, e.g., stores frequented by cherry-pickers versus stores visited by most loyal customers?

Available Data

The file [Pernalonga.zip](https://canvas.emory.edu/courses/69929/files/3281192/download) contains two tables:

1. transaction\_table.csv contains transaction history in 2016 and 2017 for close to 8,000 customers

* cust\_id – Customer ID
* tran\_id – Transaction ID
* tran\_dt – Transaction Date
* store\_id – Store ID
* prod\_id – Product ID
* prod\_unit – Product unit of measure: CT for count and KG for kilograms
* prod\_unit\_price – Unit price of the product
* tran\_prod\_sale\_qty – Quantity/units of the product in the transaction
* tran\_prod\_sale\_amt – Sales amount for the product before discounts in the transaction
* tran\_prod\_discount\_amt – Total amount of discounts applied to the product in the transaction
* tran\_prod\_offer\_cts – Total number of offers on the product resulting in the total amount of discounts in the transaction
* tran\_prod\_paid\_amt – Amount paid for the product after discounts are applied in the transaction

1. product\_table.csv contains the product to subcategory and category mapping and descriptions for about 11,000 products
   * prod\_id – Product ID
   * subcategory\_id – Subcategory ID
   * category\_id – Category ID
   * sub\_category\_desc – Subcategory name (in Portuguese)
   * category\_desc – Category name (in Portuguese)
   * category\_desc\_eng – Category name (in English)
   * brand\_desc – Brand of the product, including NO LABEL and PRIVATE LABEL

Note that customer, store and product information beyond what is available above are not provided.

Grading

Professional data scientists are expected to be domain experts implementing sound mathematical models using robust and reusable computer programs.  Your work will be graded according to the following criteria:

* Integration of domain knowledge/practicality into solution (20%)
* Creativity and mathematically sound application/execution of chosen technique/model (25%)
* Robustness and efficiency of solution/code (25%)
* Report and presentation flow, delivery, and defense (20%)
* Peer evaluation (10%)

Reports

Project written reports  and computer codes are due on February 7, 2020.  Three groups will be selected to present their reports on February 18, 2020.