Big Data Management Assignment for AMPBA

A large retail store, which sells daily groceries wants to analyze customer traits and association patterns among various products it is selling.

Dataset

The dataset contains all orders checked out by customers for about 3 month in 2015.

SKU – Stock Keeping Unit. A unique product code.
OrderNum – Order number (i.e. transaction id)
CustomerID – Unique Customer id
Sale Price – Sold Price (what the customer has paid for the item bought in the order).
SKU Description – Description of the product.
Created On – The date on which the customer has created the order.
Category – The category of the product (SKU). There are about 11 categories.

RFM Analysis

RFM stands for Recency, Frequency, and Monetary value, each referring to key customer trait. These RFM metrics are important indicators of a customer's behavior because frequency and monetary value affects a customer's lifetime value, and recency affects retention, a measure of engagement.

a) Calculate the RFM values for each customer (by customer id). The result set should have customer id, recency, frequency and monetary value columns.

RFM can be calculated as follows:

R (Recency) – Time since the last order is made by the customer. For recency calculation, use 05/08/2015 as current date. So, recency will be the number of days before the date of 05/08/2015, the customer has made his or her last order.

F (Frequency) – Total number of orders made by the customer.

M (Monetary Value) – Total spend (sale price) by the customer including all his/her orders.

For detailed description of the terms, refer the link below

https://en.wikipedia.org/wiki/RFM_(customer_value)

b) Find top 10 customers based on monetary value.
c) Find top 10 customers based on average spend by order i.e. (Monetary Value/Frequency).

d) Draw the following three scatter plots and comment on your observations from the date.
   - Monetary Value Vs. Frequency
   - Recency Vs. Frequency
   - Frequency Vs. Recency

Association Rules Analysis:

The marketing department wants to build recommendation engine for cross selling of items across categories. The recommendations should be made within each specific category only.

a) Find out top 10 association rules from any 3 different categories. There are SKUs from 11 different categories. You have the freedom to choose any three the categories.

b) The rules should be sorted based on confidence (highest to lowest) and the minimum value of lift should be 1. Please choose minimum support as applicable.

Assessment Weightage:

Evaluation will not only be based on completing the above tasks, but also the following factors.

1. Code clarity
2. Documentation (use markdown)

There may be penalty if code clarity and documentation is not proper.

Notebook Requirements:

The notebook should contain:

1. The participant name.
2. Clear description of the problem and dataset.
3. Proper sections should be created for clarity.
4. Insights and inferences should be described for understanding.
5. There should be conclusion section with summary of accomplishment and few bulleted points of lessons learnt in developing the project.

The Coding scheme for Individual Assignment is 2N-c

Deliverables:

The deliverables should be the notebook with all outputs from the complete run of the notebook and the datasets.

Do NOT submit .zip files otherwise, the submission will not be considered.