MS5106: Data Science and Big Data Analytics

Part B | Lecturer: Dr. A. Griva

Assignment scope

The scope of this assignment is to analyse retail datasets and extract useful insights and business value to support decision making. The dataset includes point-of-sales data (POS) as generated by the cashiers in a grocery store. Following you may find the dataset description and the key questions you need to answer.

Dataset description

The **first excel sheet** named "POS DATA" includes the POS data. Each basket_id refers to a cashier receipt. The date field refers to the date the transaction was performed (in data type int). The barcode refers to the unique item number, the sum_units to the number of items purchased from the given barcode, and the sum_value refers to the total value of these items. For example, in Table 1, basket_id=1103084867 contains items from two different barcodes (i.e. 800220505783 and 520139501183). In more detail, it contains 2 items from the first barcode and 1 from the latter. In addition both items of the first barcode cost 1.96€. Card_id indicates the unique number of the customer's club/loyalty card. If this value = NULL, then the customer did not use their card in this transaction.

TIP 1: be careful with the negative values in "Sum_Units" and "Sum_Value" columns.
TIP 2: be careful with the decimal values in "Sum_Units" columns. What they declare?

Basket_ID	Date	Barcode	Sum_Units	Sum_Value	Card_ID
1103084867	41379	800220505783	2	1.96	9160003751260
1103084867	41379	520139501183	1	5.349993	9160003751260
1092750793	41346	520423907421	6	1.740015	9164012915385
1106160983	41388	211069400000	1	0.749817	9162005811409
1108695491	41395	520286400380	-2	-0.6	NULL

Table 1. POS data

The **second excel sheet** named "LOYALTY" contains data regarding the cardholders e.g. age, gender, marital status, household size, number of children. If the value = NULL then the customer didn't want to fill this information.

TIP 3: (Information Quality) cardholders might have declare false details e.g. age more than 120 years old etc.

The **third excel** sheet named “barcodes” includes the ids of all the barcodes that shoppers have purchased. Also, it includes retailer’s product taxonomy/hierarchy tree (only the ids).

TIP 4: each barcode should have 12 digits.

TIP 5: you can identify the descriptions of these categories by joining the 3rd and 4th excel sheet. Be careful on what on what column (or meta-column) you will use for the join operation.

The **fourth excel sheet** named “product taxonomy” includes the descriptions of retailer’s product hierarchy/ taxonomy. This hierarchy includes 4 product levels. See figure 2 to understand what is a product hierarchy for an example:

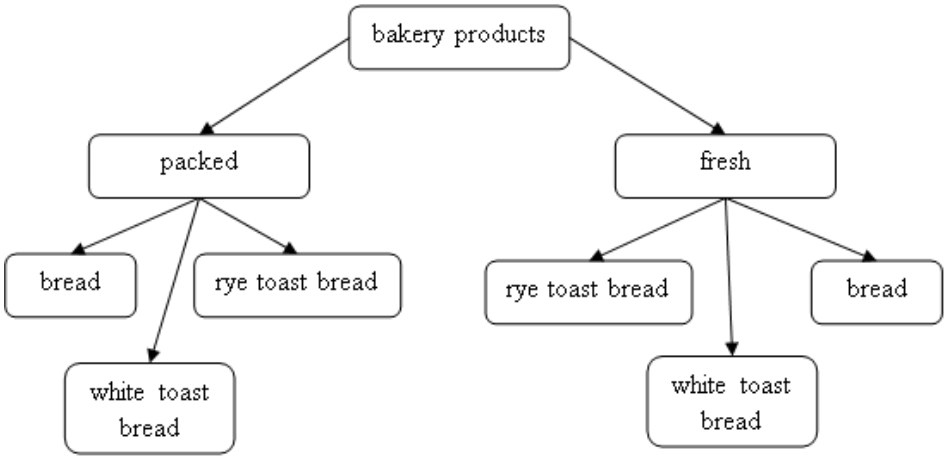

Figure 2. Product taxonomy/hierarchy

TIP 6: retailer’s product taxonomy serves operational (e.g. product replenishment) and not analysis purposes. You can (but it is not compulsory) create new product levels for this assignment.

Queries

1. Basket segmentation:
   What groups/segments of baskets that describe shopper buying behaviour can you identify?
   TIP: First you should choose the metrics and the dimensions you will use to segment the baskets e.g. product categories, basket value, both, other?

2. Open questions:
   a. Formulate **two (2)** interesting questions **that you could answer** based on the given dataset and **answer them** based on the given dataset.

   TIP: Basket segmentation (results from query 1) can be used as input in this analysis (this is not compulsory, it is just a suggestion)

   TIP: Be careful when you use dimensions that include poor quality data e.g. age, as cardholders declare wrong information about themselves.

Focus on the presentation, the visualization and the interpretation of the results.

Per query (i.e. both query 1 and query 2) you should indicate the value of your analysis e.g. how a (marketing) manager could use the extracted insights to support decision making etc. Try to propose indicative examples after presenting the insights of each query.

Team structure

4-5 members. Teams having 5 members should work more on the open questions.

Deliverable and deadline

You are asked to create a **10-minute presentation (max).** You should prepare the slides and record **a slide show with narration** (or a video). See tips below for slide show with narration:

https://support.office.com/en-gb/article/record-a-slide-show-with-narration-and-slide-timings-0b9502c6-5f6c-40ae-b1e7-e47d8741161c

Assume that you present the results to business people having a little to no knowledge on analytics. Your presentation will aid them to decide the firm's strategy and marketing plans. Do not include technicalities in the main presentation.

In this presentation include **max 5 slides** in an Appendix:

- including your data preparation tasks (e.g. cleansing operations, possible outliers, data issues etc.), and
- explaining the technical aspects of your analysis e.g. techniques, algorithms, tools used etc.

Do not include any narration in these slides.

Do not include any code in these slides.

Assessment: 20%

The first slide of the presentation should include: (i) team name, (ii) team number, (iii) student name, (iv) student id.

One person for each team is responsible to submit the assignment.

All team members should submit a "peer evaluation form".

An "Individual contribution form" is NOT required for this module.

Tools to use

You are free to select any tool/ software you want to prepare and analyse the given data.

An interesting article:

Harvard Business Review (HBR)

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/