# **INFS 5096 Customer Analytics in Large Organisations**

# **Assignment 1 – Supermarket analysis**

This assignment is worth 30 points of your final grade. Your task is to prepare the data, run an analysis, answer research questions below and write a report with your findings.

# Dataset

You are provided with transaction data from a supermarket. There are 3 years of data. Every trading day is represented by a separate file. Some days are missing as they were public holidays. Names for all variables are self-explanatory. However, if you need any clarifications, please feel free to ask on the forum.

These data are a dump from the supermarket database, and it might have some “imperfections”. It is your responsibility to properly prepared data for the analysis.

For example, there is a customer ID (from a loyalty card) attached to most transaction. However, not every customer shopping in the supermarket has a loyalty card or uses it at the register. If customer has no card, then a staff member on the register will use one of “generic” cards. As a result, it looks as there are several customers buying too many products – much more than you might expect for the “normal” customer. This is not true. These “super” customers were customers without loyalty cards and staff members on the register used default cards. Use your common sense and prepare the data accordingly. Beware, this is not the only problem with data.

# Research questions:

1. Make an introduction to the data and research to follow. Remember: your readers haven’t seen the data; you need to explain what your data is. You must “set the stage” before you start with the analysis.
2. Working with 2014 data, aggregate data by user ID and keep only users that visited the supermarket at least once in 2015 and 2013. The idea is to analyse “regular” customers in 2014 and exclude newcomers and drop-offs. Calculate the number of trips, number of purchases, total money spent, and proportion of purchases made from different departments (variable *Department\_Name*) over the year. Run cluster analysis to identify if there are any patters in the data and if it is reasonable to perform a market segmentation based on what, how much and how frequently customers buy.  
   There are no restrictions or requirements on what software and/or clustering methods to use. It is expected that you will try different methods and report your best clustering solution.
3. Use all three years of data to analyse monthly sales in the supermarket and make a prediction for total sales in January, February, and March of 2016. You need to provide expected sales and error margin for your expectation. You are free to use any techniques – time series analysis, regression analysis, neural networks.
4. It is believed in the industry that sales promotion improves actual sales and results in higher number of items sold, which leads to a higher profit even if sales are at somewhat lower prices. We cannot calculate actual profits as we don’t know wholesale prices. However, we can estimate a relationship between price change (in %) and number of items change (in %) for different products.  
   Use data for 2013 only, but full data including “loyal” and normal customers. At that time inflation was relatively low and prices were stable. So, we can assume that any price change is a marketing activity like sales promotion.

Temporary price increase might be a marketing activity too, for example manufacturer wants to pump price to make a discount promotion looks larger. **Hint 1**: start for one particular product/SKU, e.g. canned tuna, and do analysis for a single product only. Then repeat the same analysis for as many products as possible to be able to make generalisations. It is NOT expected that you will do this analysis for all products, some products might be “unpopular” and naturally have very low sales – their results would not be representative. **Hint 2**: There is a variable *Offer* marking a promotion activity. However, “promotion” is not always a price discount. It can be a better placing in the store, or a two-for-one deal, or an advertising in a catalogue. There can be “strange” relationships between price change and promotion. You can ignore the flag “promotion”, or you can include a variable *Offer* as a covariate. Either way, your main focus is on the actual prices.

Hints for working with data:

* Do preliminary testing on smaller datasets – one day, one week or one month only. When you are confident that everything works fine – run the same code on the full dataset.
* Before doing each task, think what variables you need for this task and keep only these variables. Dataset has too many variables, you don’t need all of them.
* If you own a powerful computer with a lot of RAM, then you can ignore previous hints.

# Submission

You must submit a formal report with your research findings in MS Word or PDF format. Your report will include:

1. Introduction in a business case.
2. Dataset description – number of trading days, customers, shopping trips, items, and dollar volumes. You should set the stage for the research to follow.
3. Discussions about each research question supported by required numerical outputs, tables and data visualisations.
4. Conclusion.
5. Appendix with some extra information, if required, and/or codes used.

You don’t need to submit programming code; however, you should retain copies of all assignment computer files used during development of the solution to the assignment. These files must remain unchanged after report submission, for the purpose of checking if required.

There is no requirement for word count. Your report should demonstrate completeness in covering all research questions and brevity as no one loves reading long reports. “A picture is worth a thousand words” – use data visualisations to illustrate and support your research findings.

Do not include programming code and/or code output (e.g. screenshots of R or Python codes). Your readers are not programmers, they don’t understand code, so they don’t want to see the code. Including code in the report is a poor presentation style. If you need to include results of the analysis – make a proper table.

You are given a real-life data. It might be messy and not always easy to work with. This is very important to start working on the assignment early. This way you should avoid any unpleasant surprises later.

There is a penalty for late submission – 10 points per day or part of it beyond the deadline.

If you have any questions – feel free to ask on the forum. You can discuss this exercise with me and other students. You are encouraged to share ideas but not solutions. Remember about academic integrity.