Final

**Group Project**

# Introduction

For this assignment, you will use this data set titled **Snack Chain**. This file contains real sales and promotions data from a large retail chain (of 79 stores) on 58 products belonging to four product categories (bagged snacks, cold cereal, frozen pizza, and oral hygiene products) from multiple manufacturers like Frito Lay, Kellogg, and General Mills, over 156 weeks.

The data comes from three database tables: **stores, products, and transactions**, as shown in different tabs in the spreadsheet. The different tables and fields are shown in the Glossary tab. It has over 500,000 transactions and the file is 33 MB in size. Note that this is exactly how data is structured in a corporate database. You will have to join data from the different tables (using foreign keys) to run the analytics. This is easy to do in R and R handles big data just fine. We are not interested in oral hygiene products, so you can drop this product category and all associated transactions.

# Instructions

The **transaction table has weekly information on the price and promotions of products** (e.g., whether a product was assigned a special store display in an in-store circular, or had a temporary price reduction). We want to examine the effects of these pricing and promotion strategies on total spending for that product, the number of households who purchased that product, and the number of store visits. You have to control for a variety of different things (to be determined by you), perhaps do some feature engineering, and run different models (or sets of models).

At the end of your analysis, we are interested in answers to the following questions:

1. What is the effect of promotions, displays, or being featured in the circular on **product sales** (spend), **unit sales**, and number of **household purchasers**? (3 points)
2. How do the above effects vary by **product categories** (cold cereals, frozen pizza, bag snacks) and store **segments** (mainstream, upscale, value)? (2 points)
3. What are the **five** most **price-elastic** and five **least price-elastic products**? Price elasticity is the change in sales for a unit change in the product price. (3 points)
4. As the retailer, which products would you lower the price to maximize (a) **product sales** and (b) **unit sales**, and why? (2 points)

# Submission Details & Grading Criteria

## Submission Details

Submit your results in the form of a **nicely formatted Word or PDF file and your R code as two separate files.**

Note that this may not be just a time-series project. It may combine things that you learned at different points in time during the semester. This is our last assignment, so expect it to be complex.

The instructor will not tell you what to do or how to do it. You have to create an action plan , do the right feature engineering, select the right model specifications, and document your process, analysis, and findings in a succinct, nicely formatted, nicely written Word or PDF report.

Also, submit your R code as a separate file. We are not reiterating the importance of creating a table of predictor variables, doing data visualizations and descriptive analysis, using stargazer for summarizing output, assumptions testing, and cleanliness and compactness of code and writing. Those things should be a matter of habit by now, whether or not explicitly asked for.

There is no explicit page limit, but make your solutions as short (yet complete) as you can make it.

The assignment may look daunting on the surface, but this is a typical small-scale real-world retail analytics exercise; “small-scale” because you are analyzing only 3 tables and 58 products. In real life, you may have to work with 20-50 tables and tens of thousands of products.

Please refer to this guideline to drive you through the analyses: Template for the Retail Analytics Project.docxDownload Template for the Retail Analytics Project.docx

Please see an example of a similar final report to help you with the expected format: FinalProjectReport\_SDM\_Fabric\_Softener.docxDownload FinalProjectReport\_SDM\_Fabric\_Softener.docx

## Grading Criteria

The points allocated to each question are indicative of the importance of that question.

1 point for aesthetics: Whether you presented your analysis in a nice, compact, summarized manner, without overburdening the reader with unnecessary details or analysis.

As a potential Structure you may also refer to the syllabus, namely:

* Executive Summary
* Problem Definition & Significance
* Prior Literature
* Data Source & Preparation
* Variable Choice
* Descriptive Analysis & Data Visualizations
* Data Modeling
* Quality Checks
* Actionable Recommendations
* References
* Appendix (with R code)

***Submission Details***: Submit your report in Word or PDF format via this assignment. You will have two attempts to complete this assignment. Please review the rubric below for more details on the grading criteria for this assignment.

Determine every combination by using a nested for-loop

Add a list at the end that indicates the highest score and what features went into it (as a function)

Also a list for different models

Hyper-parameter tuning with RSC and GSCV