**Research Project**

This term project, to be done in teams of 3-4 students, is the heart of this class. In this project, you will apply the statistical concepts and methods you learned in class to solving a real-world data analytics problem.

You will identify a decision-making scenario of managerial or social relevance, source appropriate data, clean and merge the data as needed, use the right models, and derive actionable insights. Several times during the semester, you will present your intermediate project work, allowing me to keep track of your progress and guide you

The final project report (10-12 pages, plus appendix) is due at the end of the semester. This report must include the following sections:

o Executive summary

o Problem definition & significance

o Prior literature

o Data source & preparation

o Variable choice

o Descriptive analysis & data visualizations

o Data modeling

o Quality checks

o Actionable recommendations

o References

o Appendix (with R code).

The project report must be sufficiently detailed, include appropriate graphics, and be of professional quality. See the project report assignments on Canvas for further details. Sample projects will be posted on Canvas for your use.

**Research Project Considerations**

In the past, I have seen teams spend so much time searching for the right project that they did not have sufficient time to do a quality job with the project. Starting early and allocating sufficient time are critical for a good project. Your project topic and data source must be approved by me before you can start working on the project. Please use authentic data sources that fully describe the data collection process, which you must discuss in your report. Kaggle or UCI dataset of unknown quality is not acceptable; such data is good for personal learning, but not for serious analysis. Please avoid classification problems as that is not the focus of this class.

Note that your professor will compare your projects with similar projects on the Internet, run a plagiarism check, and ask you to explain your work. Feel free to browse online projects and learn from them, but your work must be sufficiently original, practical, and useful. Downloading an online project, making minor modifications, and submitting as your team project will be considered “plagiarism” and get the entire project team a zero grade.

The professor is available to help you with your project throughout the semester. Please use this help. I do give extra-credit to teams who surpass my expectations, experiment with new methods and tools, and help the class learn something new. In the past, students have used my class projects to get internships and jobs. But you can only do that if your project is sufficiently good!

E-mail me the names of your team members and a tentative topic for your project before the second day of class. Choose your team members “intelligently”, making sure that you have team members proficient in R, in reading and understanding technical papers, and in writing professional-quality reports. If someone in your team does not show up for team meetings or is perpetually late with their work, you have the option to fire that person by majority vote

**To clarify expectations: Assignments and Final project (via Announcements)**

I've reiterated this several times between the last class and our touchpoint. All expected deliverables are already assigned on Canvas—please check the assignments section.

The final project is the same for every group to ensure consistency across the cohort and across years. Given our compressed schedule, I cannot allow individualized projects. (https://usflearn.instructure.com/courses/1926140/assignments/17604973)

It would be wise to begin cleaning the data as soon as possible, as this is the most time-consuming step. Starting with our next class on March 9, we will discuss solution strategies.

**Introduction**

For this assignment, you will use this data set titled Snack Chain Download 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**

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 (2 points), 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 (3 points).

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.

**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.