***Background:***

Historically marketers have had poor information sources upon which to base their decisions.  John Wanamaker once said “I know half the money I spend on advertising is wasters, but I can never find out which half.”  The problem is that sales and advertising are usually measured in large, aggregate quantities that may occur over months or quarters.  It becomes difficult to attribute specific advertising decisions to their corresponding effects on sales.  Therefore many marketers have seen decision making as intuitive and difficult to quantify.

However, the introduction of computerized technology into the marketing environment over the past two decades has resulted in new opportunities for retailer managers.  Most managers now have a wealth and variety of informational at their disposal that is unprecedented.  We can track consumers on the web as the search for products with clickstream data, record advertising exposure with single-source data, potentially use RFID tags that are linked with loyalty cards to monitor movement of consumers within a store, and use optical scanners to record actual purchases.  No longer is the problem of not having enough information, it is a problem of having too much data or more precisely lacking the tools and training that allows real-time analysis of this information.

One example of a burgeoning area in analytical marketing is demand based management which uses statistical models to predict consumer price response using historical information (Montgomery 2005).  The most prevalent type of information in retail markets is transaction data collected using optical bar code scanners which track every item purchased by a consumer at the point-of-sale.  This data potentially contains a wealth of information about how consumers respond to price and promotions.  The promise is that this information can be used to recommend optimal pricing and promotional strategies, and more broadly incorporate quantitative approaches for making marketing decisions.  This problem presents a nice microcosm of analytical marketing as a whole, since a successful solution will require quantitative analysis to turn this information into a model of consumer behavior, an analytical problem that permits the construction of an optimization algorithm, and a strategic problem for the manager in integrating these results into practice.

***Problem Statement***

This problem was motivated by challenges that supermarket retailers have been facing over the past two decades of a new type of retailer, the everyday-low price (EDLP).  Walmart is perhaps the most famous example of an EDLP retailer.  Traditionally, supermarket retailers have relied heavily upon promotional pricing to periodically drive consumer traffic into a category or the store, but typically everyday prices are kept higher, giving this pricing strategy the nickname Hi-Lo.  Hi-Lo pricing strategies are the dominant one for grocery retailers.

The retailer in this example is Dominick’s Finer Foods (DFF) with a total of 88 supermarket stores located primarily in the Chicago metropolitan market.  This market has more than eight million residents and DFF has about a 20% share of supermarket sales.  There is only one larger competitor, Jewel Foods, and then many smaller competitors.  DFF was specifically interested in understanding whether if it moved from a Hi-Lo pricing strategy to an EDLP strategy whether it could become more profitable and effectively compete against an EDLP competitor.  A more detailed discussion of the research project undertaken can be found in the article by Hoch, Dreze, and Purk (1994).

The basic tradeoff that the retailer faces is a classic one: will a substantial decrease in price be offset by a suitable increase in quantity sales?  In this exercise students are asked to answer this question by constructing a demand-based model for sales (unit movement) for the refrigerated orange juice category.  The data base was collected from the weekly store scanner data and contains information about movement (sales), price, feature, display, and profit for each item in the category.

***Category Background***

Refrigerated juice is a relatively stable category with moderate new product introductions and heavy promotion activity.  It is a large volume category, accounting for 1% of total store sales.  This category is more highly developed in urban areas, despite lower income levels.  This may be due to convenience and fewer kids in such areas.  For example, New York City has a 400 index for refrigerated juice (the average market has an index of 100).

Two-thirds of this category is driven by orange juice, 75% of this volume is “from concentrate” the other 25% of this volume is “pure premium” not from concentrate.  Juice drinks (e.g. lemonade, punch, etc.) represent 20% of the refrigerated juice category, grapefruit juice represents 4% and miscellaneous juices the remaining 9%.  In order to keep the size of the dataset smaller only refrigerated orange juice is provided.

Refrigerated juice is a price sensitive category.  In 1990, orange juice prices increased 45% at retail due to frost damage to the crop.  Sales dropped over 25% as a result of this price increase.  1991 prices and sales have recovered and returned to historical levels.

There is a natural division of products into three price-quality tiers: the premium brands (made from freshly squeezed oranges), the national brands (reconstituted from frozen orange juice concentrate), and the store brands (Dominick's private label).  There is quite a bit of disparity in prices across the tiers, which leads to large differences in wholesale costs, even though the profit margins appear similar.  An initial indication that store differences are present is the variation of market shares across stores.  Dominick's 64 ounce OJ brand has an average market share of 13.6%, but the market shares across stores range anywhere from a minimum of 5.6% to a maximum of 20.9%.

***Data***

The data is the key to any demand-based price optimization approach.  Fortunately, retailers have a readily available information source which can be used to measure demand: weekly store-level scanner data.  It should be pointed out that retailers implemented optical scanners primarily to improve their product tracking, reduce inventory costs, and reduce transaction costs.  Hence, the primary use of this data for the retailer has largely been for inventory and accounting purposes.  Which illustrates a central theme in creating effective analytical marketing strategies, namely that data must be warehoused in a form that is accessible in many different ways depending upon the user’s goals.

There are 33 UPC’s or individual types of orange juice sold at DFF, many of these a brand by size combinations.  In order to create a more manageable number of products we create five aggregates from the original UPC level data that have similar pricing and promotional strategies.  The UPCs within a product aggregate differ only by flavoring, additives, or packaging (e.g., regular, pulp, or calcium).  (Example: There are 2 UPC's that comprise the Tropicana Premium 64 oz aggregate: homestyle and regular.)  The prices within an aggregate are typically the same, but when they differ an average of prices weighted by their market shares is used.  The movement or quantity sales of each aggregate are computed as the sum of the movement (standardized to ounces).  Moreover we can still speak about profit maximization for these aggregates since we can easily map the aggregate prices back to the UPC level.

Additionally, the data is reported only for a single store; however, the entire dataset is available from the University of Chicago at [(Links to an external site.)Links to an external site.](https://www.chicagobooth.edu/research/kilts/datasets/dominicks)

[https://www.chicagobooth.edu/research/kilts/datasets/dominicks (Links to an external site.)Links to an external site.](https://www.chicagobooth.edu/research/kilts/datasets/dominicks)

In addition to shelf-price and units sold, we also have information about promotions and gross-profit margins.  Information about feature advertising in weekly newspaper fliers is provided by IRI's Infoscan, which provides an estimate of all commodity volume of a particular UPC that received feature advertisement.  In-store promotion is measured using a deal code provided in DFF's store-level scanner database.  The deal code is a dummy variable which shows whether there was a bonus-buy tag on the shelf or an in-store coupon.  Since these promotional variables are at the UPC level, averages of these variables are taken in case one UPC within an aggregate is promoted while another one is not.  These averages are weighted by their market shares.  Finally, the gross profit margin was reported by DFF and represents the ratio of the difference between price less the wholesale cost and price (e.g., gross profit margin = ( price – wholesale cost ) / price ).

The data is provided in the sheet labeled ‘Data’ in the OJ.XLS spreadsheet.  To make it easier to compare across brands all sizes are divided by the number of ounces, so prices are expressed on a per ounce basis, and movement reflects the number of ounces sold (not the number of cartons).

***Instructions:***

The questions for this exercise are based on the following data and R scripts:

[data.zip](https://canvas.cmu.edu/courses/9526/files/3580306/download?verifier=c5FTw9tEexVyjqOyJjT1T1eKlBvTbz1oZsS5yrVW&wrap=1)  (This is a zipped file with comma delimited text datasets.  You need to unzip them)

[rfj\_Script.R](https://canvas.cmu.edu/courses/9526/files/3580305/download?verifier=wb6LeNREvVBtQz9I1I6JphcSkPDvO0ctJg0a3oXn&wrap=1)  (This is a hint that you will need to extend to complete the assignment.)

This exercise is to be completed in your group as listed on Canvas.

Please provide a clear, concise, and well organized essay that addresses at least the following questions.  You are free to address other issues in the case as well.  The intent of the assignment is to have you think critically about the business problem faced in the case and how it can be solved through data mining.  Analyze the quantitative material in the case to support your answers.  Spend most of your time in defining and defending your recommendation for what should be done.

Good answers may require assumptions of facts that may not be presented in the case.  You are welcome to make these assumptions, but please state these assumptions and briefly justify why that are reasonable.  Also, you may use whatever resources you can locate to provide further information about this industry or the web in general.  Please reference your sources.

Your response must be typed, double spaced, with one-inch margins, and a 10 to 12 point font size, and must be a PDF file.  You may attach exhibits, tables, and/or graphs to support your arguments.  These supporting materials must be referenced in the text.

Split the dataset for cross-validation purposes into an estimation and hold-out sample.  Specifically, randomly flag 30% of the observations.  Additionally, you can (although you do not have to) remove the smallest products in the category (specifically only retain the products that generate 90% of the sales in the category).

***Required:***

Part 1)

1. Estimate a log-linear regression model using the refrigerated orange juice category for Tropicana Premium 64 ounce (UPC # 4850000102). Regress the natural logarithm of quantity on price, feature, and display.
2. Estimate three different models: a single model for all stores ("pooled" model), a separate model for each store ("store" model), and propose an "improved" model (hint: try to include competitive effects).  For each model compute the mean-error, root mean-squared error, and mean-absolute deviation separately for the estimation and hold-out samples.  Summarize your results and comment on the accuracy of each model.
3. Select UPC # 4850000102 in store #2 and check the adequacy of a linear regression model by plotting the residuals using a histogram and time series plot.  Comment on the adequacy of the estimate models.

Part 2)

1. What is the implied optimal price for the log-linear demand model? (You only need to write out the formula, and do not actually calculate this formula using the data.  Hint: use the formula give in Lecture #5.)
2. Using the regression results that you computed in part 1 and the formula from Question #4, compute the optimal price for each store using the wholesale costs from week #100.  Do this for each of the three models that you estimated in Question #2: {pooled, store, improved}.
3. Compare your predicted optimal prices with what Dominick’s actually charged in week #100 and summarize your results. (Hint: compute the mean and correlation between your predicted optimal price and the actual price charged.)  What model gives you the best results?  What patterns do you notice?  If you were the Dominick’s pricing manager what price would you implement?