# Analyzing IRI Marketing Dataset Yogurt Sales

Megan Mccarty, Julius Remigio, Ryan Riopelle, Syed Nazrul, Michael Galarnyk

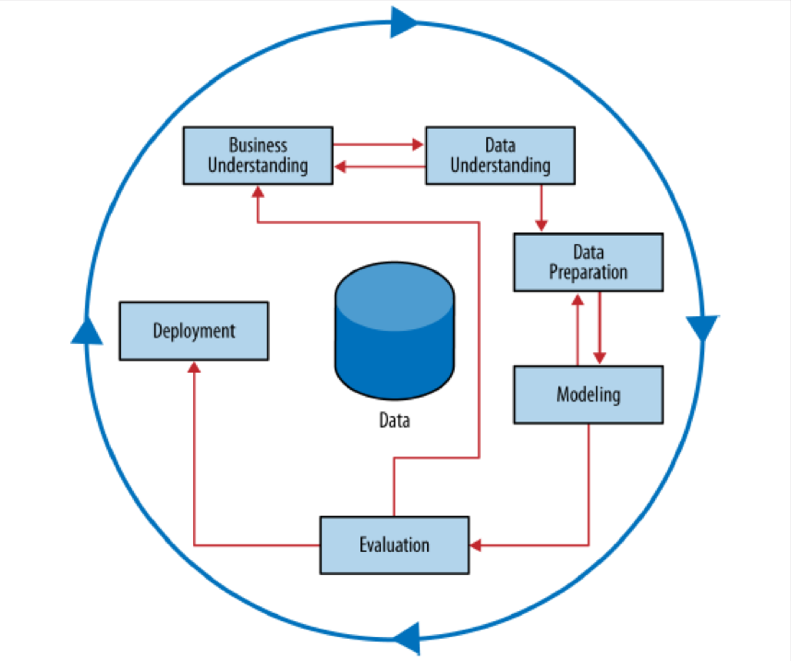
## Goals and Objectives

For our project, we analyzed yogurt sales data from the IRI dataset. The IRI dataset contains 11 years of weekly store data for chain grocery and drug stores for 30 consumer product categories in 47 markets. Be more efficient with resources.

1. Better predict whether promotions lead to an increase in sales and what particular marketing campaigns work best.
2. Better predict who to target with promotions (Julius). Do more upper or lower income people prefer yogurt?
3. Understand where geographically companies should target and have deeper inventory.
4. Predict total sales (sum of units sold/number of stores go through Ryan’s).

Methodology

The Cross Industry Standard Process for Data Mining (CRISP-DM) was applied (see Figure 1.) to help analyze the data.



**Figure 1. Schematic Illustration of the CRISIP-DM Methodology.**

Business Understanding

**Determining Business Objectives**

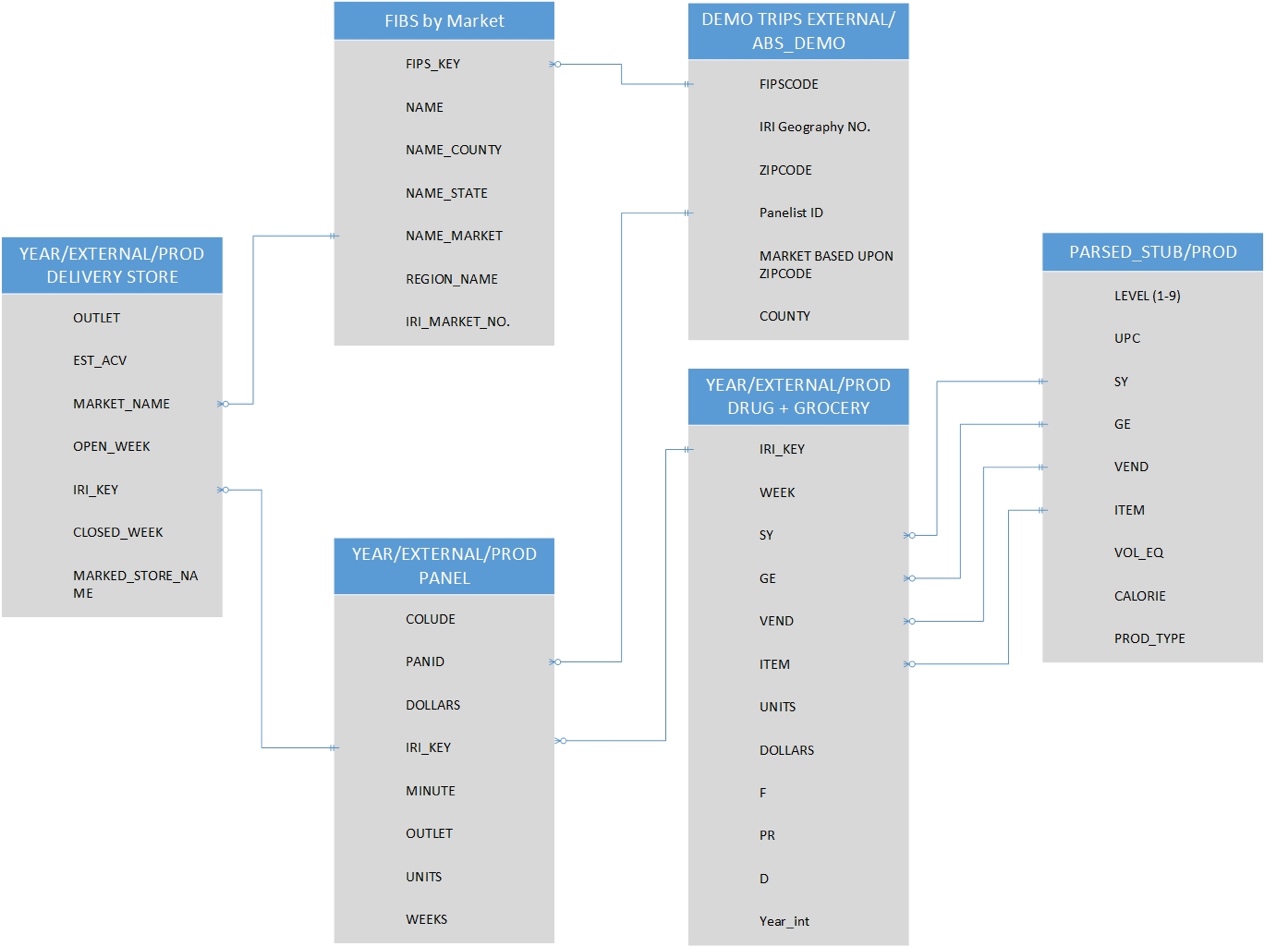
**Assess Situation**

**Determining Data Mining Goals**

Data Understanding

**Collect Initial Data**

**Describe Data**



**Figure X.** Modeling the relationships between the various data sources.

## By using the academic data set file and field description pdf, our group wanted to model the interrelationships between the various data sources to decide which data to use in our data mining life cycle.

**Explore Data:**

**Verify Data Quality:**

Panelist Data: 30% of the panelist have never purchased yogurt. The documentation doesn’t specify the criteria for grouping these panelist as part of the yogurt data. We will assume that these panelists are representative of those who did not purchase yogurt.

Grocery Stores:

Issue number 1 is due to the massive size of the data, we had to randomly sample the data. Issue number 2 is the population disparity between the actual census population and the specific cities that have data available. This could greatly throw of the data. Another factor is the census data from 2010, this time frame does not match the IRI data timeframe which spans 2001 to 2011. This makes dividing by the total population extremely inaccurate, so the per capita table listed above should most likely be disregarded accept for general inquiries about the data.

Demo Data: Year 1 and 2 were missing significant amounts of data.

Data Preparation

**Selecting Data:**

Our group decided on the following sets of data to analyze.

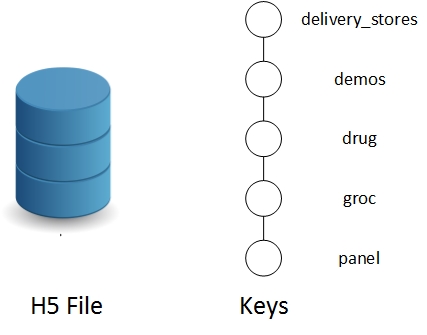
Panelist Data: Initial analysis of the data showed us that we had to use the data from 2008-2011. Panelists give information about…

Delivery Stores: Knowing where the yogurt is being sold is important for analysis.

Drug and Grocery Stores: Seeing how sales differ from store type to store type is important for supplying stores with adequate inventory across the country.

Demo Data: Demographic data is always important.

**Clean Data:**



**Figure x.** ETL process used for transforming the IRI data into a more usable format.

Modeling and Evaluation

**Select Modeling Technique** (Modeling Technique, Modeling Assumptions)

**Generate Test Design**

**Build Model Parameter Settings**

**Assess Model**

## Assess situation: talk about issues we ran into.

Megan insert bullshit here.

2007 didn’t make sense. Ran into problem we couldn’t use something we wanted to use. Panelist don’t necessarily buy yogurt.

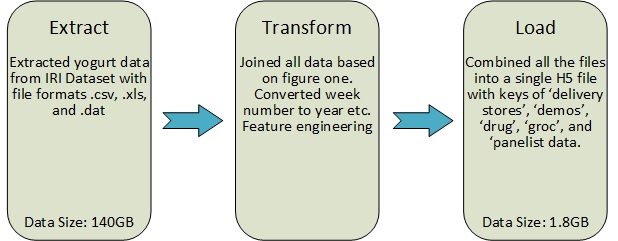
Data didn’t

Data Preparation: feature selection and any feature engineering (did you add columns). Also include number.

Modeling: Geography isn’t a model. Go into data understanding. Grid search parameter tuning to go into.

Evaluation: Cross Validation

Deployment:



**Figure x.** ETL process used for transforming the IRI data into a more usable format.

# Figure x. shows the ETL used for combining the data form the many different file formats into a single H5 file. From the roughly 140 gigabytes of data, our group reduced the data to roughly 1.8 gigabytes. Additionally, while a lot of the data was in csv, xls, and .dat file formats, our group organized it all into 1 h5 file which has made the data modeling process easier. The keys in the H5 files are ‘delivery stores’, ‘demos’, ‘drug’, ‘groc’, and ‘panelist data’.

<http://datastreams.co.kr/en/wp-content/uploads/sites/3/2015/12/Data-Integration-.png> (data integration)

<https://en.wikipedia.org/wiki/Extract,_transform,_load> (ETL)

<http://datastreams.co.kr/en/wp-content/uploads/sites/3/2015/12/Data-Integration-.png>

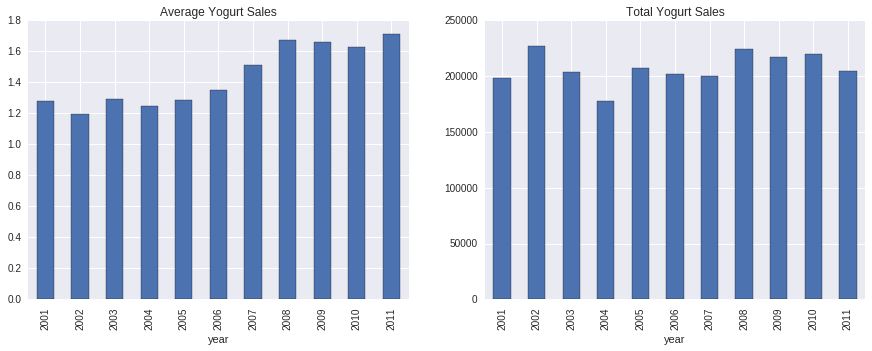
(ETL process image)

Modeling and Evaluation

Using panel data, we used the WEEK attribute to calculate the year after 1979. Based on this, we managed to calculate the yogurt sales. We see the average and total yogurt sales remain relatively unchanged.

(note missing y axis. X axis years need to be horizontal for readability. Absolutely no units, also not a Figure caption below yet)

Understanding. Superscript.

****

We also managed to analyze the day and time of the week of yogurt sales using the minute column of our panel data. For this analysis, we only managed to look at data between 2008-2012 as minute data was missing before 2008.

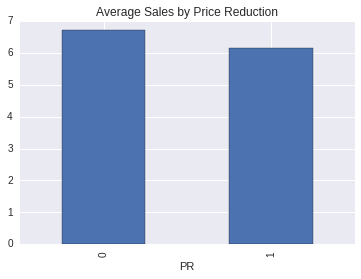
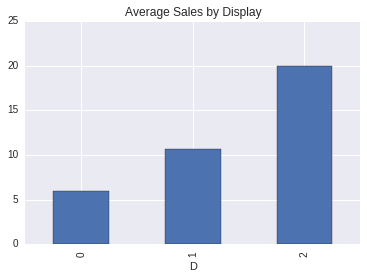


**Figure 2. Monthly, Daily and Weekly Yogurt Sales.**

Based on this plot, we see that people tend to buy yogurt closer to the weekends and during the afternoons.

**Sales Promotions**

We analyzed the effects of advertisements on sales using a 2-tailed p-test with 5% tolerance. Using this test, it is clear that the display size and price reduction had a statistical significance on increasing yogurt sales at drug and grocery stores.

  
    **Fig. Sales Promotions at Drug Stores**