Analyzing IRI Marketing Dataset Yogurt Sales

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



**Analyzing IRI Marketing Dataset**

Table of Contents

[1.0 Goals and Objectives 2](#_Toc453262152)

[2.0 Methodology 3](#_Toc453262153)

[3.0 Business Understanding 4](#_Toc453262154)

[3.1 Determining Business Objectives 4](#_Toc453262155)

[3.1.1 Assess Situation 4](#_Toc453262156)

[3.1.2 Identify Project Objectives 4](#_Toc453262157)

[3.1.3 Determining Data Mining Goals 4](#_Toc453262158)

[4.0 Data Understanding 4](#_Toc453262159)

[4.1 Collect Initial Data 4](#_Toc453262160)

[4.1.1 Describe Data 4](#_Toc453262161)

[4.2 Verify Data Quality 5](#_Toc453262162)

[5.0 Data Preparation 6](#_Toc453262163)

[5.1 Selecting Data 6](#_Toc453262164)

[5.2 Clean Data 6](#_Toc453262165)

[6.0 Modeling 7](#_Toc453262166)

[6.1 Assess situation 7](#_Toc453262167)

[6.2 Select Modeling Technique 8](#_Toc453262168)

[6.2.1 Modeling Technique 8](#_Toc453262169)

[6.2.2 Modeling Assumptions 8](#_Toc453262170)

[6.2.3 Generate Test Design 8](#_Toc453262171)

[6.3 Build Model Parameter Settings 8](#_Toc453262172)

[6.4 Assess Model 8](#_Toc453262173)

[6.4.1 Cross Validation 8](#_Toc453262174)

[6.5 Evaluation 9](#_Toc453262175)

[7.0 Deployment 11](#_Toc453262176)

[8.0 Conclusion 11](#_Toc453262177)

Table of Figures

Figure 1 3

Figure 2. Crisp Model 4

Figure 3. Modeling the relationships between the various data sources 5

Figure 4. H5 File Description 7

Figure 5. ETL Process Diagram 8

Figure 6. Average Yogurt Sales 9

Figure 7. Monthly, Daily, Weekly Yogurt Sales 10

Figure 8. Sales Promotions at Drug Stores 10

# Goals and Objectives

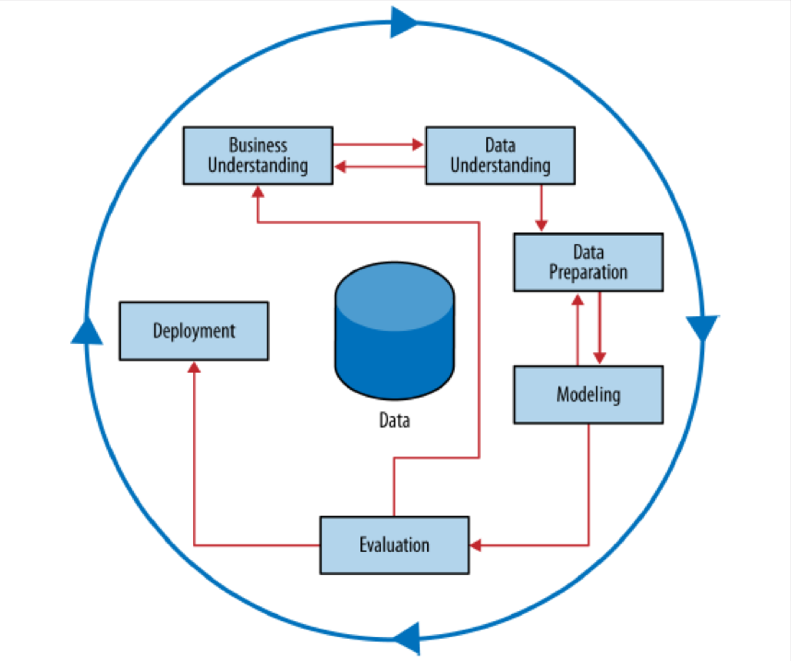
For our project, we analyzed yogurt sales data from the IRI dataset. The IRI dataset contains 11 years of weekly store data for chain grocery and drug stores for 30 consumer product categories in 47 markets. The business objective of our work is to analyze the IRI Dataset in an effort to identify trends and influential features in the sale of yogurt. Through the use of multiple predictive modeling techniques, we have developed the capability for proactive evaluation of marketing resources

Desired Capabilities

1. Predict the impact of new promotions on future sales.
2. Allow focused marketing by identifying the driving demographics in yogurt sales.
3. Aid inventory decisions by understanding geographic sales trends.
4. Predictions of future sales to better inform business decisions

# Methodology

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



Figure

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

# Business Understanding

## Determining Business Objectives

This analysis is driven primarily by a marketing business objective of

### Objectives

Analyze the IRI Marketing Dataset in an effort to identify trends and influential features in the sale of yogurt. Develop the capability for proactive evaluation of marketing resources through the use of predictive modeling

### Assess Situation

The evaluation of the yogurt data includes data science methodologies and algorithms that can be utilized to extract hidden information from the data. These trends can be utilized answer questions about the data, and predict future outcomes that can be used to directly address our business objectives and improve marketing.

### Data Mining Goals

Data mining is a process that contains multiple techniques for exploring data based on data science methodology. An important problem in data mining is the development of model based off of specific features in the data. Developing an efficient method for utilizing these features for classification or regression is the ultimate goal. It is also important to address quality problems within the data before the data is passed to the model. This includes removing outliers in the data and creating or removing features.

### Project Plan

The project plan for data analysis was performed according to the CRISP-DM model. The project plan is provided below for each step of the process as we explore, manipulate, and deploy predictions based on the specific business needs.

Figure . Crisp Model

# Data Understanding

Data understanding is integral the overall success of the project. The processes for data understanding are listed below.

## Initial Data Collection

The data was obtained from IRI Worldwide, a company that focusing in market research. The company provides clients with consumer data, shopping data, and retail market intelligence and analysis focused on the consumer packaged goods and (CPG) industry.

## Describe Data

The data that was provided was fairly scattered across multiple folders and contained information for various products that are sold at grocery stores and drug stores across the U.S. The data was broken up by years, store locations, store type, products, product types, demographic information, geographic information, and panel data obtained from specific stores at specific locations. This data was translated using Linux commands, Bash, and Python into a format that could be better utilized by the team (HDF5) . This structure is shown below as a database format of tables and relative columns within each table.

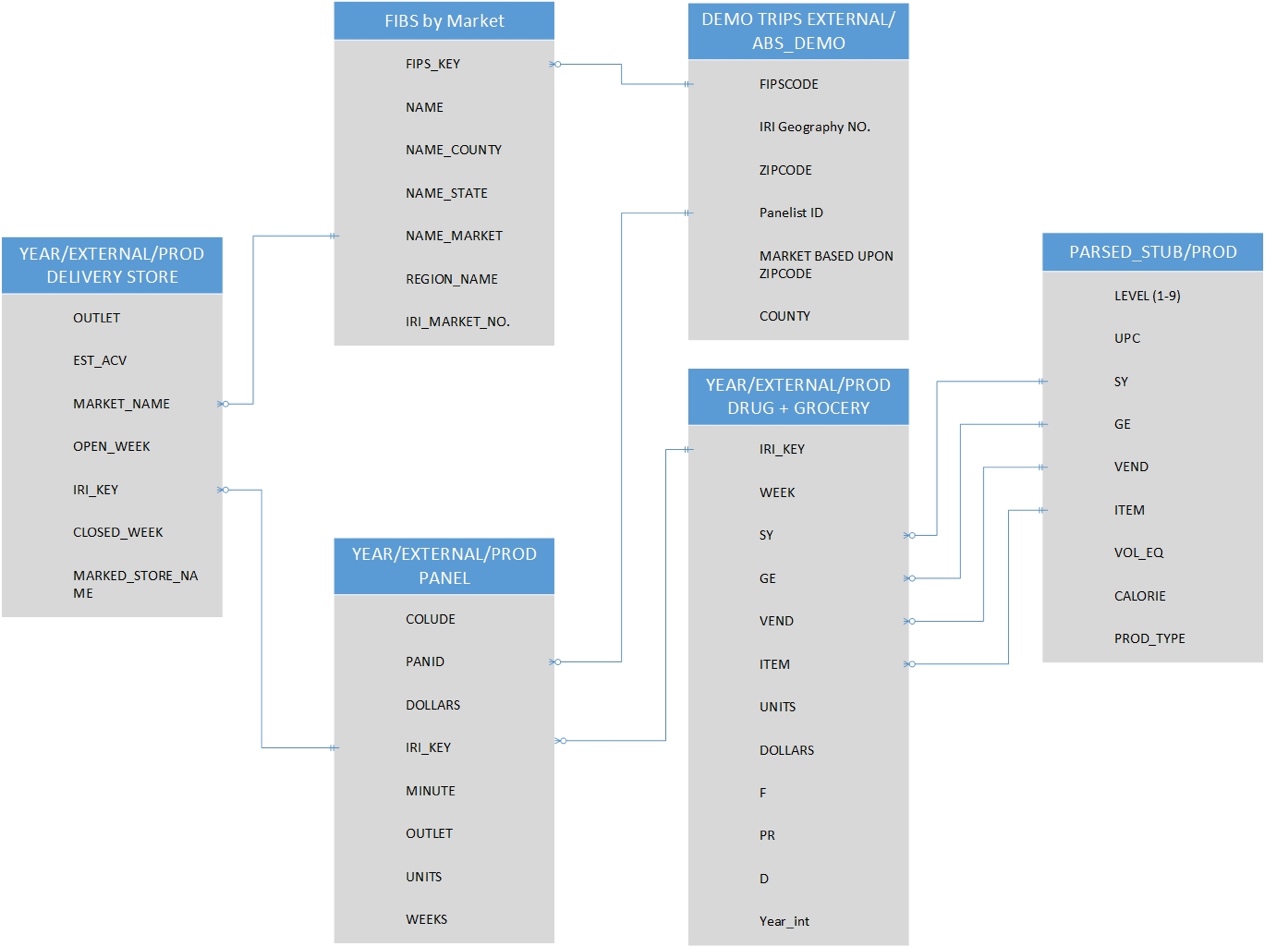


Figure . Modeling the relationships between the various data sources

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.

## Data Exploration

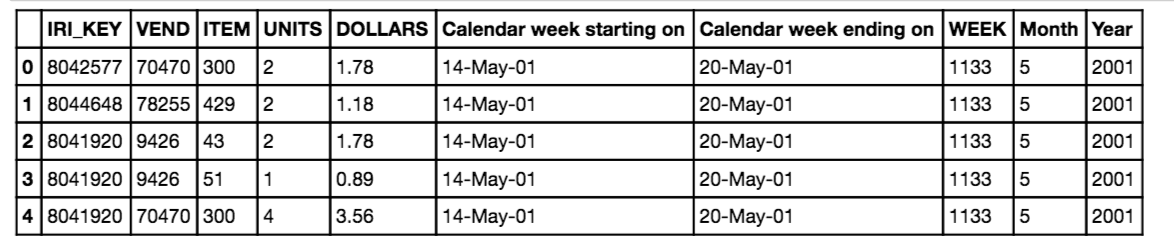
The entire IRI Dataset consists of multiple tables which may play a role in consumer trends. These tables include delivery (delivery\_stores), demographics (demos), drug store (drug), grocery (groc), and panel data. Each of these tables contains important information with data specific to its table name.

### Demographics Exploration Using K-Means (Julius will fill out)

### Drug Store Sales Exploration

In order to obtain geographic data from the HDF file, we linked the delivery store table to the drug store table using the common ‘IRI\_KEY’, which describes the specific ID for each store in the table.

Table . Drug Store Table



Each IRI\_KEY is listed on multiple rows with specific data pertaining to the units sold, dollars sold, week, month, and year. In order to get summary statistics for the geographic data on drug stores, we summarized mean sales units by year. We also counted the number of total available data points across the 10 year span. Both of these were graphed and are shown in figures X.X.

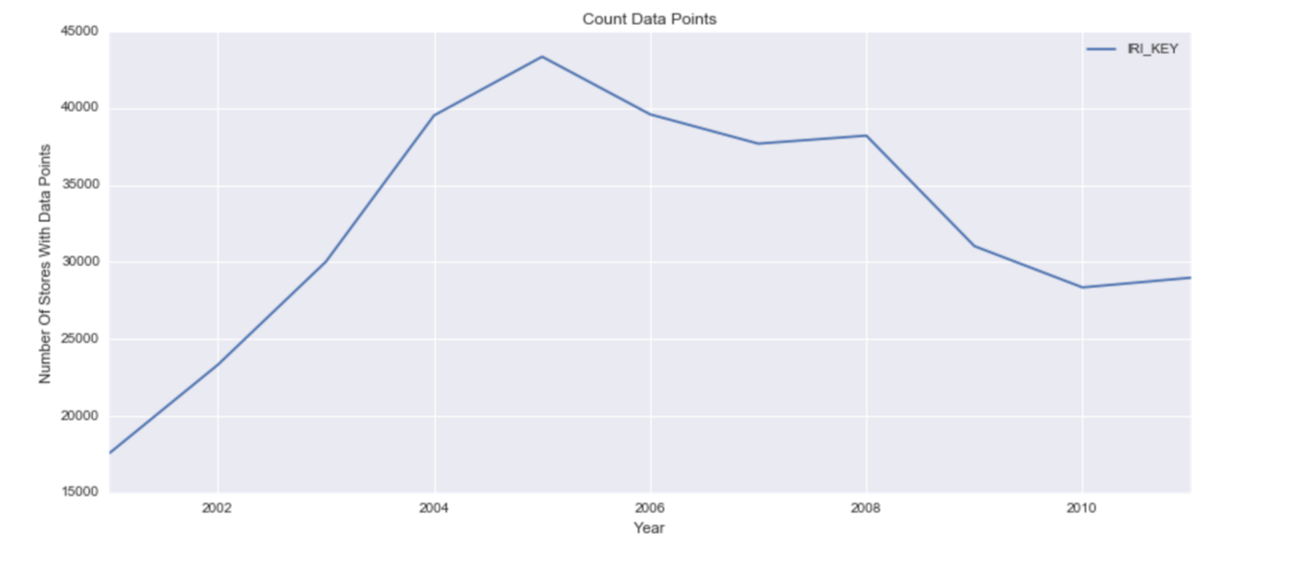


Figure . Count Data Points

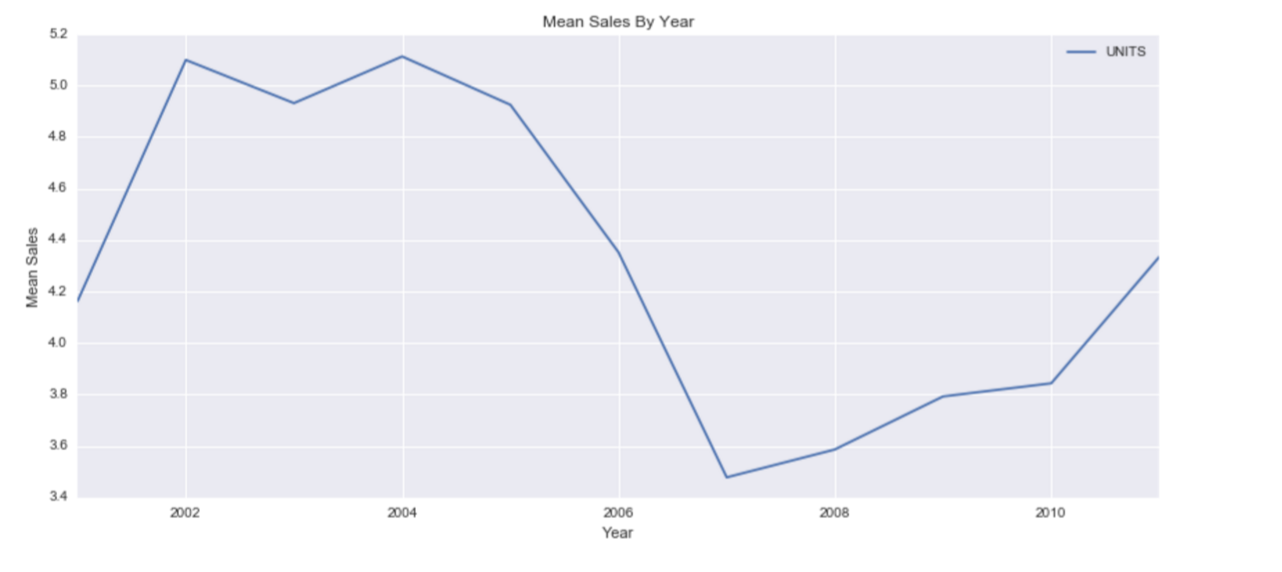


Figure . Mean Sales By Year

We see that the total peak point for sales was reached in 2004 for the yogurt. There is a dramatic reduction in mean sales over the years from 2004 to 2007. We do not currently know the cause of this dramatic decline. If we cross reference the mean sales with the total available data points, there is minimal correlation between the two. This provides evidence that the decline is independent of the number of available data points.

Mean unit sales per month was also investigated, this table is provided below.

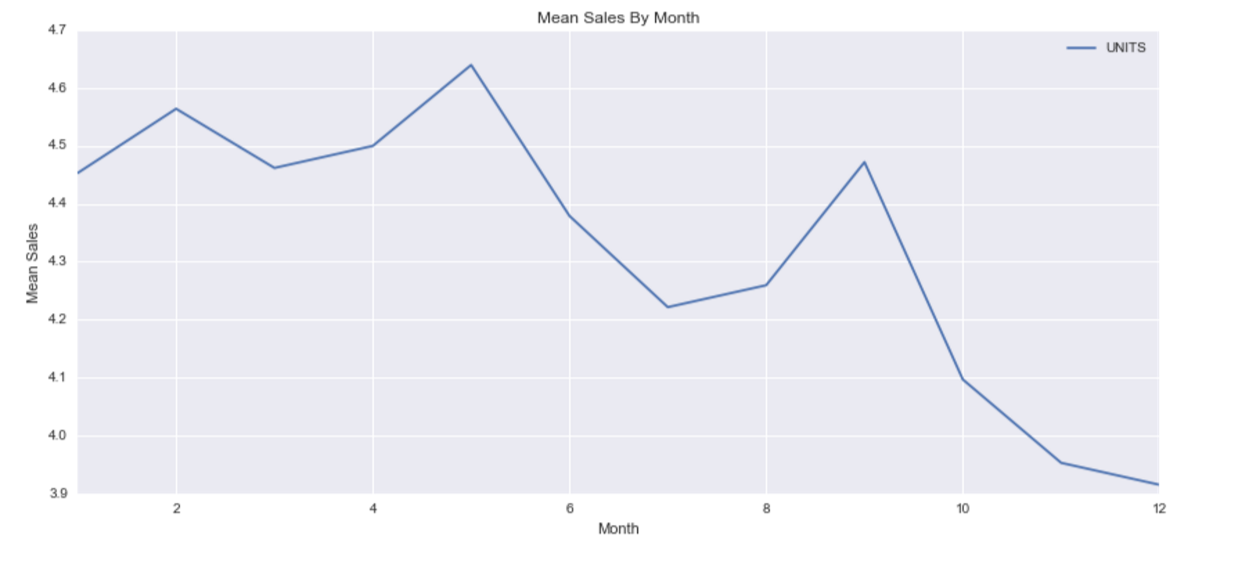


Figure . Mean Sales By Month

The data looks as though it is pretty consistent across all of the months with sales experiencing a slight low during months 10-12 (October-December). We presume that this decline in sales is due to a reduced incentive to buy yogurt during cold winter months as warmer comfort foods are most likely preferred by customers.

Mean unit sales by week were also investigated for weekly trends. This graph has a great deal of noise due to the number of samples across a year. Later we will investigate how PCA can be utilized to reduce noise in the data.



Figure . Mean Sales By Weeks Of Year

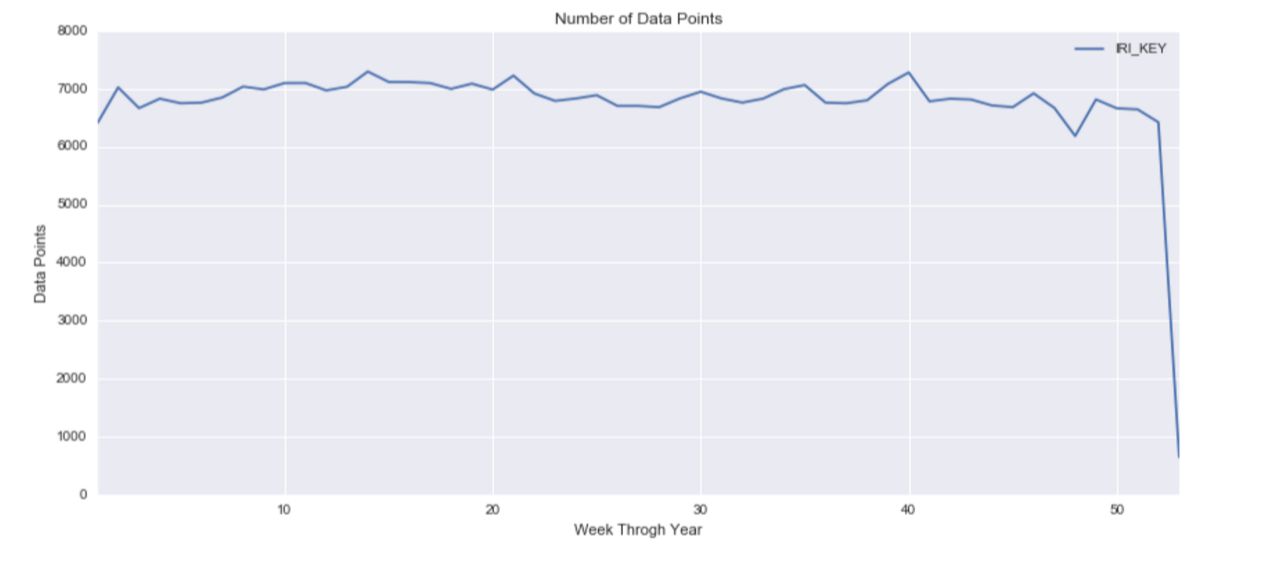


Figure . Number of Data Points

The data above shows a trend that corresponds to the monthly data points. A slight decline in sales is present in the weeks that fall during the October through December months. For reasons stated above, we believe that this trend may be do to consumer preference for yogurt during the warmer months in comparison the winter months.

### Grocery Store Sales Exploration

Similar trends were observed in the grocery store sales, but the data was so large for the actual sales data that that only small portions were able to be observed at one time. The H5 file that was created contains this data. In order to receive accurate predictions random sampling was used across the entire data set to obtain a representative sample. Further charts and graphs are provided in the notebooks and are not shown here because the descriptive charts follow the same trends as the ones shown above.

## 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.

### Drug Stores

Multiple assumptions we made for the drug stores. Many of these assumptions may not accurately portray the data and should be considered when analyzing the data trends listed in the data exploration section. These data assumptions may be less important in the overall machine learning classification of the data due to the fact that the algorithms pics up on trends that may not be apparent to the human observation.

The first is the population disparity between the actual census population and the specific cities that have data available. It was unclear if the data from each city represented specific drug stores or if it represents a fraction of the total drug stores in the city. If only a small percentage of stores were included in the data sampling, the data for total sales could be greatly skewed. The data for mean sales on the other hand would still be accurate. In this report, we provide both total sale and mean sales for the analysis.

Another factor is the census data from 2010, this time frame does not exactly match the IRI data timeframe which spans 2001 to 2011. This makes dividing by the total population inaccurate, but only by a small percentage increase in the population between 2010 and the time the original data was sampled.

# Data Preparation

## Selection

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.

## Cleaning

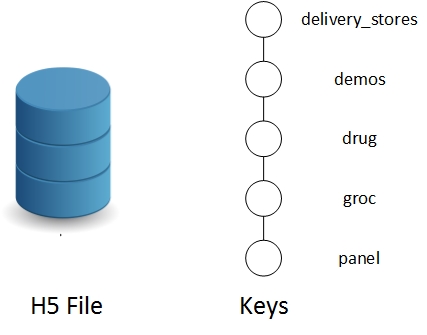


Figure . H5 File Description

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

## Feature Engineering

### Clustering Demographic Information (Julius)

The Demographic information was initially clustered as an exploratory analysis to discover specific trends in the data that could be used for later regression and classifications models. Demographic trends were discovered using K-means algorithms. These clusters are described below.

Describe the clusters…

### Geographic Information for Drug Stores

After investigating the trends in the drug store data across time, we joined the drug store sales with the store information table. The resulting graph and columns are shown below. The specific columns that were kept are listed here: ('IRI\_KEY', 'VEND', 'ITEM', 'UNITS', 'DOLLARS', u'EST\_ACV', u'Market\_Name',u'Month', u'Year','WEEK', u'name','Calendar week starting on', 'Calendar week ending on').

Sales by market graph shows the total sales per each market area aggregated using a group by. This bar graph helps give a general sense for where most of the sales are being made for yogurt across drug stores. We can see that Chicago has the highest total sales for yogurt by a large margin compared to all other areas. *As mentioned in the evaluation section of this report, it is unclear what the sample percentage is for the city, i.e. how many stores were sampled out of the total number of stores in the city.* For this report we will provide both total sales and mean sales data. The mean sales data can provide a better representation of the data, as the mean is not dependent on the total number of stores sampled.

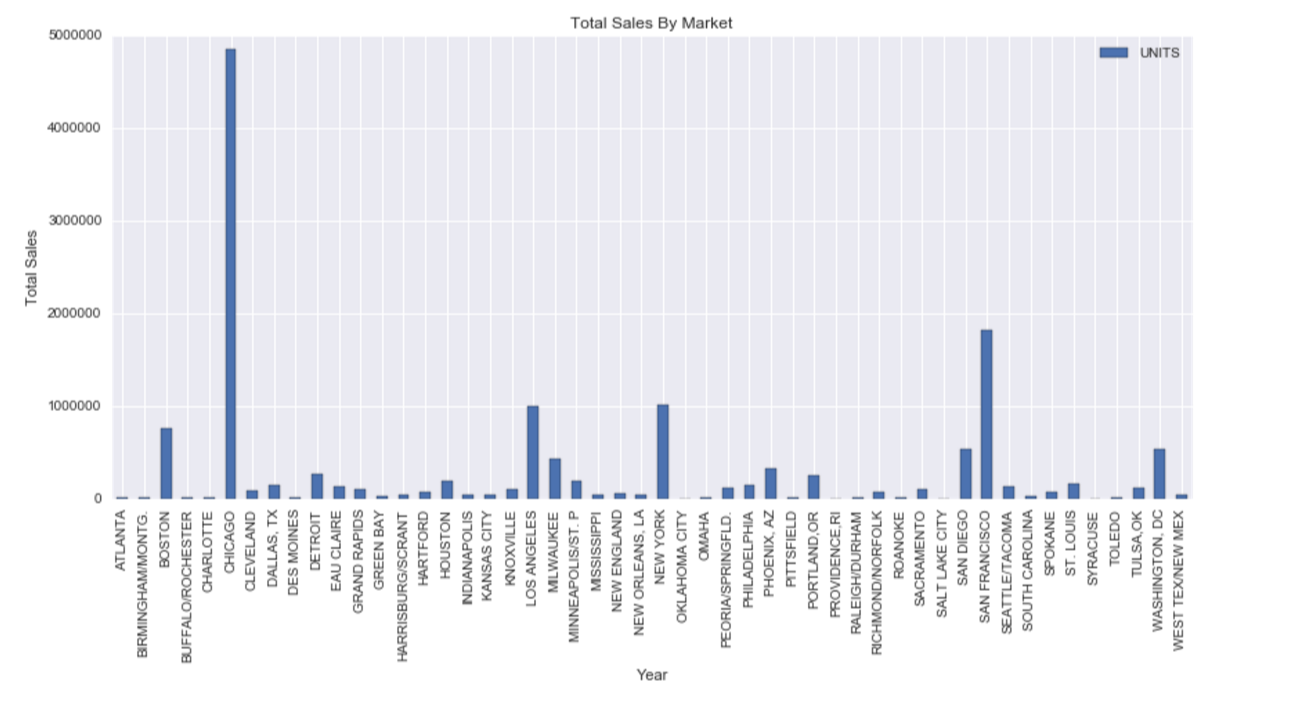


Figure . Total Sales By Market

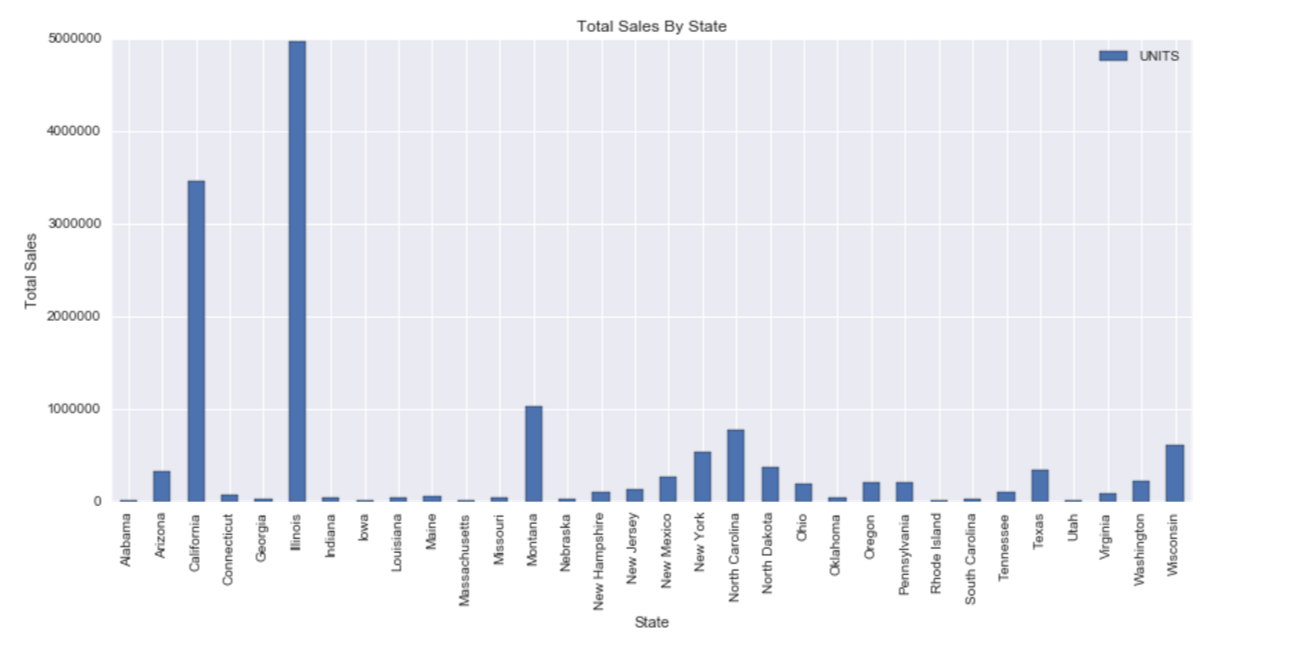


Figure . Total Sales By State

We sales data shown in figure X.X shows that the highest total sales for yogurt occurs in Chicago, San Francisco, New York, and Los Angeles. This is consistent with our assumption that larger cities would sell more yogurt due to a larger consumer base. Similarly the total sales were graphed by state, for the same conclusion.

We wanted to further investigate this trend, by graphing the trends across states and regions of the country. In order to do this, we mapped the geographic locations to each state and grouped based on states. The following graph shows the mean sales by state. We can see that this graph is almost identical to the graph above, as the data tended to be distributed equally across all states, i.e. each state has only one or two cities that were listed within in it.

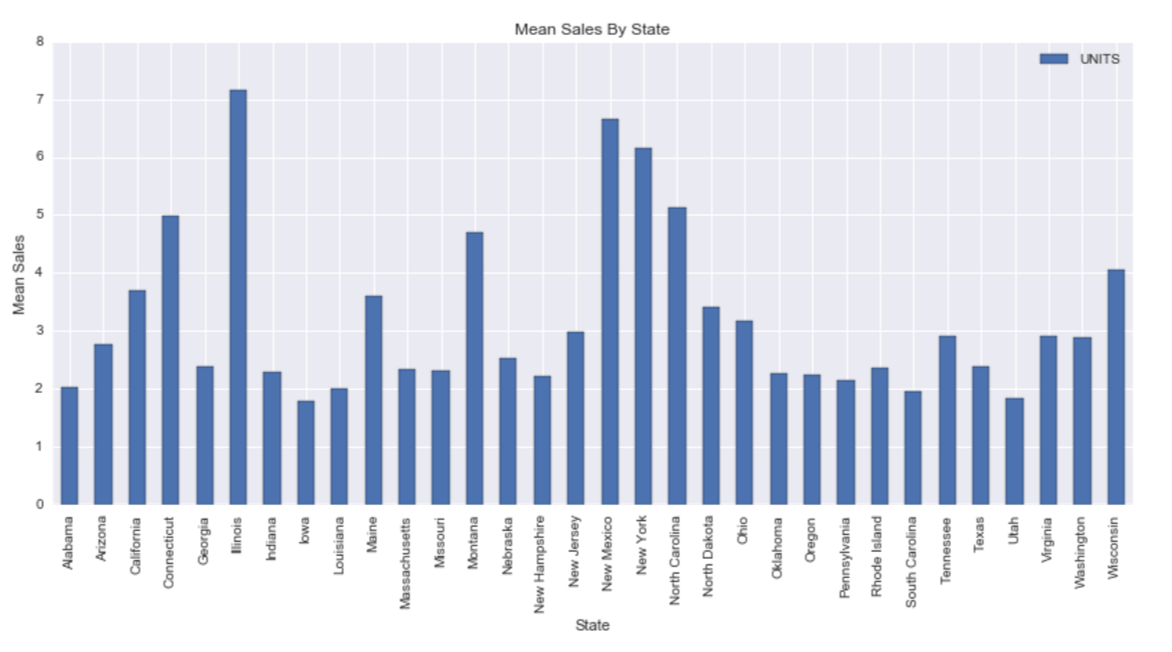


Figure . Mean Sales By State

This figure shows that in general, stores sell more yogurt to people in the states of Illinois, New Mexico, New York, North Carolina, and Connecticut. Further analysis may be warranted to determine how the marketing campaigns differ across the locations. These states listed above could be used as model states for their marketing campaigns.

Form here we chose to group the sales trends by region of the country. In order to do this, we looked the US census data 2010 for information on how to group the states by region.

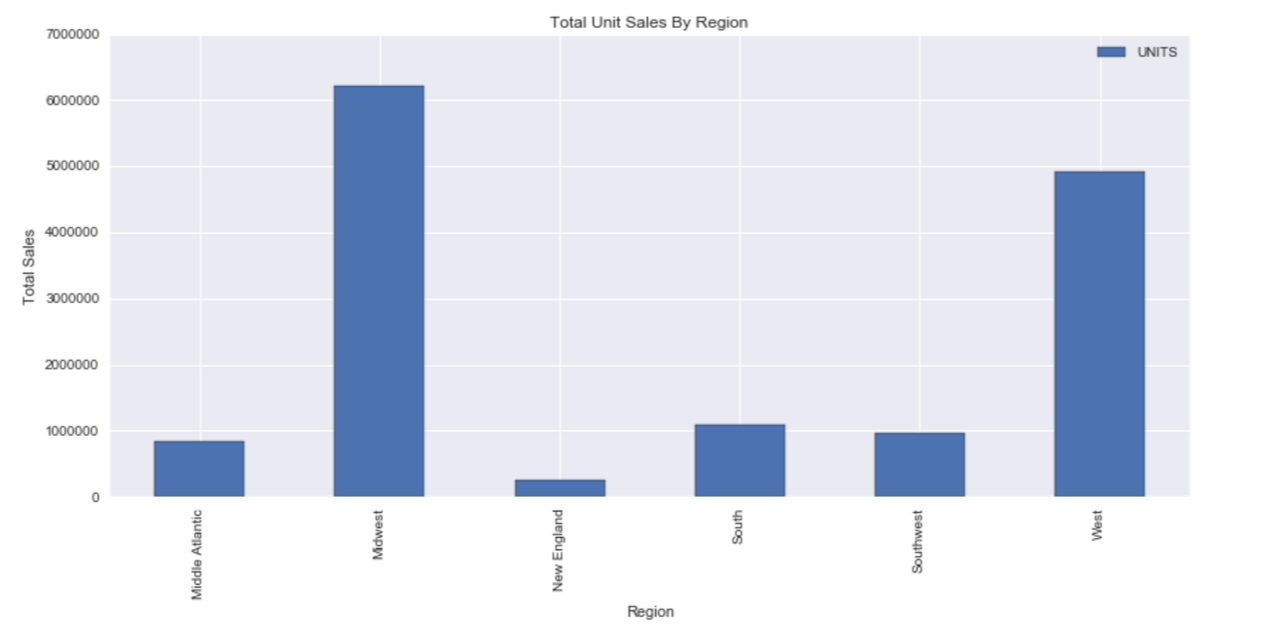


Figure . Total Unit Sales By Region



Figure Mean Unit Sales By Region

In the table for sales by region we can see that the top two spenders in yogurt are located within the west and Midwest. This is a total amount and could be due to a larger population in those areas. The mean unit sales figure shows on average how many sets of yogurt are sold across each region. The Midwest and the South come up as the top two mean sales locations. This information may be much more useful as population does not have an opportunity to skew the data.

Next we would like to know the per capita ratio of sales per person in each state we can divide the sales by state population. In the following table we graph the normalized values of sales per population (population provided was from US Census 2010).

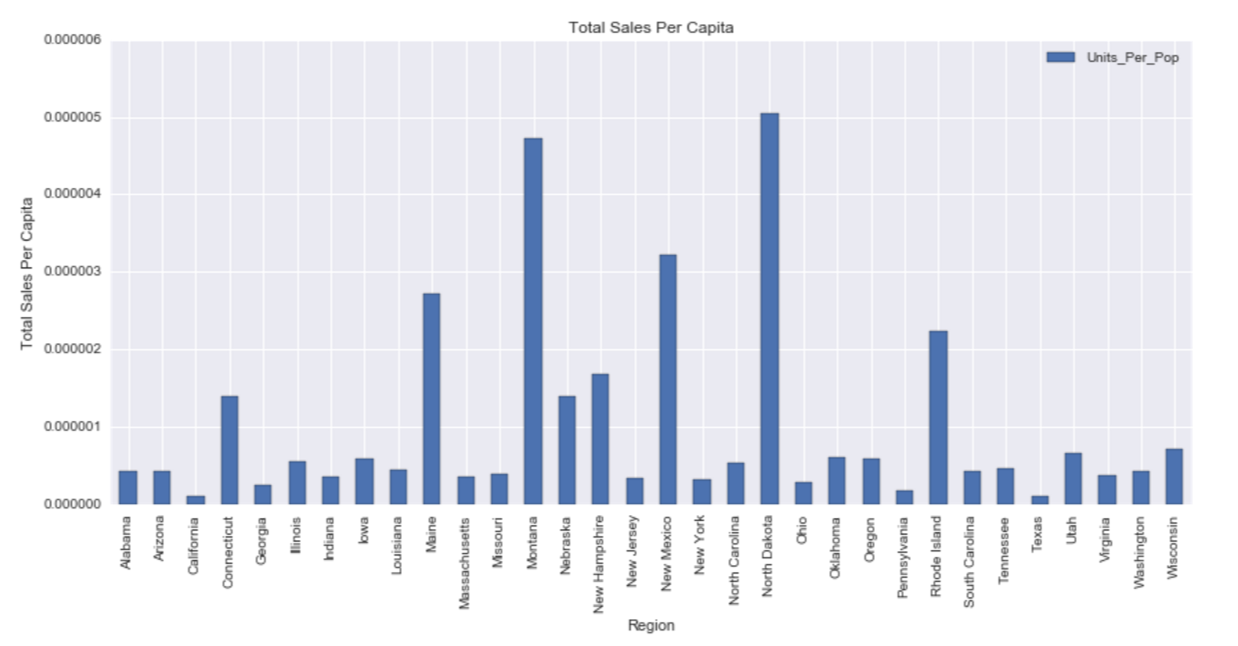


Figure . Total Sales Per Capita

In this figure we can see that North Dakota, Montana, New Hampshire, and Rhode Island come up for the highest amount of sales per population. There many be many issues with this data because little is known about the original distribution of stores that were sampled. The city or location does not directly correlate with the state, and therefore the data may be skewed significantly, but we chose to keep this graph in the report just for reference.

Next, average weekly sales for each state is computed. This figure shows the states with the highest distribution. This figure is much more useful than total sales per capita because it takes the mean instead of total sales. Here Illinois, New Mexico, and New York, and Tennessee come up as the highest consumers for yogurt. This is consistent with the average mean unit sales portrayed above.

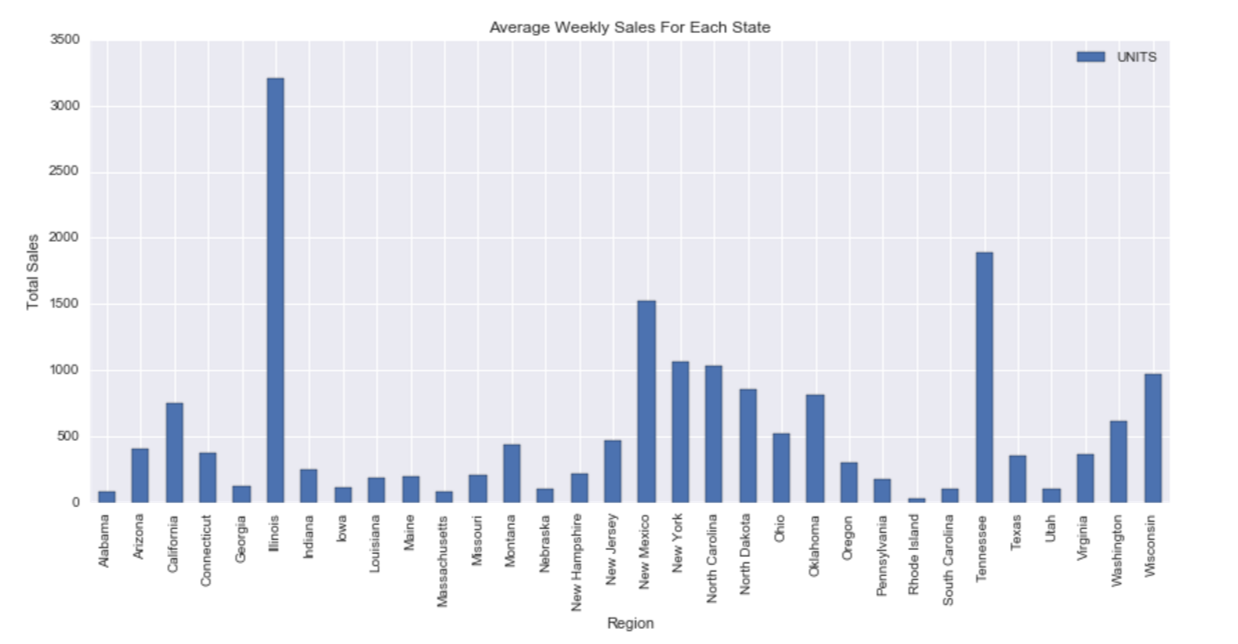


Figure . Average Weekly Sales

### Geographic Information for Grocery Stores

A similar method was developed for the grocery stores. Features were grouped by region, state, and linked with population. These features are used later in the model which predicts the number of sales using combined information from all of the tables. This is explained further in the feature selection chapter.

## Data Integration (merge)

Data integration was a key component to the overall structure of the data. Each table held keys which were used to join each table together. The store delivery table contained a column labeled “IRI Key” that was provided to join to the drug and grocery store information. Similarly, the panelist data was joined to demographic information was linked by “Panelist ID”. The parsed stub data was connected using the columns listed as “SY”, “GE”, “Vend”, and “Item”.

Outside data from the U.S. Census was obtained through the government census website. This information was used estimate population, and geographic regions for the geographic data. States were manually linked to the specific “Market Areas” listed in the store delivery table.

# Modeling

## Select Modeling Technique

Multiple machine learning models were used to address the specific questions posed in the business objectives section. Machine learning algorithms can be broken down in to supervised learning, and unsupervised learning; these models include Linear Regression, Logistic Regression, Decision Tree, SVM, Naive Bayes, KNN, K-Means, Random Forest. Our team utilized a majority of these algorithms in the development of all of our models.

Our models include three different subsets of classification models. These models are listed below.

1. PCA analysis on store data
2. Demographic clustering model for exploration (mentioned in exploration section)
3. Sales classification using random forest and linear SVC classification
4. Sales promotions predictions using random forests
5. Sales regression model based on lasso

The modeling assumptions are described in the following chapters.

## Modeling Assumptions and Feature Selection

### PCA Analysis on Store Data

Principal component was utilized in order to examine the weekly purchasing trends of people across a year for the drug stores. A similar model outcome is expected for the PCA run on the grocery store data set due to similar sales patterns. The first 3 dimensions have an extreme amount of variability and are hard to interpret. The PCA analysis resulted in the minimal data, so it was not used within the overall model.



Figure 17. PCA Analysis On Weekly Sales Data

### Demographic Clustering Model

The demographics data consists panelists whose purchases were tracked for one or more years. A subset of these purchases were categorized as yogurt and the associated panelist information was also provided. This include nominal features relating to each panelist’s household.

Prep and Cleaning Steps for Random Forest Classification:

1. Combine files into single data object for sales and panelists
2. Identify samples with missing features
3. Years 1 and 2 had many features missing, so they were excluded from this analysis. Samples with any missing features or NaN were dropped.
4. Panelist information was originally standardized. These values were mapped back to the original nominal descriptors. The nominal features were then converted to binary features.
5. Panelist demographics were joined to the panelist sales information.
6. This new data structure was then joined to product feature data.
7. An analysis was done to verify that every purchase included at least one yogurt product. The analysis showed that 30% of the data in the yogurt category had no yogurt purchases.
8. A distinct list of panelists was then created and labeled to show who purchased yogurt and who did not. This data was randomly sampled to produce a label-balanced set of classification data.
9. A random forest classifier was produced to predict these labels based on available features.

Random Forest Classification Model and Methodology:

The classification data was split into the following parts:

* Train – 50%
* Test – 25%
* Validate – 25%

Train is used to train the initial model. Validate was used to tune the hyper-parameters of the random forest classifier. Test was to sore the initial model. The final model was validated using a 10k Cross Validation.

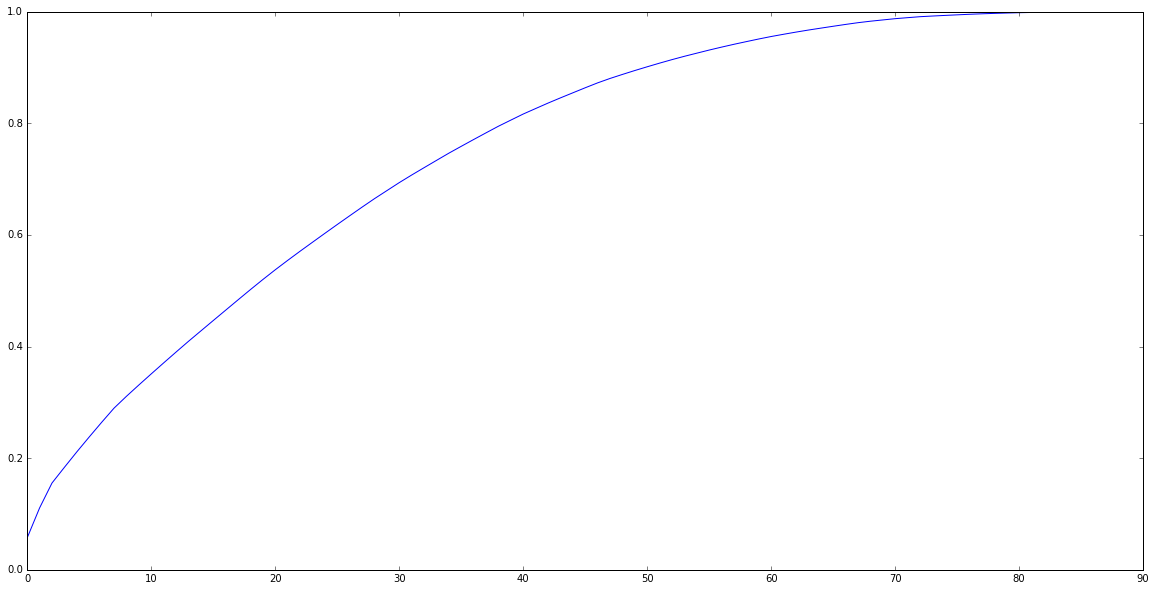
Results:

* Initial Model Score: 73%
* Tuned Model Score: 75%
* Cross Validation score: 58%

Detailed results of this analysis can be found in the ipython notebook: Final - Demo Explore

The cross validated classifier was used to rank demographic feature importance.

Feature Importance Graph:



Top 20 Features:

[('ALL\_TVS', 0.057266909496007205),

('CABL\_TVS', 0.05364124104430728),

('Family Size', 0.04462171105259242),

('Number of Dogs', 0.028103212437407103),

('Number of Cats', 0.02765960246692521),

('Combined Pre-Tax Income of HH\_$25,000 to $34,999 per yr',

0.026717754686462737),

('Education Level Reached by Female HH\_Some high school',

0.026181756998910476),

('Education Level Reached by Female HH\_Graduated high school',

0.02545415055017671),

('Combined Pre-Tax Income of HH\_$55,000 to $64,999 per yr',

0.02121184139959722),

('Age Group Applied to Female HH\_65 +', 0.020261691254809176),

('Combined Pre-Tax Income of HH\_$35,000 to $44,999 per yr',

0.01989990221268738),

('Female Working Hour Code\_Full time, > 35 hrs./wk.', 0.019672635578323668),

('Education Level Reached by Female HH\_Technical school',

0.019436590608864088),

('Education Level Reached by Male HH\_Graduated high school',

0.019214332476458883),

('Combined Pre-Tax Income of HH\_$20,000 to $24,999 per yr',

0.018651100542961273),

('Female Working Hour Code\_Not employed', 0.01863362722624224),

('Male Working Hour Code\_Part time, < 35 hrs./wk.', 0.018549213122606528),

('Type of Residential Possession\_Owner', 0.018534346574399636),

('Type of Residential Possession\_Renter', 0.018398409305321033),

('Education Level Reached by Male HH\_Some high school', 0.018005895684901674)]

### Sales Classification Model (Sedat)

CL

Pick 10000 random rows from grocery store data

Sales before 2007 were stable

Creating dummies in Products data

Merge grocery store and products data

Calculate per unit price

Set UNITS and UNIT\_PRICE so that they are 1 if greater than median, else 0

Classification

Multiple models were used

We will pick population data based on geographic location

While Random Forest provides better accuracy, SVC is more stable in 10-fold cross validation.

SVC also needs less features (using Lasso) than Random Forest. In real world, SVC would be more practical as we will ask a limited number of questions.

### Sales Promotions Model (Megan)

## Model Tuning

### Cross Validation

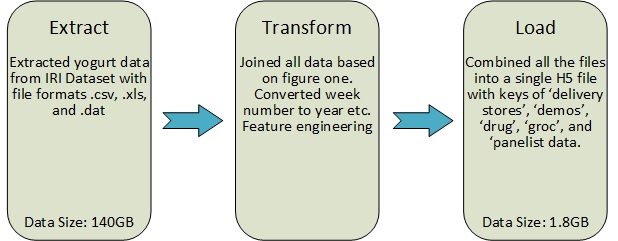
Figure 18. ETL Process DiagramFigure 18. ETL Process Diagram

Figure . ETL Process Diagram

Figure 18ETL 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)

# Evaluation

## Review Process

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.

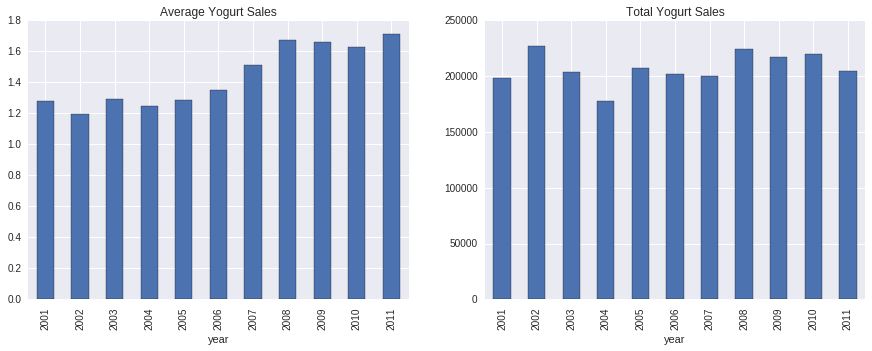
****

Figure . Average Yogurt Sales

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 . Monthly, Daily, Weekly Yogurt Sales

**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.

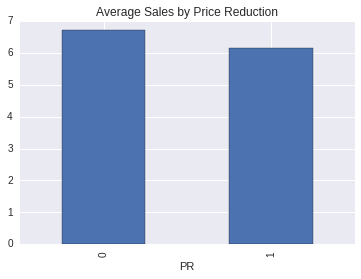
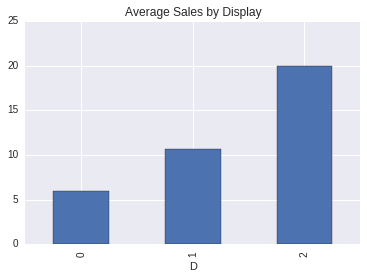


Figure . Sales Promotions at Drug Stores

    **Fig. Sales Promotions at Drug Stores**

## Alignment of Results with Objectives

## Next Steps

# Deployment

## Business Use

# Conclusion