Analysis of Yogurt Sales from IRI Marketing Dataset

DSE 220 Final Project

Megan McCarty
Julius Remigio
Syed Nazrul
Ryan Riopelle
Michael Galarnyk

IRI Dataset Reference:
Bronnenberg, Bart J., Michael W. Kruger, Carl F. Mela. 2008. Database paper: The IRI marketing data set. ***Marketing Science***, **27**(4) 745-748.

Executive Summary

Methodology

This project employed the Cross Industry Standard Process for Data Mining, commonly known as CRISP-DM. There are six major phases although the sequence of the phases is not strict and moving back and forth between different phases is always required. These phases and the sequence are described in the diagram below.

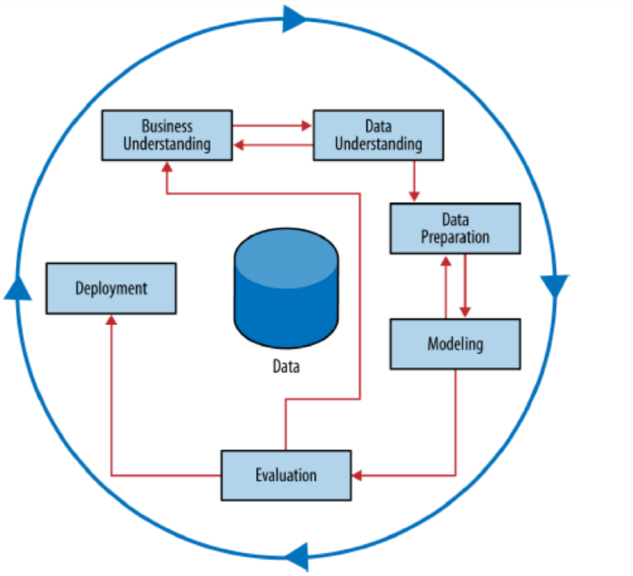


Figure 1. Schematic Illustration of CRISP-DM Methodology

Business Understanding

The initial phase of the project focused on understanding the project objective and requirements from a business perspective, and then converting that into a data mining problem and an initial plan to achieve our objectives.

Objectives

The primary objective was to analyze the IRI Marketing Dataset in an effort to identify trends and influential features in the sale of yogurt. Then to develop the capability for proactive evaluation of marketing resources through the use of predictive modeling. After review of the data and resources, the team decided to focus on developing four 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 decision by understanding geographic sales trends
4. Predictions of future sales to better inform business decisions

Access Situation

The IRI Marketing dataset contains store sales and consumer panel data for 30 product categories. It is used to enable modeling of important economic and marketing metrics for companies and policy makers. For this project, the team decided to focus on yogurt sales. As a result, we had the following data types at our disposal:

- Drug and Grocery Store Weekly Yogurt Sales Data
- Consumer Panelist Transaction Data
- Consumer Panelist Demographic Data
- Yogurt Product Attribute Data
- Store Attribute Data

While available, the team declined the use of a super computer and instead decided to manage the above data via HDF5 files. Hierarchical Data Format Files (HDF5) allowed the team the enjoy the benefits of a database without requiring external server resources. It is a very mobile format which facilitated greater collaboration. For the analysis, the team employed machine learning modules available in Python.

The team identified the following risks going into the project:

- The size of the grocery store weekly yogurt sales data was too large for personal laptop RAMs. The team used a representative sample.
- Not all keys between data tables linked. Team performed the analysis on the subset of the data that could merged and accepted the potential information loss as a small risk.
- An underlying risk of the whole project was that the IRI data may not be an accurate representation of the markets for which data is provided. Thus we accepted that any conclusions are skewed by the data were are working with.
- The python machine learning module Scikit-Learn isn't capable of handling categorical data naturally. The team accepted any risk associated with the binary encoding of categorical data.

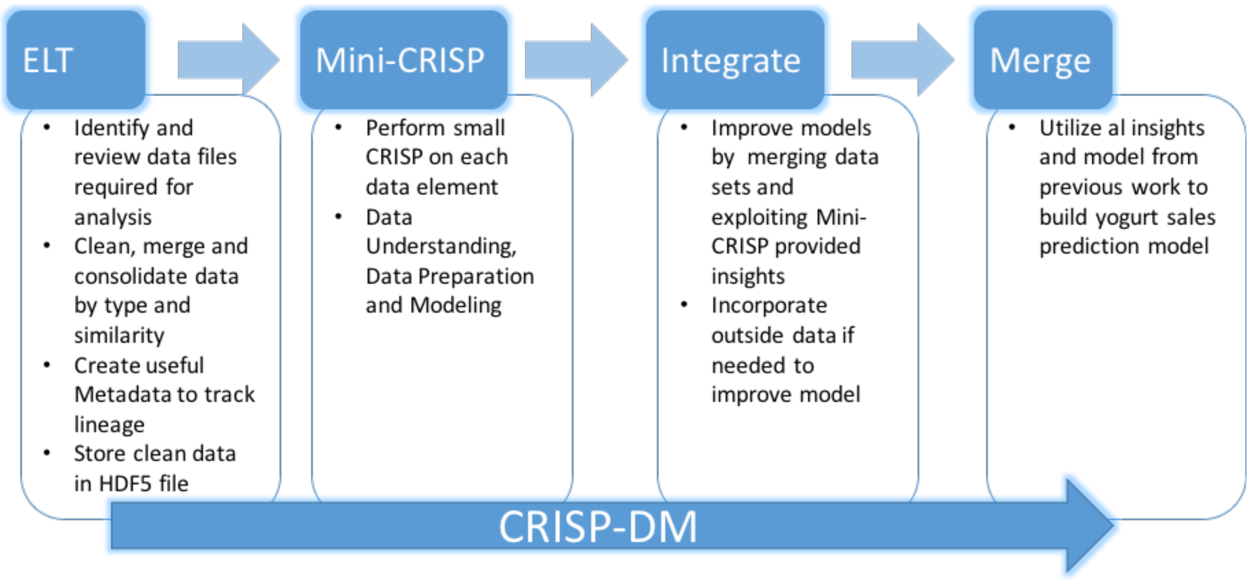
Data Mining Goals

The primary goal for data mining is to discover underlying trends in the data that can be used to build a predicative model. To obtain an in-depth understanding of all the data types and how they work together, the team focused on five specific areas.

- Sales Promotions – Determine wither sales promotions have a statistically significant impact on sales
- Panelists – Identify demographics that contribute to higher or lower yogurt sales
- Products – identify any yogurt attributes that tend to sell better or worse
- Stores – Determine if there exist geographic trends in the sales data
- Time Series – Determine if there exists daily, weekly, or yearly sales patterns

Project Plan

As mentioned previously, the team followed the CRISP-DM process for this project. The initial project plan was to examine each data element individually, integrate them to improve understanding and then merge all the elements to build a sales model. Each individual data element was worked following the CRISP-DM method (referring to this stage as "Mini-CRISP" in the figure below.) Figure 2 gives a visual representation of the project plan.



ELT

- Identify and review data files required for analysis
- Clean, merge and consolidate data by type and similarity
- Create useful Metadata to track lineage
- Store clean data in HDF5 file

Mini-CRISP

- Perform small CRISP on each data element
- Data Understanding, Data Preparation and Modeling

Integrate

- Improve models by merging data sets and exploiting Mini-CRISP provided insights
- Incorporate outside data if needed to improve model

Merge

- Utilize al insights and model from previous work to build yogurt sales prediction model

CRISP-DM

Figure 2. Project Plan

Data Understanding

Initial Data Collection

This project examined yogurt consumer trend data obtained from IRI Worldwide, a company who provides clients with consumer data, shopping data and retail market intelligence and analysis focused on the consumer packaged goods. The IRI dataset contains 11 years of weekly sales data from grocery, drug and mass stores for 30 consumer product categories in 47 markets. Files pertaining to yogurt sales was scattered across the multiple year and product folders of the IRI Dataset. The structure shown in Figure 3, displays the various files types found in the dataset and how the relate to each other.

The major initial data collection challenges were similar files distributed over various folders, different versions of similar data and different file formats and headers. The data pertaining to yogurt sales was extracted from the dataset, translated and assembled into a Hierarchical Data Format (HDF5) File using Linux commands, Bash and Python. Analysts was able to extract 3,649 files, totally 140GB and combine them into a single 1.8GB HDF5 file.

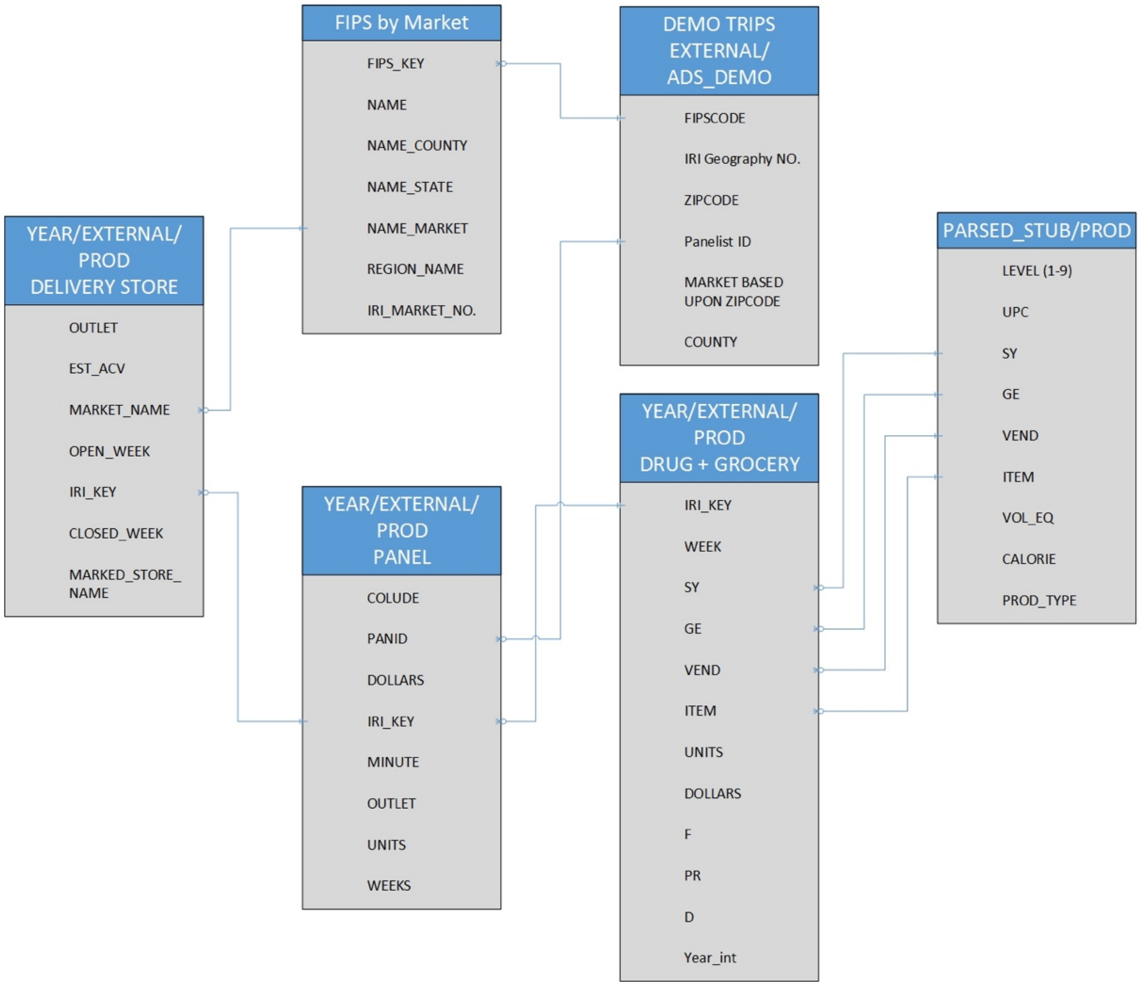


Figure 3. Modeling the relationships between the various data sources

Data Description

The HDF5 file housing all of the yogurt data contained five tables. Figure 4 below gives a brief description of each of these tables.

Yogurt Data Table	HDF5 key	Description
Drug Store Transaction	drug	Total weekly sales of yogurt items purchased from drug stores. Each record represented an item and a store.
Grocery Store Transactions	groc	Total weekly sales of yogurt items purchased from grocery stores. Each record represented an item and a store.
Store Attributes	delivery_stores	Information about each of the stores reporting who contributed their transactions to the dataset.

Consumer Panelist Transactions	panel	Transactions made by consumer panelists.
Consumer Panelist Demographic Data	demo	Demographic data for each of the consumer panelists
Product Attributes	ptype	The attributes of each of the yogurt items

Figure 4. Yogurt Data Tables Description

Data Exploration

Sales Promotions

One of the core objectives of this project was to build a model that predicted the influence of sales promotions. During initial data exploration, it was noted that sales appear to be higher in the presence of promotions. The first step was to determine wither they are actually higher or due to random chance.

Analysts defined promotion success as any single or combination of price reduction, advertisements or displays that result in weekly sales that is more than one standard deviation away from the mean. A mean and standard deviation was calculated for each store and product pair to enable direct comparison. A new binary field was created that served to label the record as having a successful promotion or not. The table below displays the percentage of successful weeks for each type of promotions. And while not always successful it does show that there is a trend.

Sales Promotions	Percent of Successful Weeks
Large Advertisement	36%
Medium Advertisement	33%
Advertisement With Retailer Coupon	50%
Any Advertisement	38%
Minor Display	31%
Major Display	47%
Any Display	43%
Price Reduction	29%

Figure 5. Successful Sales Promotions

Panelists Demographics

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. The K-Means technique was used to do clustering on this data in an unsupervised manner. Figure 6 below shows the elbow curve.

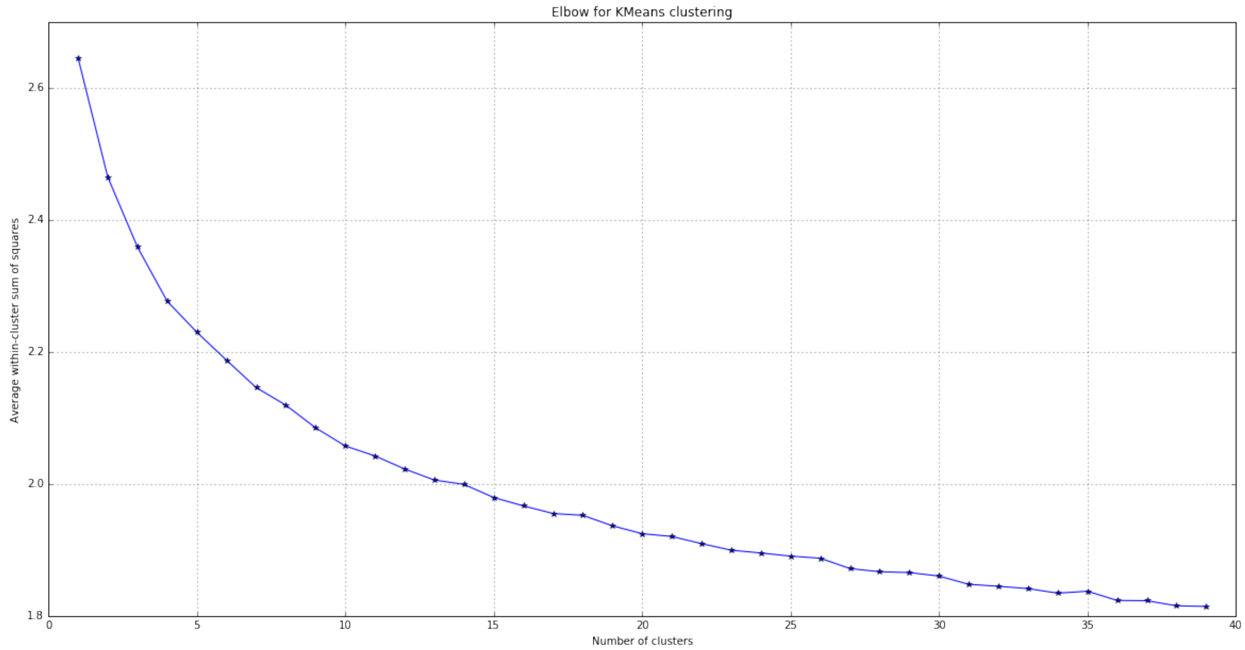


Figure 6. Elbow Curve for K-Means

Analysts opted to use K-Means with 12 clusters after examining the above elbow curve. Figure 7 contains a sample of one of these clusters. Not surprising, given the elbow curve, the team was unable to achieve stabile clusters.

```
Cluster: 0 Count: 18857
ALL_TVS            2.0
CABL_TVS           2.0
Family Size        2.0
Number of Cats     0.0
Number of Dogs     0.0
Name: mean, dtype: float64

Age Group Applied to Female HH                                   65 +
Age Group Applied to Male HH                           No such person
Children Group Code                         Family size>0 yet no children
Combined Pre-Tax Income of HH                  $25,000 to $34,999 per yr
Education Level Reached by Female HH                 Some high school
Education Level Reached by Male HH                                N/A
Female Working Hour Code                         Full time, > 35 hrs./wk.
Male Working Hour Code                           Part time, < 35 hrs./wk.
Marital Status                                                 Single
RACE3                                                           White
Type of Residential Possession                                 Renter
```

Figure 7. Sample Cluster

Panelist Transactions

Analysts examined the consumer panelist transaction over time. Figure 8 shows that the average and total yogurt sales for panelists remains relatively unchanged.

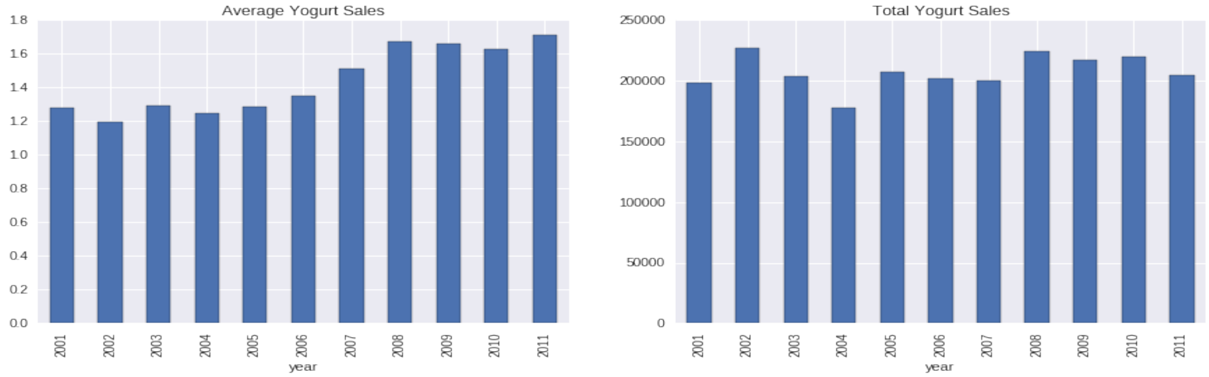


Figure 8. Average and Total Yogurt Sales

Analysts also examined panelists transactions by month, week and hour of the day. Only records between 2008 and 2012 contained the MINUTE field, thus those were the only records included in this exploration. Figure 9 shows the results. While slight changes, yogurt sales appear consistent between months and weeks. There is variability in sales throughout the day, but for the most part, they seem to occur during standard business hours.

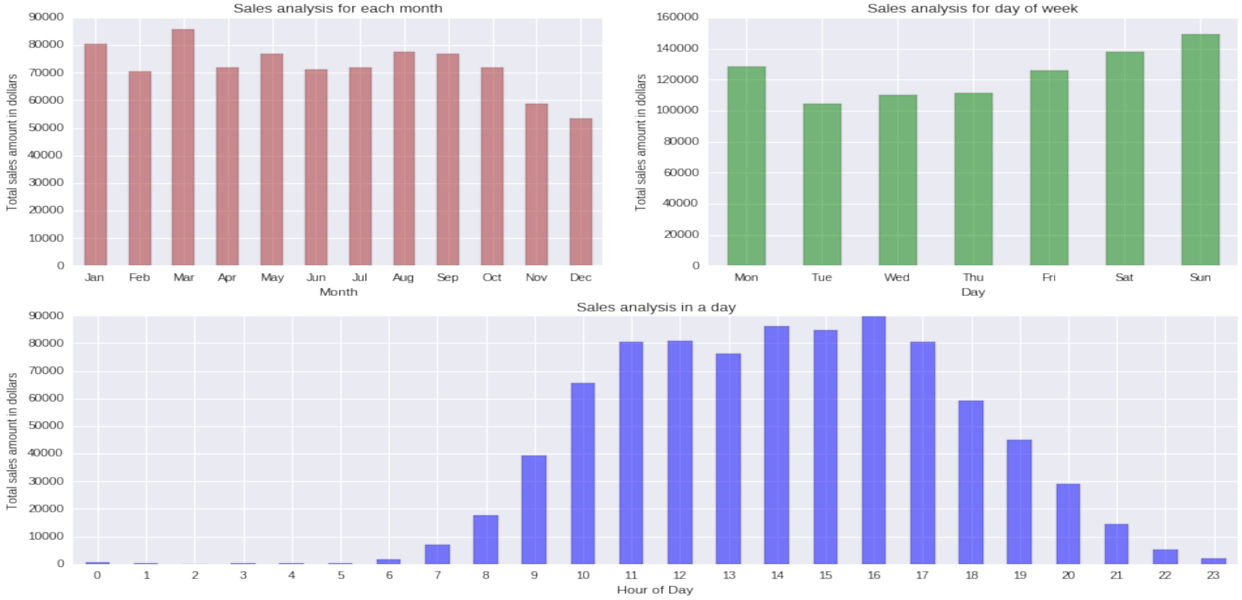


Figure 9. Monthly, Weekly & Hourly Yogurt Sales

Store Sales Over Time

Analysts examined the drug store and grocery store yogurt sale transactions over time. For brevity, only the results for drug stores are shown. Grocery stores have the same general trends and their plots can be viewed in the Python notebooks accompanying this report. Figure 10 displays the mean sales by year.

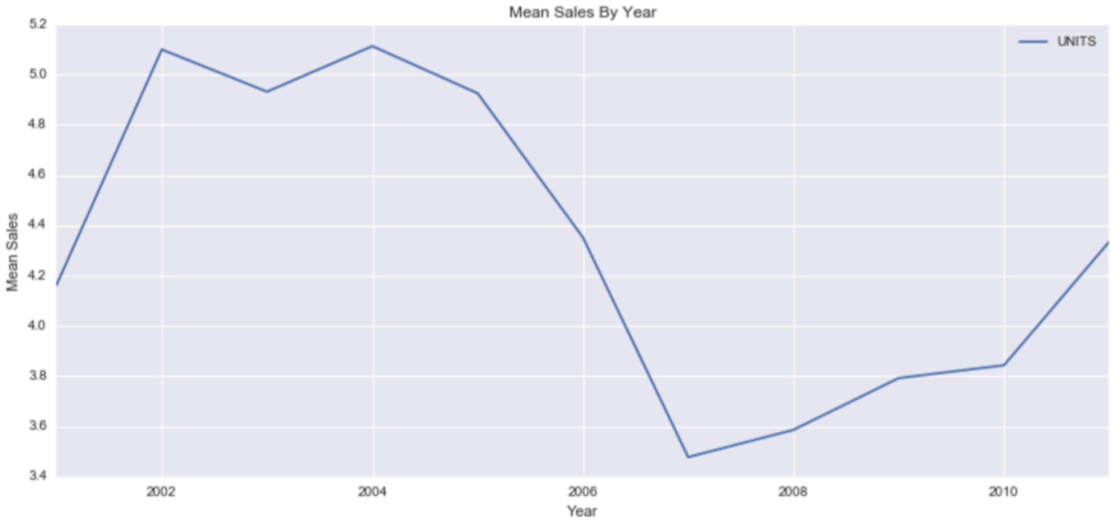


Figure 10. Mean Sales by Year

Yogurt Sales peaked between 2002 and 2004, then decline between 2005 to 2007. The cause of the decline is unknown. The team cross referenced the means with the available data points to ensure this trend was not the result of decrease in data points.

Mean yogurt sales by month are displayed in Figure 11. The sales are fairly consistent across all months with sales experiencing a low during October through December (months 10-12). Analysts theorized that the decline in sales was due to reduced incentive to buy yogurt during cold winter months as consumers may be more interested in warmer comfort foods.

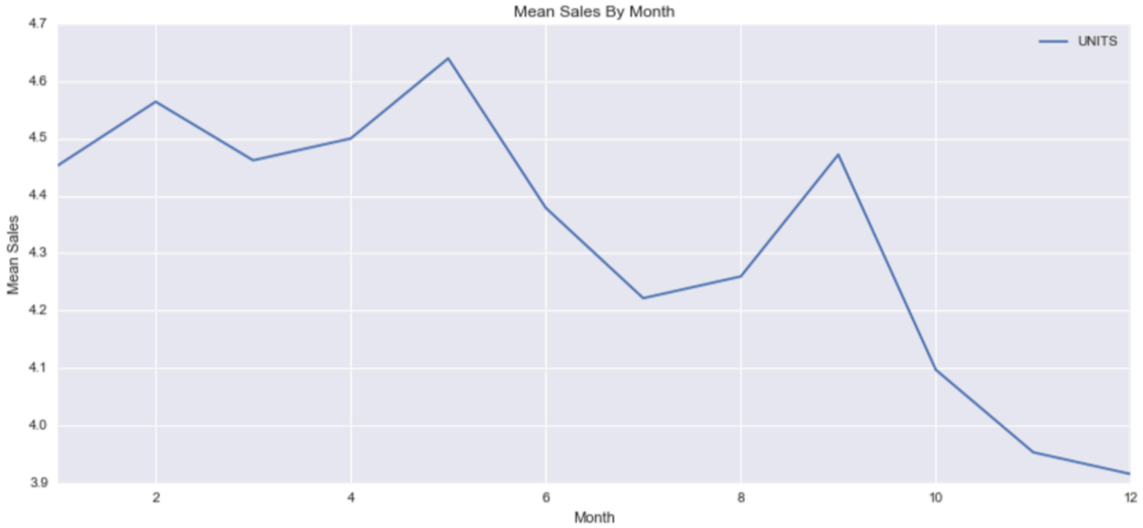


Figure 11. Mean Sales by Month

Mean yogurt sales by week are displayed in Figure 12. One can see from this graph that there is a fair amount of noise in the data; which isn't surprising as it contains multiple stores and yogurt items. The analysts attempt to use PCA for noise reduction will be presented later in the Feature Engineering section.

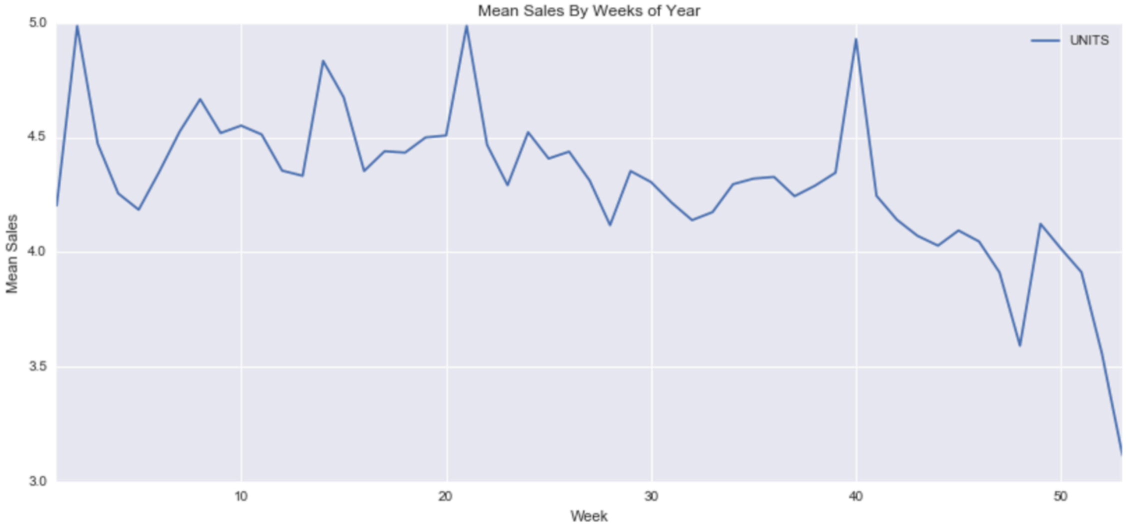


Figure 12. Mean Sales by Week

Geography

The store attribute table was merged with the drug store transaction table in an effort to explore how yogurt sales vary over geographic regions. Figure 13 contains a graph of the mean sales by state. The 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.

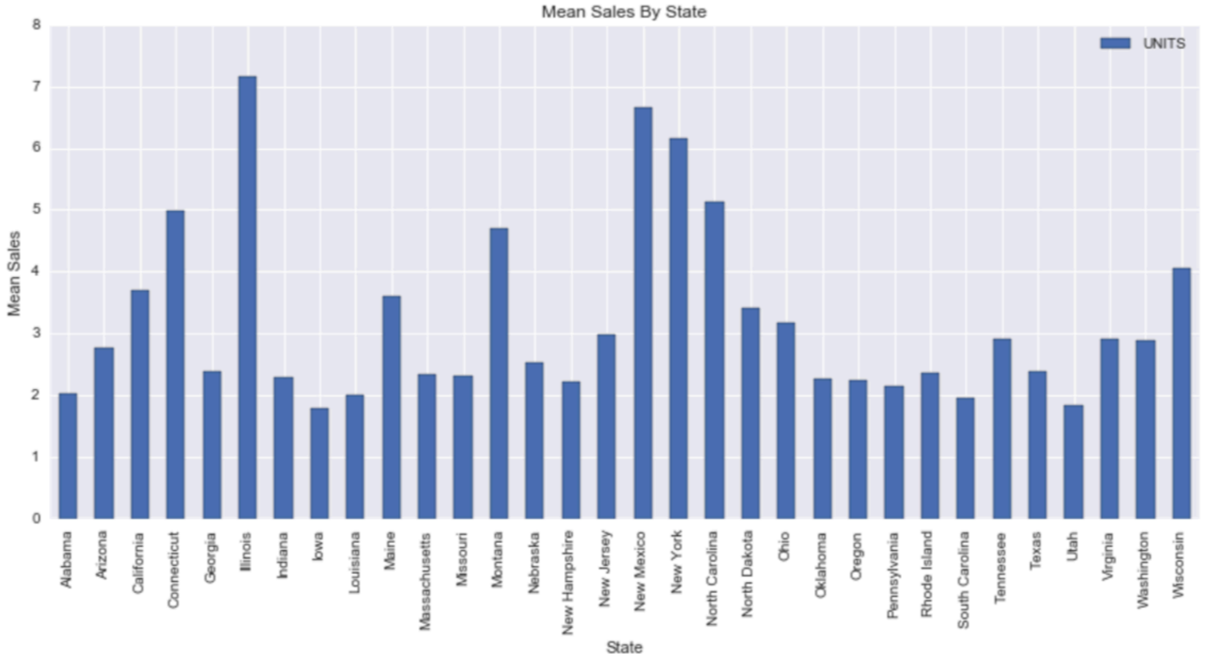


Figure 13. Mean Sales by State

Next, analysts chose to group the yogurt sales by U.S. regions (as defined by the census). Figure 14 contains the graph of means sales by region. The West and Midwest are the top yogurt consumers.

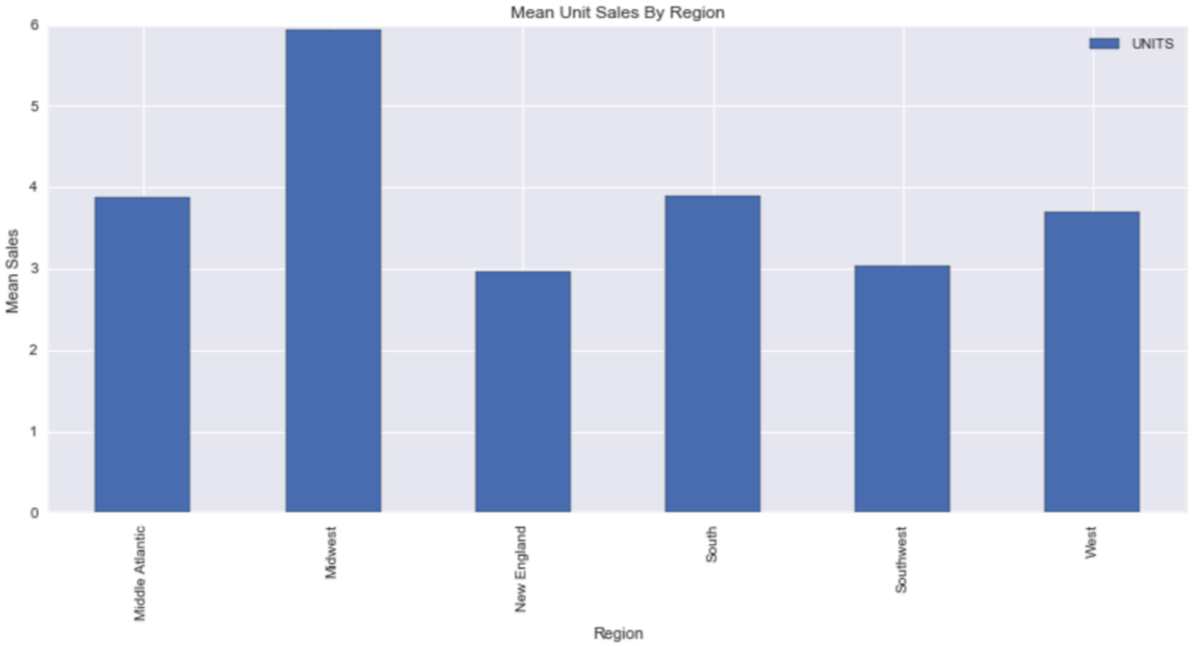


Figure 14. Mean Sales by Region

Analysts realized that the trends previously viewed may be skewed by population. The team decided to explore wither the higher sales were simply due to higher population and not because people in those states had an above average love of yogurt. Data from the US Census in 2010 was obtained and merged with the sales data. This was done to calculate yogurt sales per capita. The team recognized that populations change over time and that a store's state population may not directly relate to their consumer base. But decided that using 2010 state populations as a snapshot was sufficient for this coarse data exploration. Figure 15 contains the results.

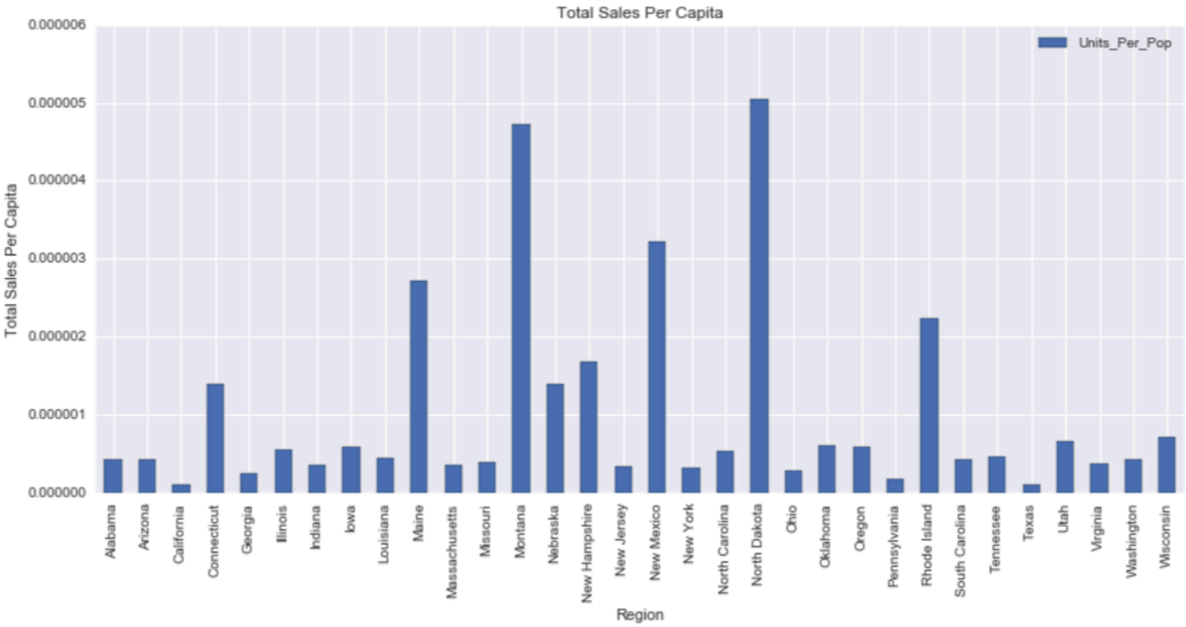


Figure 15. Yogurt Sales Per Capita

From the figure it appears the North Dakota, Montana, New Hampshire and Rhode Island are the top yogurt lovers in America. As previously noted, there may be issues with our single year state population approach. Thus we do not draw any conclusions from this graph but it does provide strong evidence that store location affects yogurt sales.

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.

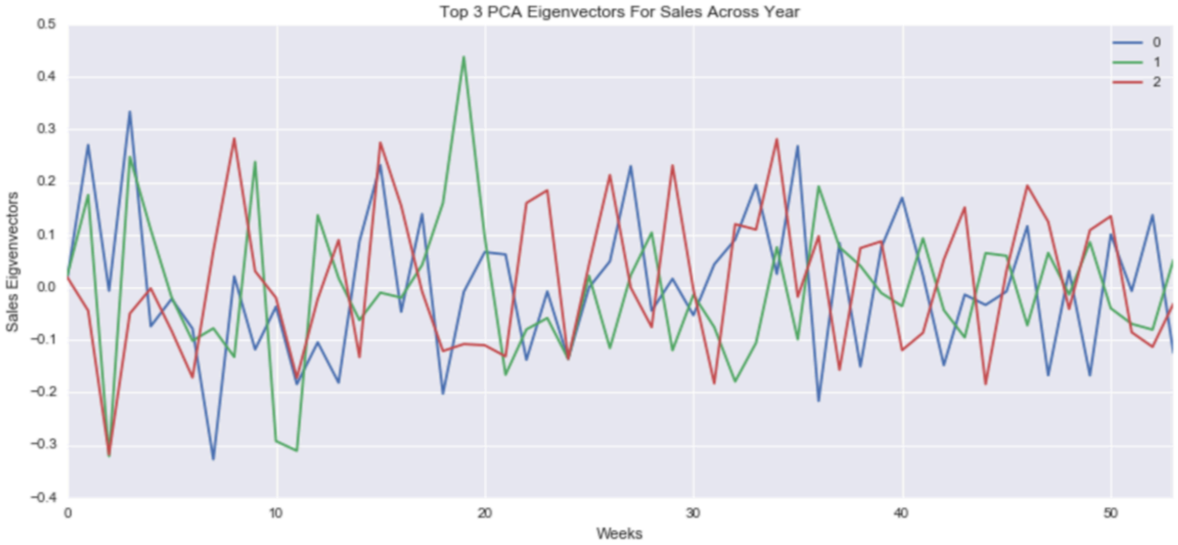


Figure 16. PCA Analysis on Weekly Sales Data

Products

Analysts examined the drug store and grocery store transactions by yogurt attributes. For brevity, only those attributes that appear to have discerning characteristics are discussed here. Figure 16 shows the mean yogurt sales by type, style, category (L2) and brand (L3). These are shown because each appear to have a class whose mean sales surpasses the rest. Analysts were hesitant to draw any direct conclusions from this data as there is no information regarding the survived stores inventory. These trends could simply be a result of limited supply of alternatives. However, the team did believe these trends provide strong motivation for product attribute inclusion in any sales predication models. This will be discussed later in the Feature Selection section.

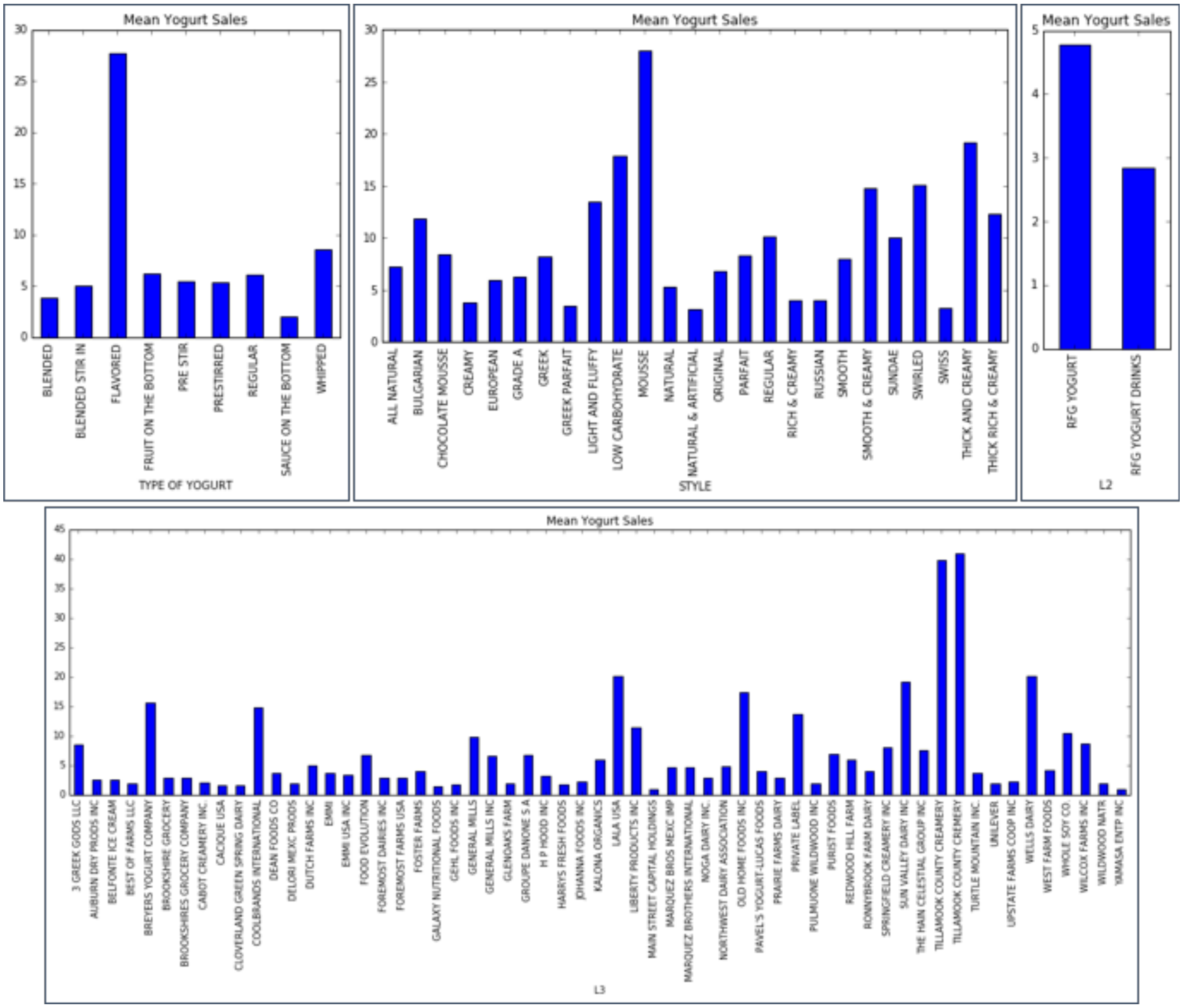


Figure 17. Yogurt Sales by Attributes

Data Preparation and Modeling

The team produced three models, referred to here as Demographics Model, Sales Promotions Model and Sales Model. In this section we will step through the data preparation and model creation for each one.

Demographics Model

For every panelist whose purchases were tracked for one or more years there is a record in the demographics data table. This table contains nominal features relating to each panelist's household. It was discovered that despite IRI dataset indications to the contrary, not all panelist identified as yogurt consumers actually purchased yogurt. Analysts used that fact to our advantage and constructed a model that predicts whether a consumer purchases yogurt based on their demographic information.

The following are the steps taken to prepare the necessary data for modeling:

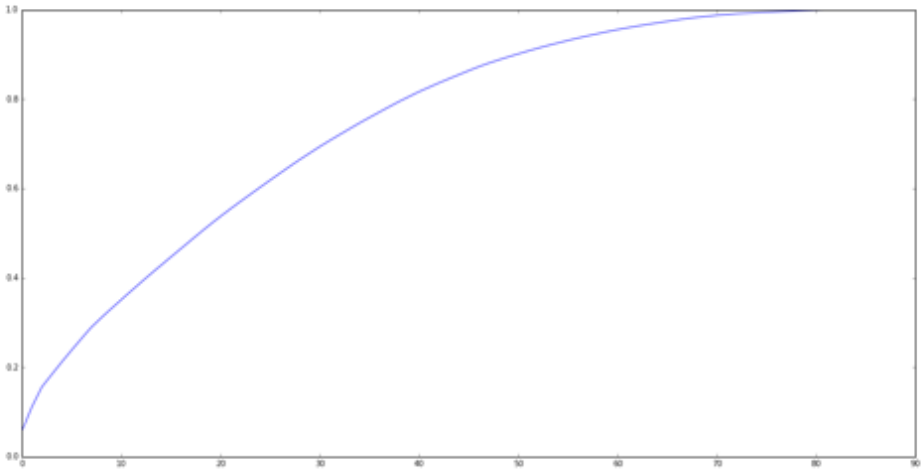
1. Merge the Panelist demographic table and the Panelist sales transaction table into a single data object. This allowed analysts to view the associated demographics of the person performing each transaction.
2. Several demographic features were missing data in years 2001 and 2002, so those years were excluded from analysis. Any other samples with several missing demographic features or NANs were also excluded.
3. The nominal demographic features were binary encoded to accommodate the skikit-learn machine learning functions.
4. Product Attributes table was joined to the data object. This allowed analysts to view the type of product purchased in each transaction.
5. An analysis was performed to verify that every transaction included at least one yogurt product. Despite the fact that these transactions were all contained within yogurt category of the IRI dataset, the analysis revealed that 30% of the transactions contained no yogurt purchase.
6. A distinct list of panelist was created and labeled to show who purchased yogurt and who did not.
7. The distinct list was randomly sampled to produce a label-balanced set of classification data.

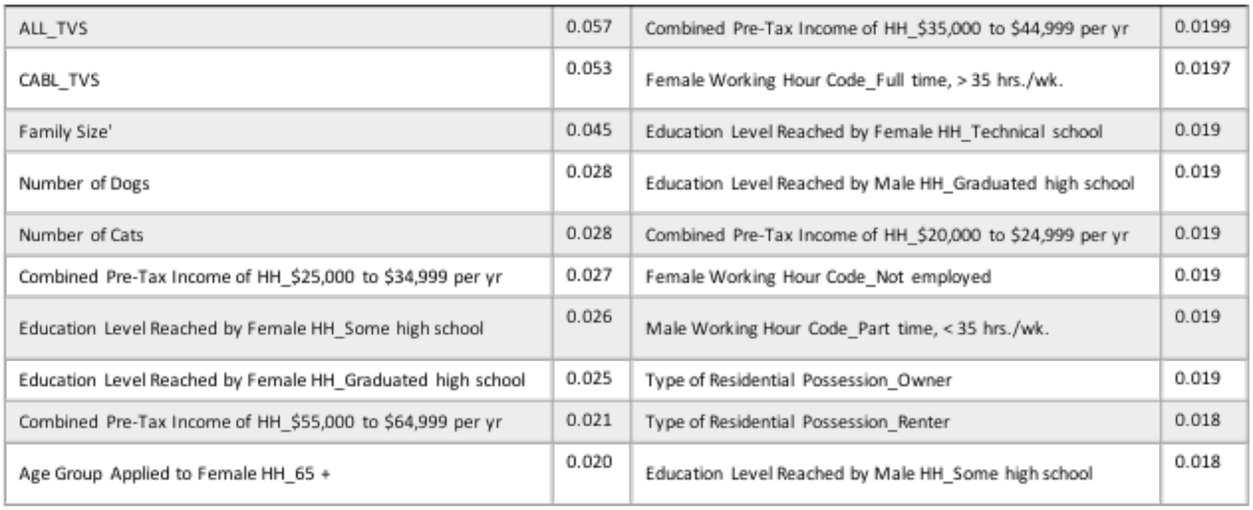
A Random Forrest Classification model was built to predict wither a panelist was a yogurt consumer or not based on their household demographics. The classification data was split into train (50%), test (25%) and validate (25%) sets. The training set was used to train the initial model. The validate set was used to tune the hyper-parameters of the Random Forrest Classifier. The test set was used to score the initial model. The final model was validated using 10-Fold Cross Validation. Results of the model were:

- Initial Model Score: 73%
- Tuned Model Score: 75%
- 10-Fold Cross Validation Score: 58%

Please note that detailed results of this model can be found in the notebook Final – Demo Explore.

The cross validated classifier was used to rank demographic feature performance. Figure 17 shows the feature importance graph and the top 20 features.





ALL_TVS	0.057	Combined Pre-Tax Income of HH_$35,000 to $44,999 per yr	0.0199
CABL_TVS	0.053	Female Working Hour Code_Full time, > 35 hrs./wk.	0.0197
Family Size'	0.045	Education Level Reached by Female HH_Technical school	0.019
Number of Dogs	0.028	Education Level Reached by Male HH_Graduated high school	0.019
Number of Cats	0.028	Combined Pre-Tax Income of HH_$20,000 to $24,999 per yr	0.019
Combined Pre-Tax Income of HH_$25,000 to $34,999 per yr	0.027	Female Working Hour Code_Not employed	0.019
Education Level Reached by Female HH_Some high school	0.026	Male Working Hour Code_Part time, < 35 hrs./wk.	0.019
Education Level Reached by Female HH_Graduated high school	0.025	Type of Residential Possession_Owner	0.019
Combined Pre-Tax Income of HH_$55,000 to $64,999 per yr	0.021	Type of Residential Possession_Renter	0.018
Age Group Applied to Female HH_65 +	0.020	Education Level Reached by Male HH_Some high school	0.018

Figure 18. Demographic Model Top 20 Features

Sales Promotions Model

One of the core objectives of this project was to build a model that predicted the influence of sales promotions. Earlier in this paper, we explained how analysts verified the existence of successful sales promotions, i.e. ones that caused a statistically significant increase in sales. In this section we will walk through how the model was constructed.

The following are the steps taken to prepare the necessary data for modeling. It should be noted that multiple steps below prepare data for modeling that ultimately were discarded during feature selection.

1) Drug store and grocery store data was concatenated. This gave us one data object containing all yogurt sales transactions. Two new binary features was created that identified each record as originally belonging to drug stores or grocery stores.
2) Grouped by Store and Product to obtain the mean and standard deviation for weekly units sold by each store for each product. Then created a new feature which will serve as the classification label. It is binary and indicates wither the weekly unit sales is more than one standard deviation away from the mean.

3) Created two additional binary features that identify when any advertisement or any display is used, respectively. Converted advertisement and display categorical features into a series of binary encoded features.
4) Imported the modified version of the IRI provided Week Converter and merged it with the data object. This provided additional features regarding time, i.e. Year, Month and Week of the Year.
5) Added additional feature price which is the per unit price of the yogurt item in the transaction.
6) Merged table with Product Attributes table to obtain additional features regarding the yogurt items. Tables were merged using an inner join. This created a subset of only those records whose UPC code linked with a row in the Products table. Also only included those product attributes who had at least 50% non-missing data.
7) Merged table with Store Attributes to obtain the State feature for each store.

A Random Forrest Classification model was built to predict if a promotion was successful or not. Recall that we define a successful promotion as one where that week's unit sales is more than one standard deviation away from the mean. Analyst built the model in stages; first with only promotion features and then adding in new feature incrementally. Below are the results of each attempt.

Model Features	Initial Score	10-Fold Cross Validation Score
Sales Promotions	79.17%	79.05% (+/- 0.3%)
Added drug/grocery	79.21%	79.08% (+/- 0.3%)
Added price	78.64%	77.50% (+/- 1.5%)
Added Product Attributes	78.24%	50.63% (+/- 22.0%)
Added store identification key	78.52%	74.65% (+/- 5.7%)
Replaced store key with state	78.63%	77.44% (+/- 2.0%)

Figure 19. Random Forrest Iterative Features Results

Analysts tuned parameters using grid search and then did final 10-fold cross validation. The final model achieved 78% accuracy.

It should be noted that additional feature selection on the product attribute table was attempted. A Random Forrest model was built to identify the most important features. Unfortunately, because of the binary encoding required by sklearn the results proved difficult to interpret.

Sales Model

Per the project plan, analysts attempted to incorporate influential data entities into a single sales prediction model. Two model were attempted; a regression model and a classification model.

Regression

This model attempted to predict the weekly unit sales of yogurt. Potential features for this model were all the attributes our analysis to this point had identified as influential. More specifically, sales promotions, product attributes, month, year, state, price and outlet. The goal was to accurately predict the unit sales of yogurt. Analysts attempted both Lasso and Ridge regression.

Unfortunately, both attempts were unsuccessful as they returned 32.76% and 43.84% accuracy, respectively. Support Vector Regression was also attempted but it required too much computational time to be a viable option.

Classification

The goal is to predict wither or not future weekly sales will be above the national median. For this model, analysts attempted to use sales promotions, geographic, product and demographic data for prediction.

The following are the steps taken to prepare the necessary data for modeling.

1) In the sales transaction data, the sales promotion field referring to advertisements was dropped. The categorical display field was converted to a binary field indicating if any display was present or not. The price reduction field was maintained.
2) Table was merged with Product Attribute table. The categorical attribute fields were converted to a series of binary fields
3) Unit price was calculated (formula was Dollars/Unit)
4) Table was merged with store attribute to obtain the state feature. Using that field, state population and median income were added to the table. The median income field was converted to binary indicating income was above or below general median. In data exploration, it was theorized that densely populated cities tend to have higher average yogurt sales for each store and that low income groups are more likely to buy yogurt.
5) Created a new binary field that indicates wither the sales were above or below the median. This field will serve as our classification label.

Analysts employed a Lasso model to identify 34 out of the resulting 518 features as influential. A Random Forrest Classification and Linear SVC model were used to build the model. Figures 19 and 20 shows the individual model score and the results of their 10-fold Cross Validation. While Random Forrest achieved a higher score, cross validation revealed that Linear SVC is more stable.

Model	Score	10 Fold Cross Validation
Random Forrest Classification	80% +	Unstable
Linear SVC	72% +	Stable

Figure 20. Sales Prediction Model Results

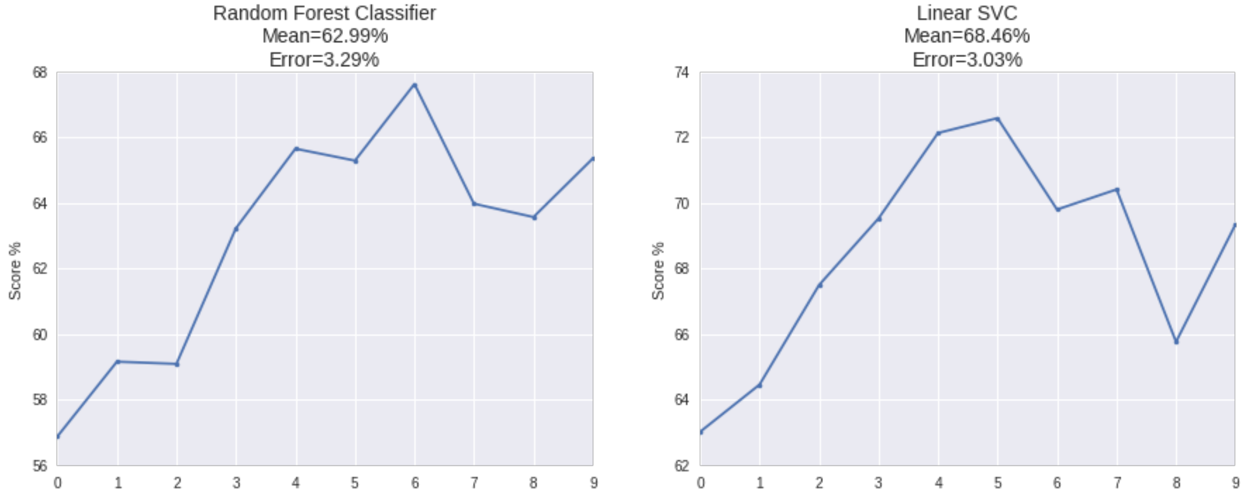


Figure 21. Results of 10-Fold Cross Validation

Evaluation

The team was able to address our initial objectives listed below with various machine learning algorithms.

- Predict the impact of new promotions on future sales
- Allow focused marketing by identify the driving demographics in yogurt sales
- Aid inventory decision by understanding geographic sales trends
- Predictions of future sales to inform better business decisions

We successfully created a model that predicts the impact of sales promotions and were able to gain insights into the driving demographic factors in yogurt sales. While we weren't able to utilize them in a model, we did discover geographic trends in sales data. Two attempts were made to achieve the last objective. Both a classification and regression model were built to predict future sales using multiple elements of our data. The regression model didn't achieve sufficient accuracy but the classification model was successful. A high percentage of our results successfully align with our objectives.