Team Froyo

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

Objective

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

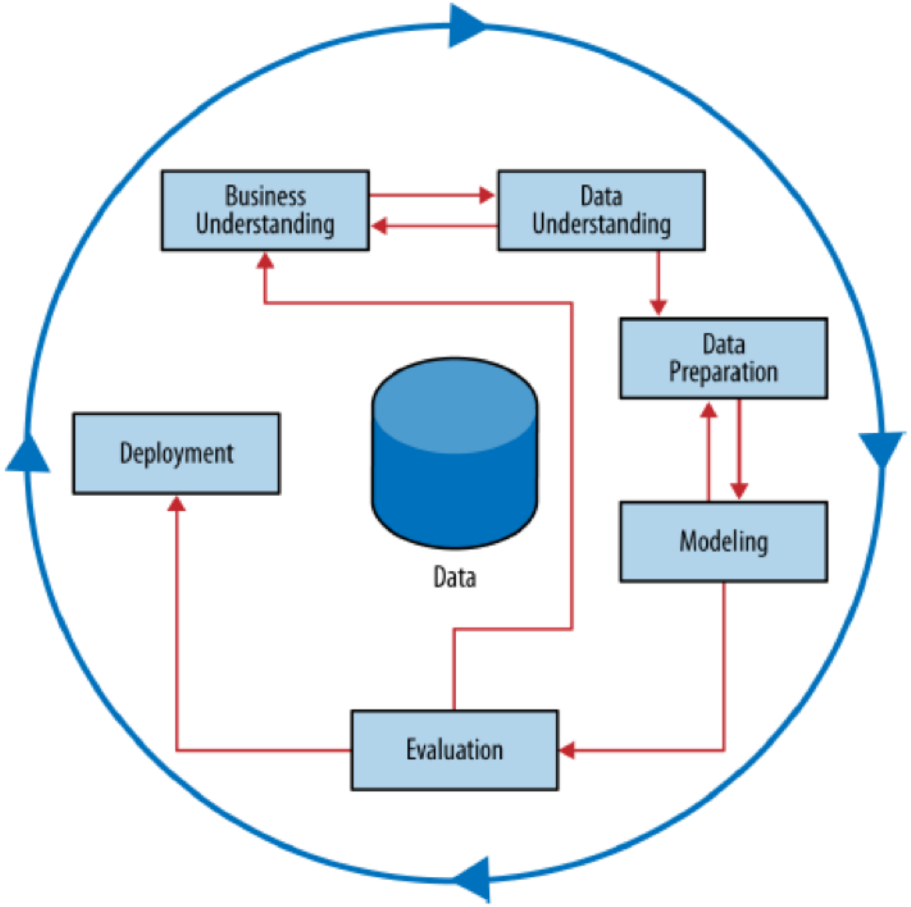
Desired Capability:

- 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

Methodology

Cross Industry Standard Process for Data Mining (CRISP-DM)

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment



CRISP Data Mining Process

Assess Situation

IRI Marketing Data Set

- Dataset containing store sales and consumer panel data for 30 product categories.
- Enables modeling of important economic and marketing metrics for companies and policy makers.

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

Resources

- Python Machine Learning Packages
- Hierarchical Data Format files (HDF5)
  - Benefits of database without requiring external server resources

Risks and Contingencies

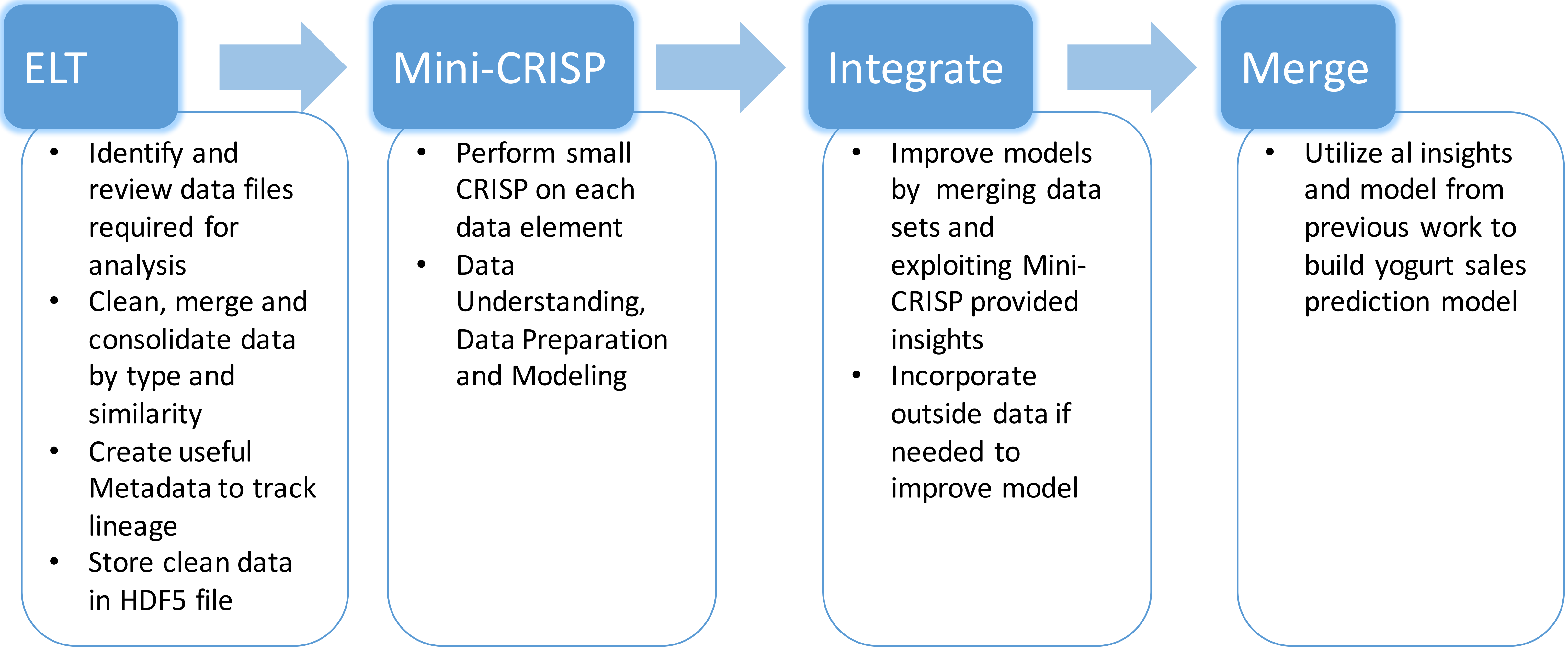
- Size of grocery store data was too large for personal laptop RAM. We used representative samples.
- Unable to link all keys between tables so performed analysis on subsets that could be merged. Analysts accepted potential information loss as small risk
- IRI data my not accurately represent the markets for which data is provided and may skew some of our conclusions.
- Scikit-learn can’t handle categorical data well which we mitigated by binary encoding.

Data Mining Goals

Primary Goal: Discover underlying trends in the data that can be used to build a predictive model

Promotions	• Prove wither the promotions, displays and/or price reductions has a statistically significant impact on sales
Panelists	• Identify panelists demographics that contribute to higher yogurt sales
Products	• Identify single or combinations of yogurt attributes that tend to sell better or worse
Stores	• Determine if there exists geographic trends in the sales data
Time Series	• Determine if there exists daily, weekly, monthly or yearly sales patterns

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

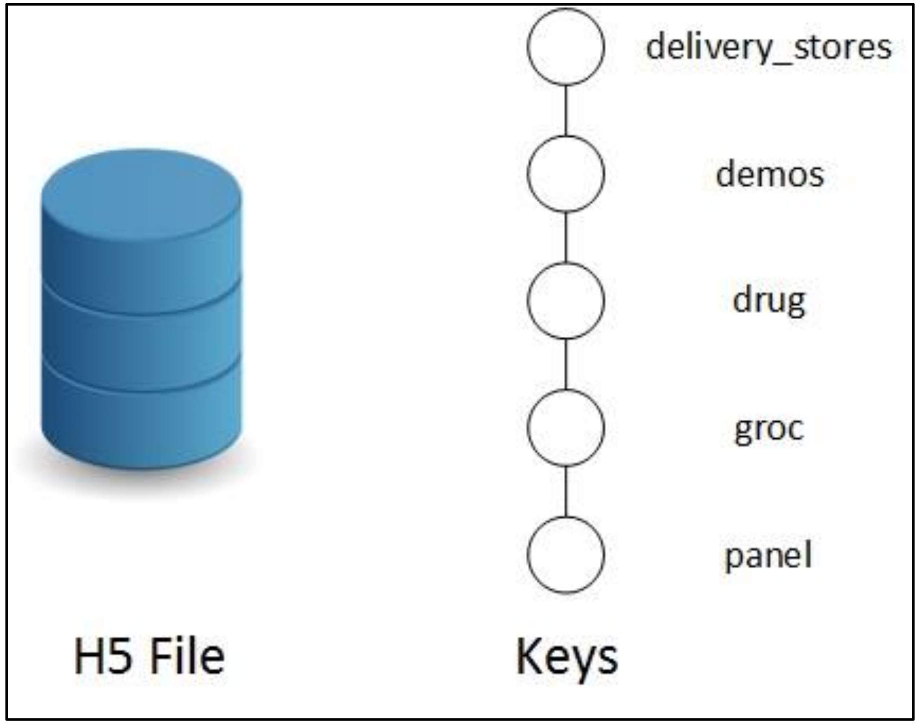
- Improve models by merging data sets and exploiting Mini-CRISP provided insights
- Incorporate outside data if needed to improve model

Merge

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

Big D. Challenges

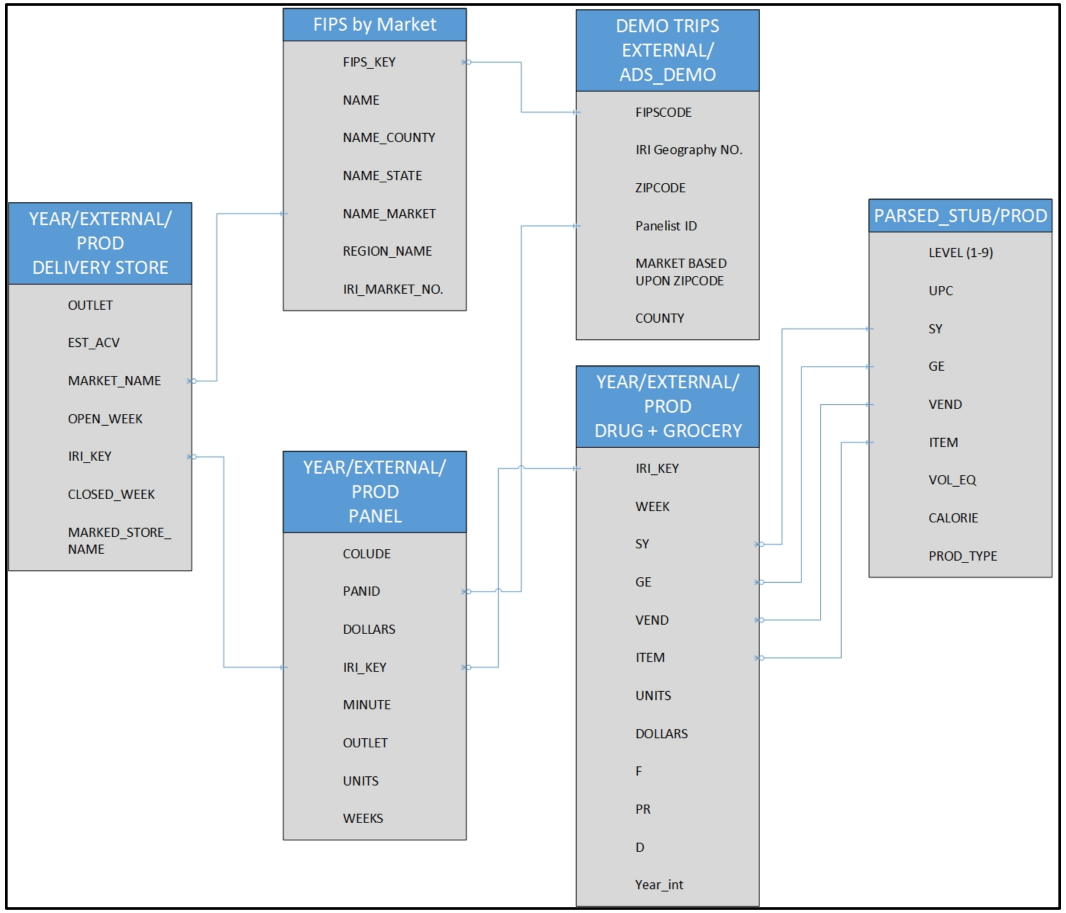
- Disparity
  - Similar files distributed over various folders
  - Different version of similar data
  - Different file formats and headers
- Volume
  - 140 GB Uncompressed ➞ 1.8 GB Compressed
  - 141,394,709 rows just for grocery yogurt sales
- Variety
  - 3,649 files ➞ 1 file
  - Demographics, Panelists, Trips, Stores, Sales, Products, etc.



Organization of Single H5 File

Data Description

- Store Descriptors
- US County Codes (FIPS)
- Panelist Demographics
- Panelist Sales
- Store Sales Data
- Product Descriptors



YEAR/EXTERNAL/ PROD DELIVERY STORE
OUTLET
EST_ACV
MARKET_NAME
OPEN_WEEK
IRI_KEY
CLOSED_WEEK
MARKED_STORE_ NAME

FIPS by Market
FIPS_KEY
NAME
NAME_COUNTY
NAME_STATE
NAME_MARKET
REGION_NAME
IRI_MARKET_NO.

YEAR/EXTERNAL/ PROD PANEL
COLUDE
PANID
DOLLARS
IRI_KEY
MINUTE
OUTLET
UNITS
WEEKS

DEMO TRIPS EXTERNAL/ ADS_DEMO
FIPSCODE
IRI Geography NO.
ZIPCODE
Panelist ID
MARKET BASED UPON ZIPCODE
COUNTY

YEAR/EXTERNAL/ PROD DRUG + GROCERY
IRI_KEY
WEEK
SY
GE
VEND
ITEM
UNITS
DOLLARS
F
PR
D
Year_int

PARSED_STUB/PROD
LEVEL (1-9)
UPC
SY
GE
VEND
ITEM
VOL_EQ
CALORIE
PROD_TYPE

Relationships between various Data Categories

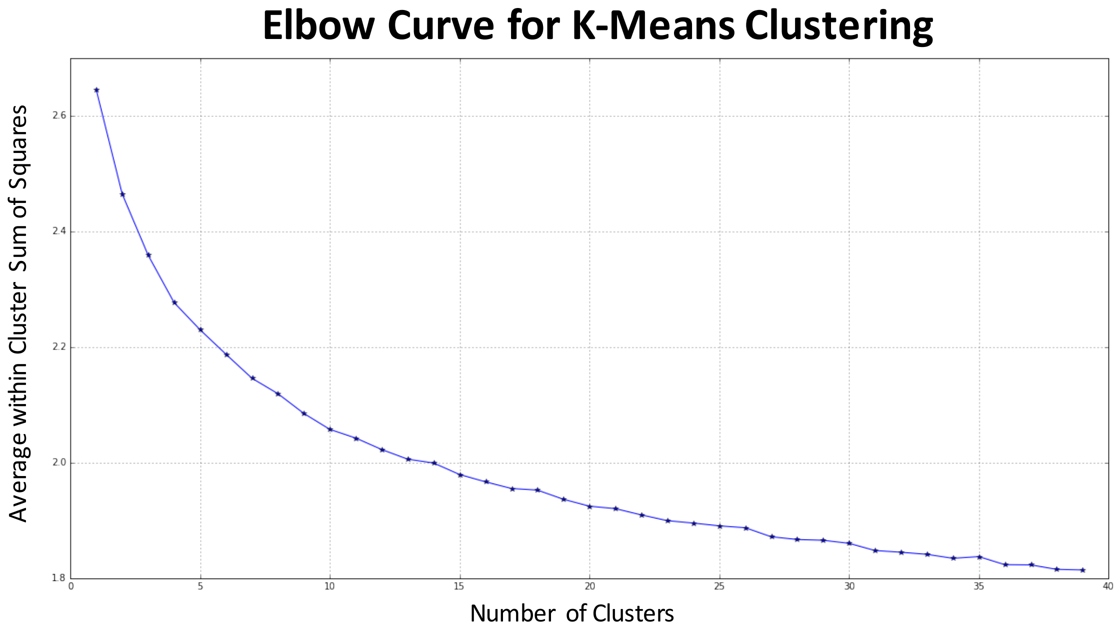
Panelists Demographics
Mini-CRISP

Business Understanding:

- Can we predict who purchases yogurt?

Data Understanding:

- K-Means Clustering used to categorize panelists into market groups
- 12 Clusters were identified
- Unable to achieve stable clustering



Panelists Demographics
Mini-CRISP

Data Preparation:

- Not all panelists were yogurt consumers. Merged panelist transactions with products table to identify those panelists who were yogurt consumers and those that were not.

Modeling:

- Random Forrest Classification
  - Features selected from panelist demographic data.
  - Label is either panelist purchases yogurt or they do not
- Model Accuracy 75% after parameter tuning

Top 5 Features:

Has a TV

Has Cable

Family Size

Number of Dogs

Number of Cats

Evaluation:

- 10-fold Cross validation was 58% (+/- 7%)

Sales Promotions

Mini-CRISP

Business Understanding:

- Can we predict the success of a sales promotion?

Data Understanding:

- Promotion Success = Sales during the week where an advertisement, display or price reduction was run are more than one standard deviation from the mean

Sales Promotions	Percentage 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 %

Panelists Demographics
Mini-CRISP

Data Preparation:

- Calculate the mean weekly sales for each product in each store. Identify those weeks with successful promotions

Modeling:

- Important features were the various sales promotion fields. Attempts to improve model by incorporating product attributes, store attributes, price, outlet, time of year were unsuccessful
- Random Forrest Classification
  - Label was successful promotion or not
- Grid Search Parameter Tuning

Evaluation:

- 10 fold cross validation
- Model Accuracy 78%

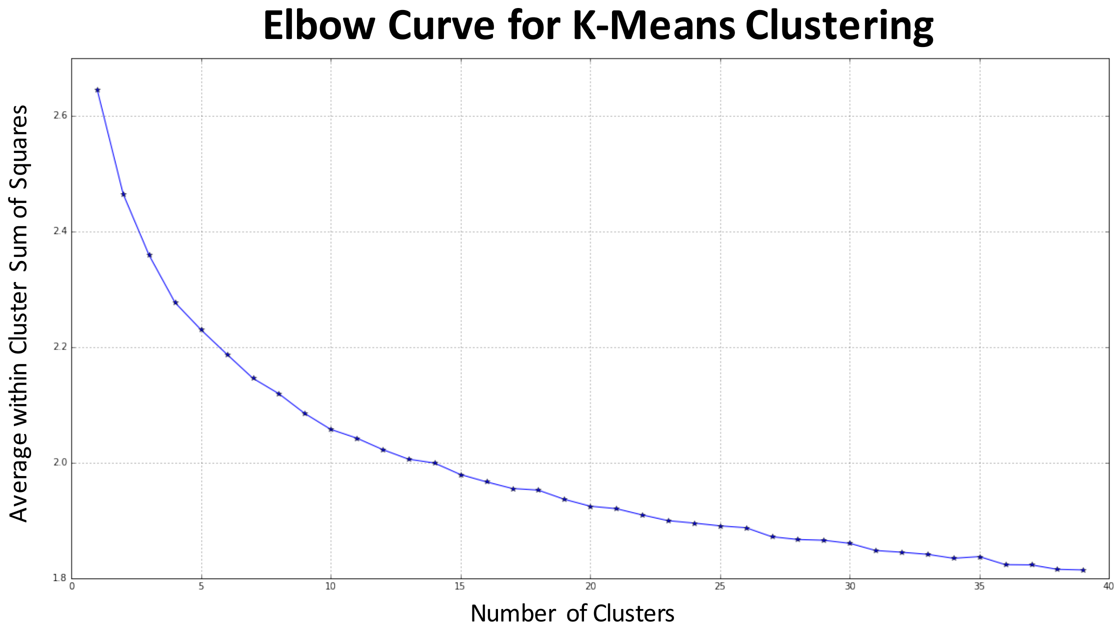
Geography, Products, Sales Date

Data Exploration

- The team examined yogurt product attributes, stores geography and sales dates for potential modeling
    - C?

Data Understanding:

- K-Means Clustering used to categorize panelists into market groups
- 12 Clusters were identified
- Unable to achieve stable clustering



Unit Sales Prediction
CRISP

Business Understanding:

- Can we predict next week yogurt sales?

Data Understanding:

- The team examined yogurt product attributes, stores geography and sales dates for underlying sales trends
- Identified trends include:
  - States with higher mean yogurt sales than others
  - Drop in yogurt sales during fall and winter months

Unit Sales Prediction
CRISP

Data Preparation:

- Merge multiple datasets that were identified as contributors to yogurt sales

Modeling:

- Two Modeling Methods were used: Classification and Regression
- <u>Regression</u>
  - Build model that predicts the weekly unit sales
  - Lasso, Ridge and SVR
  - Feature selection and parameter tuning performed
- <u>Classification</u>
  - Build model that predicts wither next week's sales will be higher than national mean
  - Random Forrest, SVC
  - Feature selection and parameter tuning performed

Unit Sales Prediction

CRISP

Evaluation:

- Regression
  - Despite feature selection and parameter tuning efforts, attempts to build a regression model were unsuccessful
  - Only 44% Maximum Accuracy Achieved

- Classification

Model	Random Forest	Linear SVC
Score	80%+	72%+
10 Fold CV	Unstable	Stable

Conclusion

Project Results:

- Predict the impact of new promotions on future sales
- Explore the driving demographics in yogurt sales
- Understanding geographic sales trends
- Prediction of future sales