Dairy Shop Sales Analysis

Data provided by Nielsen

Gokturk Demir, Ji Hun (James) Lee,
and Weihuang (Steven) Xie

Data Overview and Engineering

- Household Data: Customer Profile
  - 60000 x 58; e.g. income, household size, education, race, employment
- Purchase Data: Purchasing data specific to each UPC
  - 2 mil x 7; e.g. price and quantity
- Product Data: Product Features
  - 10000 x42; e.g. yogurt, cereal, etc
- Trip Data: Record 1 million tractions made by 50,000 households
  - 1,000,000 x 8; important variables: amount spent
- Joining these tables for analysis
  - Joined via foreign key: household code, upc, trip code
  - Advanced analysis requires weaving multiple tables into one fabric
- Analysis to be done on this dataset:
  - Unsupervised Learning to find clusters within household and product
  - Supervised Learning to find relationships among sales, price, promotion, yogurt

Household Segments

RFM Model --Scoring

One Problem with Data: Time frame for each customer's transaction is different

Solution: RFM is a customer analysis model especially for purchase information:

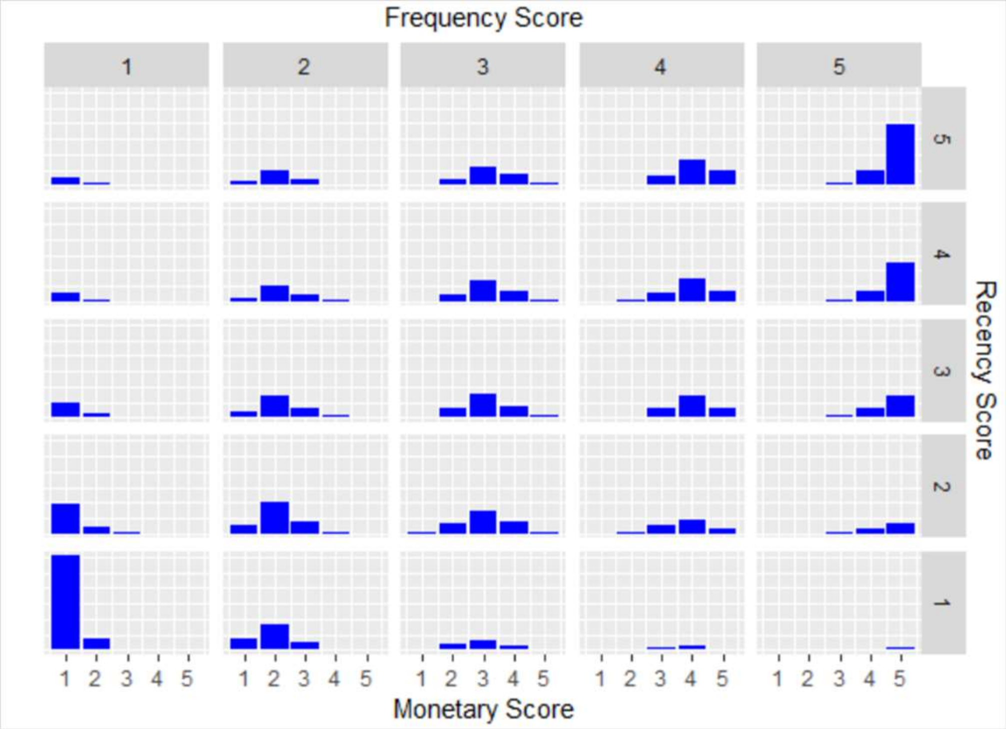
1. Recency(R): How recently did the customer purchase?
2. Frequency(F): How often you purchase?
3. Monetary Value(M): How much they buy?

RFM Model Output:

	customer_id	recency_days	transaction_count	amount	recency_score	frequency_score	monetary_score	rfm_score
	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>	<int>	<dbl>
1	2000000	3	7	17.8	5	1	1	511
2	2000076	25	25	68.3	3	3	3	333
3	2000112	41	13	32.5	2	2	2	222
4	2000126	0	154	321.	5	5	5	555
5	2000129	119	3	3.98	1	1	1	111
6	2000179	20	18	116.	3	3	4	334
7	2000185	49	30	87.6	2	4	3	243
8	2000213	11	38	70.6	4	4	3	443
9	2000235	6	109	273.	5	5	5	555
10	2000258	21	8	41.5	3	1	2	312

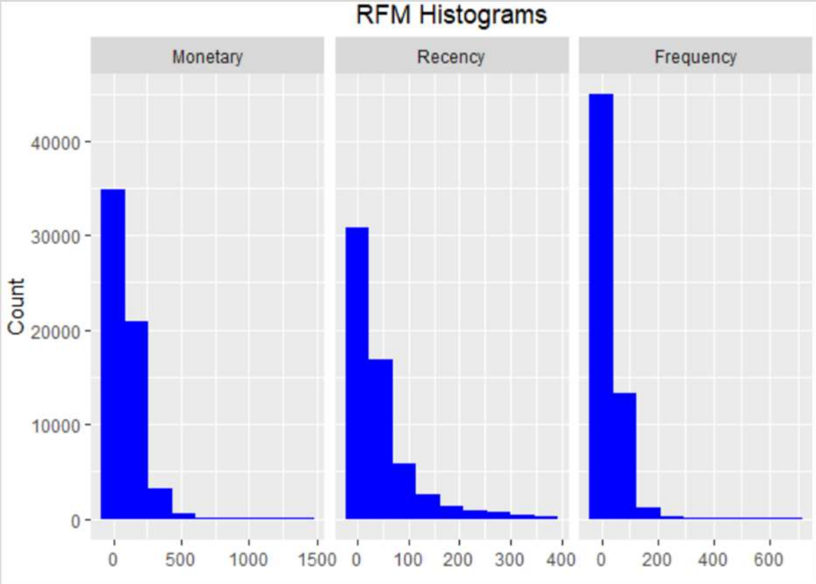
Household Segments: Data Description i

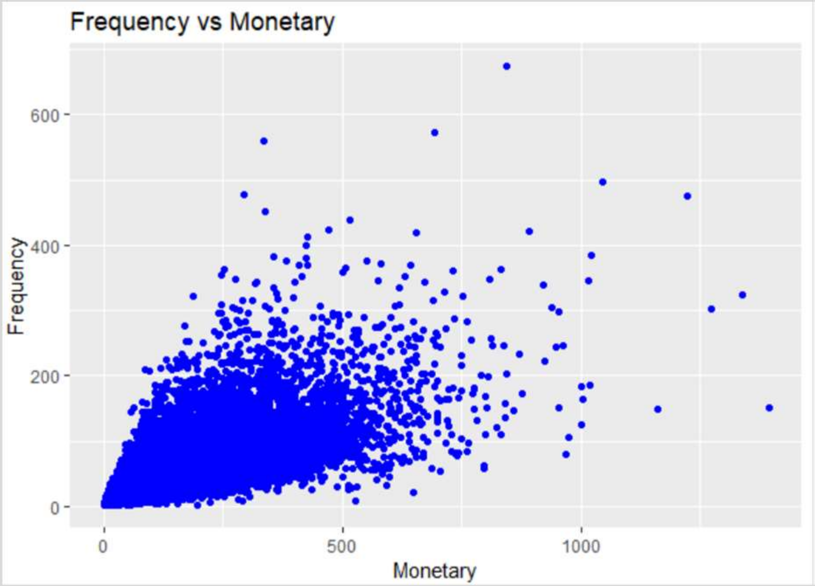
RFM Model --Overview



Household Segments: Data Description II

RFM Model --Overview





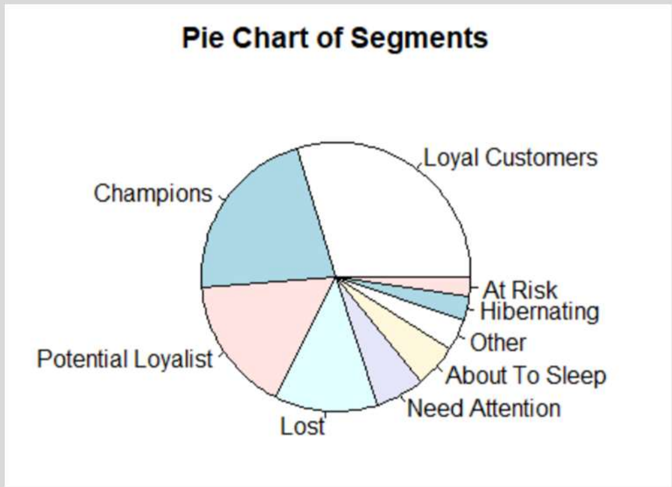
Frequency and Monetary tend to have positive relationship;
Most people spent 0-250 dollars; 100 times a year; within last 50 days;

Household Segments: Segmentation

RFM Model --Segment

Segment	Description	R	F	M
Champions	Bought recently, buy often and spend the most	4 - 5	4 - 5	4 - 5
Loyal Customers	Spend good money. Responsive to promotions	2 - 5	3 - 5	3 - 5
Potential Loyalist	Recent customers, spent good amount, bought more than once	3 - 5	1 - 3	1 - 3
New Customers	Bought more recently, but not often	4 - 5	<= 1	<= 1
Promising	Recent shoppers, but haven't spent much	3 - 4	<= 1	<= 1
Need Attention	Above average recency, frequency & monetary values	2 - 3	2 - 3	2 - 3
About To Sleep	Below average recency, frequency & monetary values	2 - 3	<= 2	<= 2
At Risk	Spent big money, purchased often but long time ago	<= 2	2 - 5	2 - 5
Can't Lose Them	Made big purchases and often, but long time ago	<= 1	4 - 5	4 - 5
Hibernating	Low spenders, low frequency, purchased long time ago	1 - 2	1 - 2	1 - 2
Lost	Lowest recency, frequency & monetary scores	<= 2	<= 2	<= 2

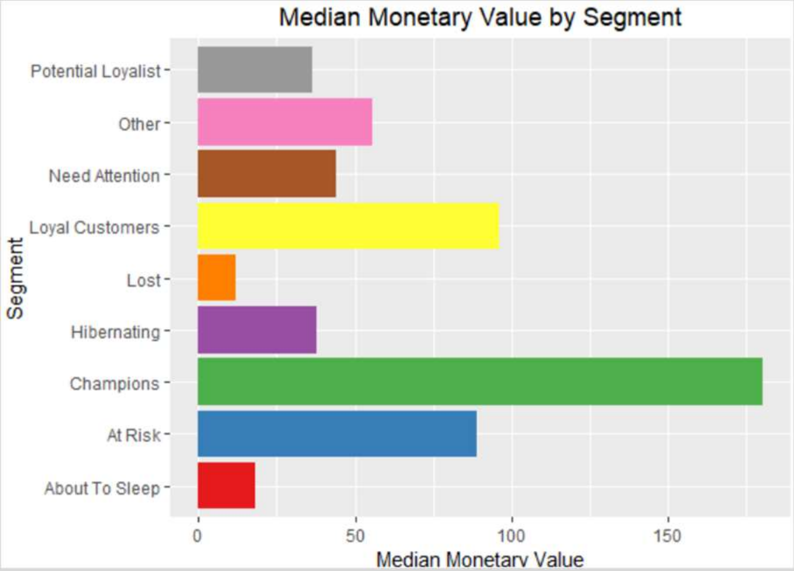
https://cran.r-project.org/web/packages/rfm/vignettes/rfm-customer-level-data.html

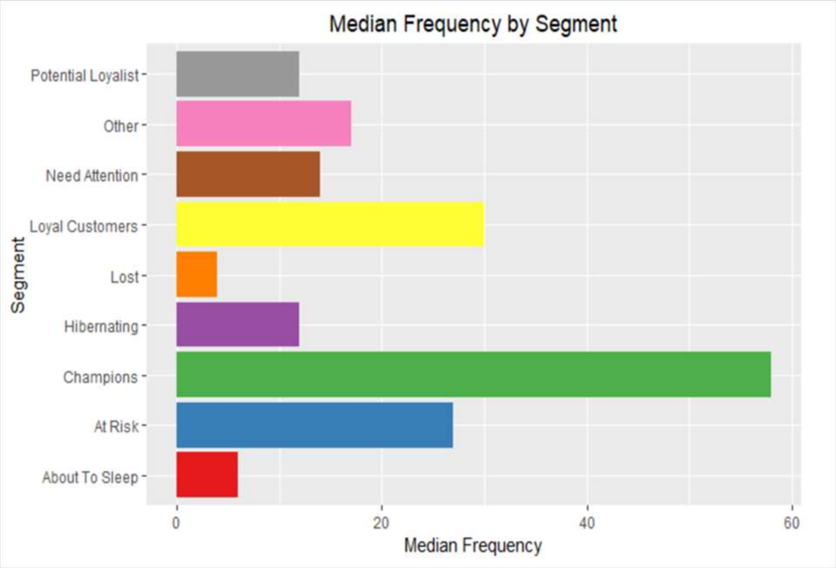


Champions and loyal customers are the majority; good amount of potential loyalist; Also significant amount of lost customers

Household Segments

RFM Model --Segment Characteristics





Champions and loyal customers contribute the main sales

Attributes Driving Sales

(for Champions and Loyalist Segment)

SEM Model --Only Numeric Attributes(ML)

Latent Variables:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
income =~						
household_incm	1.000				5.712	1.000
size =~						
household_size	1.000				1.352	1.000
age =~						
ag_nd_prsnc_f_	1.000				4.080	1.469
female_head_ag	0.114	0.006	17.716	0.000	0.465	0.191
male_head_age	0.105	0.006	16.686	0.000	0.430	0.136

Regressions:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
total_sales ~						
income	1.392	0.102	13.605	0.000	7.953	0.077
size	16.038	0.446	35.953	0.000	21.688	0.211
age	0.006	0.089	0.070	0.944	0.025	0.000

Generally, income and household size are driving the overall sales of our main customer.

Attributes Driving Sales

SEM Model --All Attributes(DWLS)

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  f1 =~
    household_size    1.000
  f2 =~
    household_incm    1.000
  f3 =~
    female_head_ag    1.000
    fml_hd_mplymnt    0.455    0.017   26.227    0.000
  f4 =~
    male_head_dctn    1.000
    mal_hd_mplymnt    0.684    0.005  146.787    0.000
    male_head_age     2.716    0.032   86.010    0.000

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  total_sales ~
    f1               19.835    0.248   80.016    0.000
    f2                1.190    0.068   17.547    0.000
    f3                1.141    0.251    4.542    0.000
    f4                6.601    0.519   12.709    0.000
```

```
Thresholds:
                   Estimate  Std.Err  z-value  P(>|z|)
    fml_hd_mplym|1   -1.363    0.007 -186.716    0.000
    fml_hd_mplym|2   -0.810    0.006 -139.810    0.000
    fml_hd_mplym|3   -0.657    0.006 -118.349    0.000
    fml_hd_mplym|4    0.255    0.005   49.124    0.000
    mal_hd_dctn|t1   -0.682    0.006 -122.028    0.000
    mal_hd_dctn|t2   -0.663    0.006 -119.137    0.000
    mal_hd_dctn|t3   -0.572    0.005 -105.003    0.000
    mal_hd_dctn|t4   -0.078    0.005  -15.099    0.000
    mal_hd_dctn|t5    0.476    0.005   89.022    0.000
    mal_hd_dctn|t6    1.301    0.007  184.104    0.000
    ml_hd_mplymn|1   -0.682    0.006 -122.028    0.000
    ml_hd_mplymn|2   -0.545    0.005 -100.685    0.000
    ml_hd_mplymn|3   -0.476    0.005  -89.014    0.000
    ml_hd_mplymn|4    0.673    0.006  120.639    0.000
```

Income, household size , female & male head status(education, employment, age) are driving the overall sales of our main customer.

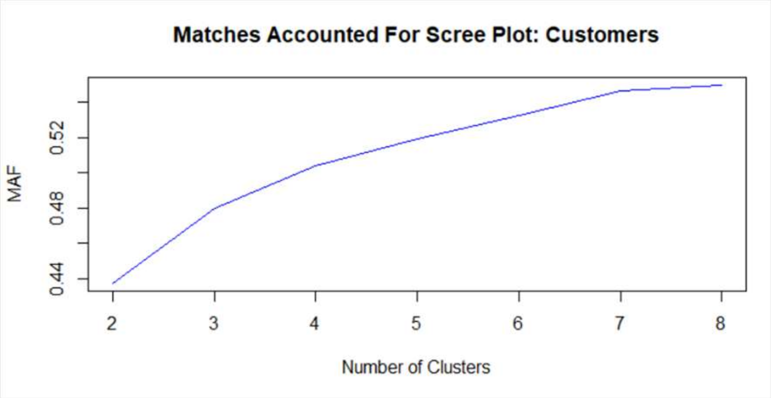
Main Products Driving Sales

	product_descr	product_types	total_sales_sum	total_quant_sum
1249	YOGURT	4257	3570626.92	2877248
1248	SPCL K MG OAT & HNY	6	551131.20	182196
1247	GREAT GN DATE RS/PEANUT	4	372331.20	115472
1246	YOGURT/ADDITIVE	162	359674.91	377590
1245	HNY/NUT CHR	24	237072.61	66835
1244	CHR	21	206459.25	64980
1243	CN HARVEST	6	178199.68	49326
1242	RS BRAN	165	154297.38	54756
1241	FRST FLK	157	153548.39	53609
1240	OM SQ	11	152714.25	46463

We will focus on Yogurt and do more analysis specific to customer who purchase yogurt. .

HOUSEHOLD CLUSTERING: METHODOLOGY

- Segmentation of households in panelists table
  - All column values are categorical
  - Focus on top 1000 projection factor households
- Use K-Modes (analogue of K-means) and LCA
  - Our complexity parameter: **number of cluster**
  - Validation for the **number of clusters**
  - 4 Clusters!



HOUSEHOLD CLUSTERING: RESULT

1) $100 +K Annual income, no children under 18, White/Caucasian , High School Graduate, Single
2) $100 +K Annual income, 0-12 years old children, White/Caucasian, Some college, Married
3) $70-99K Annual income , no children under 18, White/Caucasian , High School Graduate, Divorced/Separated
4) $100 +K Annual income, no children under 18, White/Caucasian, High School Graduate, Married

YOGURT CLUSTERING

flavor_code	style_code	type_code
Vanilla	Regular	Low-Fat
Plain	Regular	Low-Fat
ST	Regular	Low-Fat

- There are so many different types of yogurt, so we want to know the main types (=clusters)
- Why are segments' type code and style code the same?
    - It's because data are predominantly regular type (not organic) and low_fat
- How can we find segmentation when data is dominated by one category?
    - Ignore the dominant type, and then cluster the rest to find diversity among segments to avoid overgeneralization

Which segments are the most price sensitive to yogurt?

- National/Regional Level
  - Daily transaction data in the whole year 2011; aggregate yogurt and non-yogurt sale data with price and quantity
  - Model: $logsale_t = \beta_0 + \beta_1 * logprice(yogurt_t) + \beta_2 * logprice(rest_t) + e_t$
  - Overall price sensitivity of yogurt: inelastic; coefficient = 0.6 < 1 (national level)
  - More or less similar elasticities depending on the region (nine regions)
  - Coefficient is significant but simple linear regression is ill-fitting
    - Shapiro-Wilks test: short-tailed normal distribution of residuals
    - Residual vs Fitted value plot: heteroskedasticity - diminishing residuals as fitted values increase
  - Variance stabilizing transformation: box-cox transformation (lambda=0.5 for interpretation purpose)
    - New coefficient value: 0.4
- Segment Level
  - Same model as national level; some segments suffer ill fit for regression
  - Variance stabilizing transformation
  - Price elasticity is the highest among the segment 3 (~.71), but not by large margin than the rest
  - All segments we have made are insensitive to yogurt price increases

Which marketing mix is the most important to sale of a general yogurt?

- Segment Level: select 1000 households with the highest projection factors
- Marketing Mix Aggregate: Bayesian Linear Model
  - Model (segments 1 through 4): MCMCregress(sale1 ~ price1 + promotion1 + region1)
  - Using credible interval in the Quantile Summary to interpret the “statistical significance” of coefficient to see whether the interval contains 0
  - Similar result obtained by OLS
  - Result: price and promotion significant for all segments but with varying degrees. Promotion most effective for the third segment!
- Marketing Mix Household: Bayesian Hierarchical regression
  - Divide sale, price, promotion into intervals and make them factor variables
  - Hierarchical MCMC: regression on individual household
  - MCMChregress(fixed=sale ~ price + promotion + region,random=~price +promotion+region, group=”household_id”,r=6,R=diag(6))
  - “Elasticities” of sale obtained with respect to price, promotion, and region at individual level

Which household segments are likely to choose which type of yogurt (segment)?

- Simple Model: Contingency table?
    - One way to answer is using conditional distribution: E.g. P(yogurt=1|household=1)
- Choice modeling of yogurt segment based on household segment
    - Multinomial Logit Model: mlogit(yogurt ~ household)
    - Perspective: A very simple model with one predictor only but 4 levels : Essentially multinomial regression on contingency table
    - Result: No significant predictors; in other words, <u>No household segment is likely to strongly favor some particular yogurt segment</u>
- More Complex Model?
    - Problem with the aggregate model: <u>Segmentation is too reductive</u> and k-mode clustering may not be the best model for the type of data in which some level dominates all others
    - Or, national level data is <u>not specific enough</u> of data for model to capture important trends
    - mlogit(yogurt_segment ~ income + household_size + education + race + employment + region)
- <u>At national level, there is no predictive relationship between household and yogurt segments</u>

Summary

1. We create 8 segments of the overall households based on RFM model. The main customer segments are the “champaign” and “loyal customer” groups. In these groups, income, household size and male status are the key sales drivers.

2. We clustered the customers and yogurt, which is the best selling product.We found out that our customers have 4 different clusters and yogurt has 3.

3. Key Results: 1) We found household segments that are the least price sensitive to yogurt 2) We find marketing mix preference at each household level 3) We find no predictive relationship between household segment and yogurt segment