Marketing Analytics Course Final Project

——Based on Nielsen Yogurt 2012 Dataset

Wendy Gao
June, 2017

What problems are we going to solve for today?

- Data Introduction and Variable Information
- Data Exploration
- Customer Analysis: Panelist Segmentation
- Sales Analysis: Marketing Mixed Model, Time Series Model and Pricing Model
- Conclusion
- Limitations and Further Improvement

Data Introduction:

Nielsen Data from RCC, Yogurt and Cereal Purchasing

Panelist

60538 rows, 58 variables:

Household income, size, composition, Age and presence of children, Type of residence, Female/male head age, employment, education, occupation, etc.

Purchase

2069789 rows, 7 variables:

UPC, Quantity, Total price paid, Coupon value, Deal flag UC, etc.

Trip

1095213 rows, 8 variables:

Household code, Purchase date, Retailer code, Store code, Total spent, etc.

Product

13681 rows, 16 variables:

Flavor, form, formula, container, style, type, brand, variety, etc.

Business Problems:

Nielsen Data from RCC, Yogurt and Cereal Purchasing

Panelist & Purchase

- Who are our customers?
- Panelist Segmentation
- Which cluster has the most purchase?
- Promotion Plan

Purchase & Trip

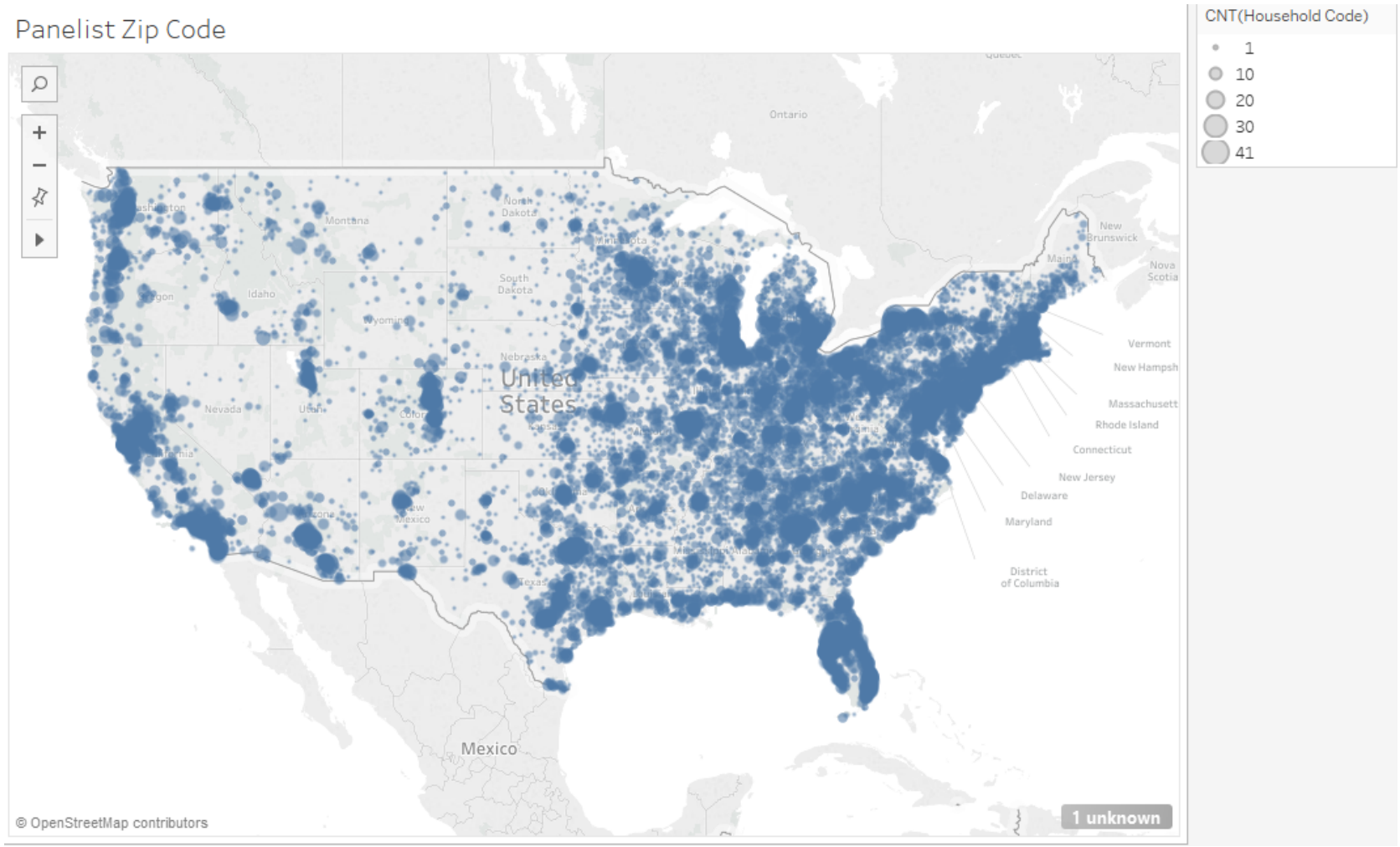
- Sales Analysis of Overall Purchasing Quantity
- Sales Analysis of Certain Type Product
- Sales Volume Prediction based on Seasonality

Trip & Product

- Price Elasticity
- What is the Unit Price for Yogurt that could generate largest profit
- How to improve Promotion Activities based on Current Price

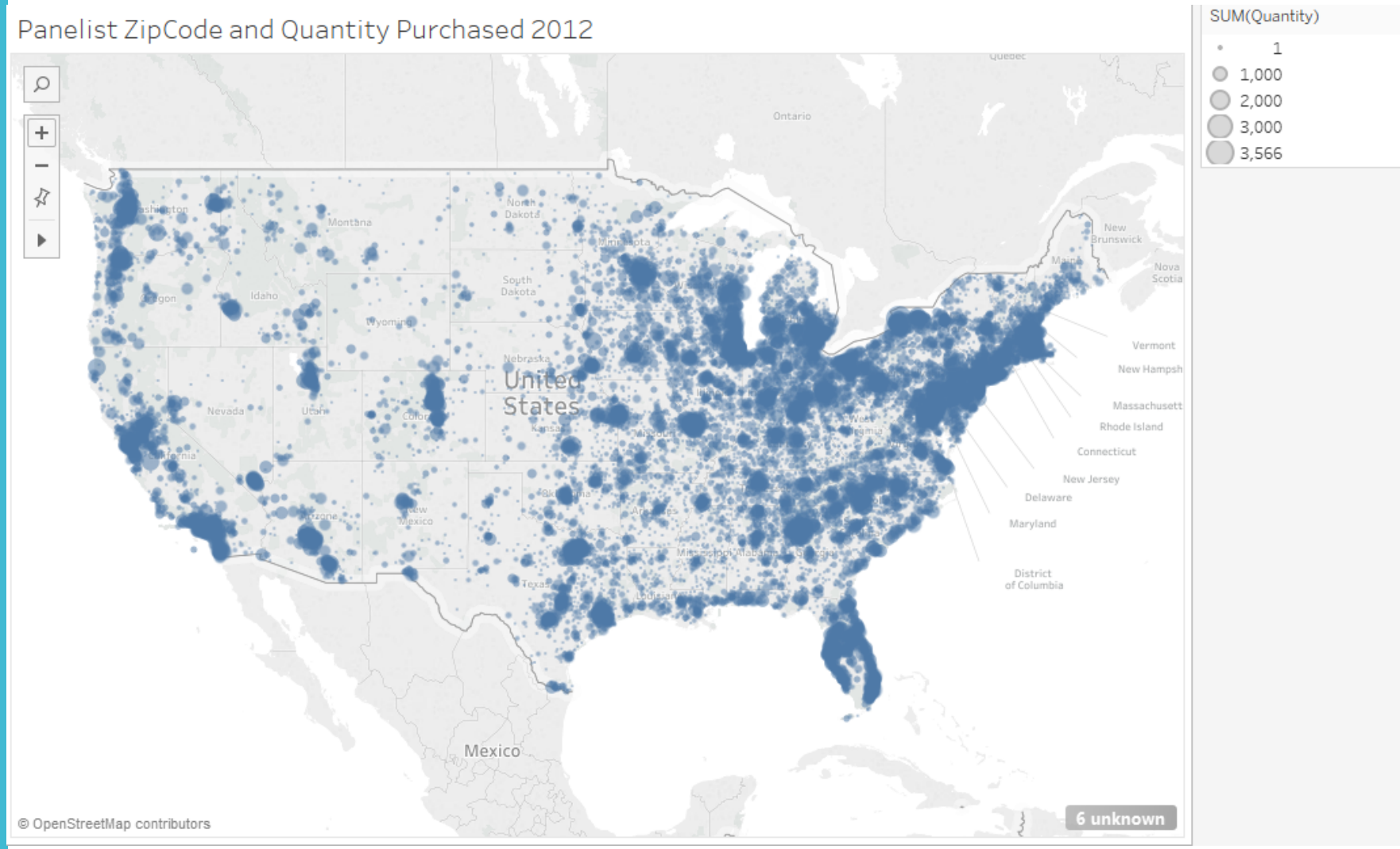
Data Exploration:

Panelist Information - Location



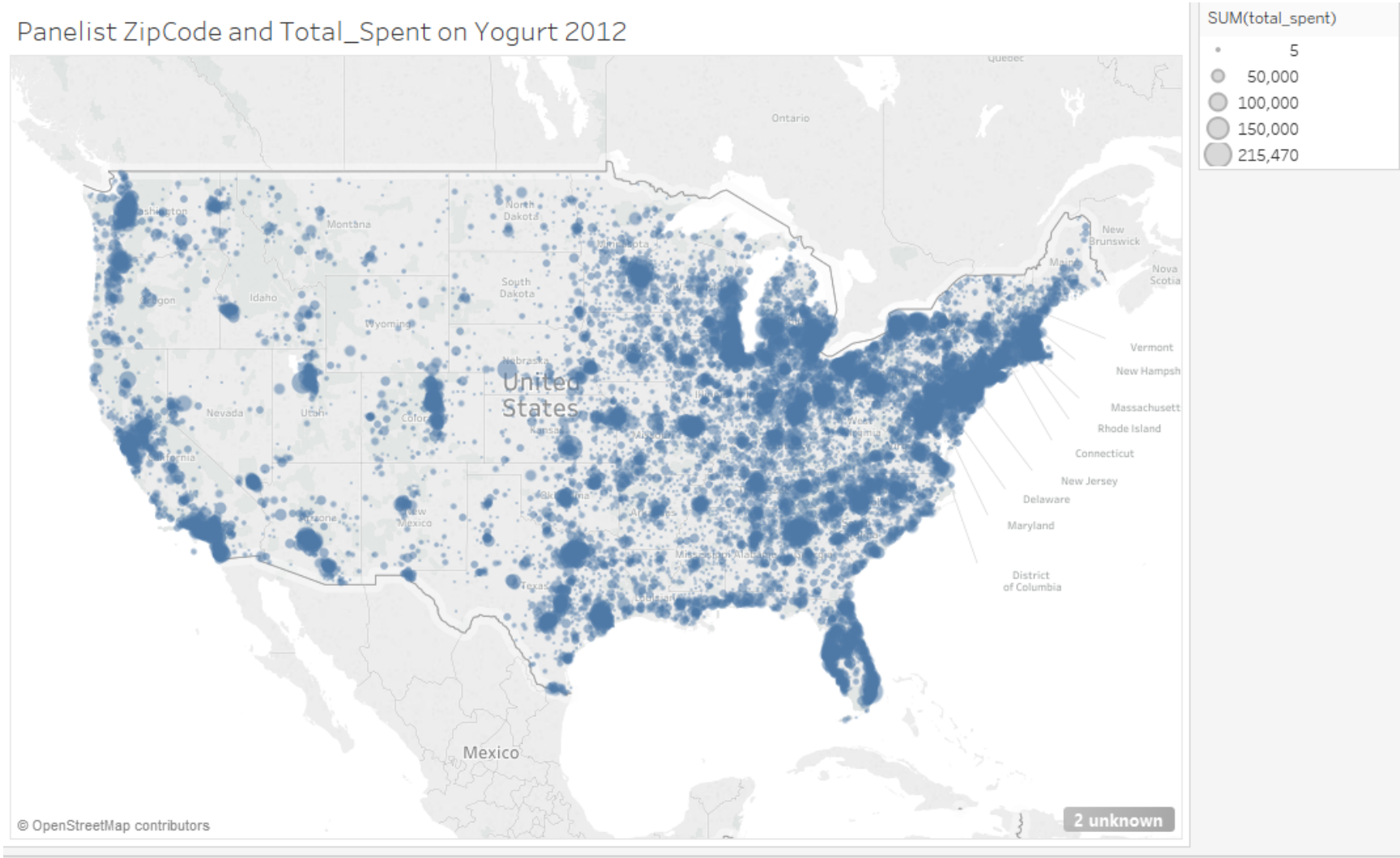
Data Exploration:

Panelist Information - Location & Purchasing Quantity



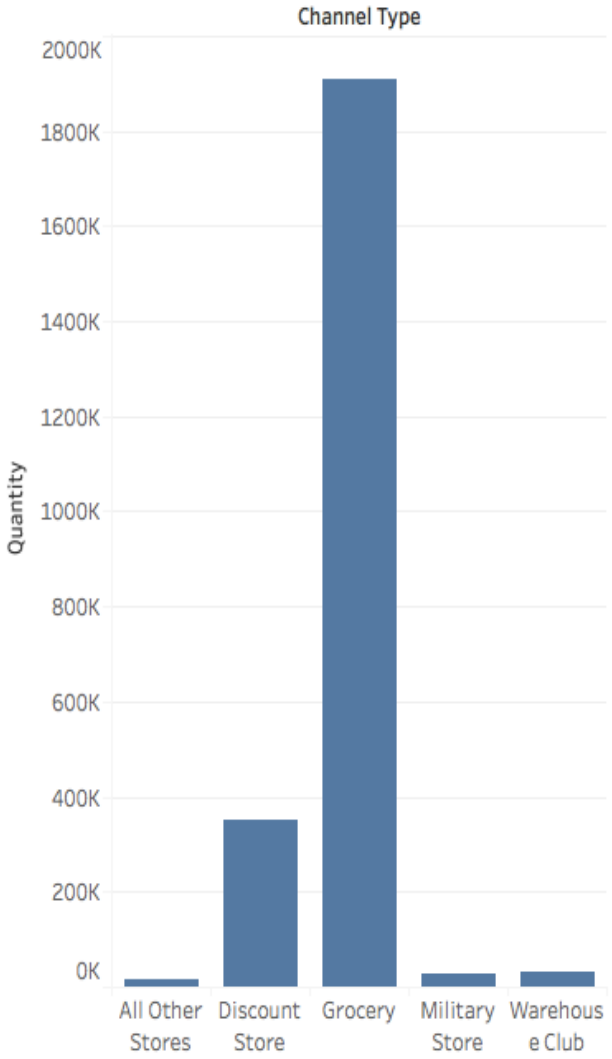
Data Exploration:

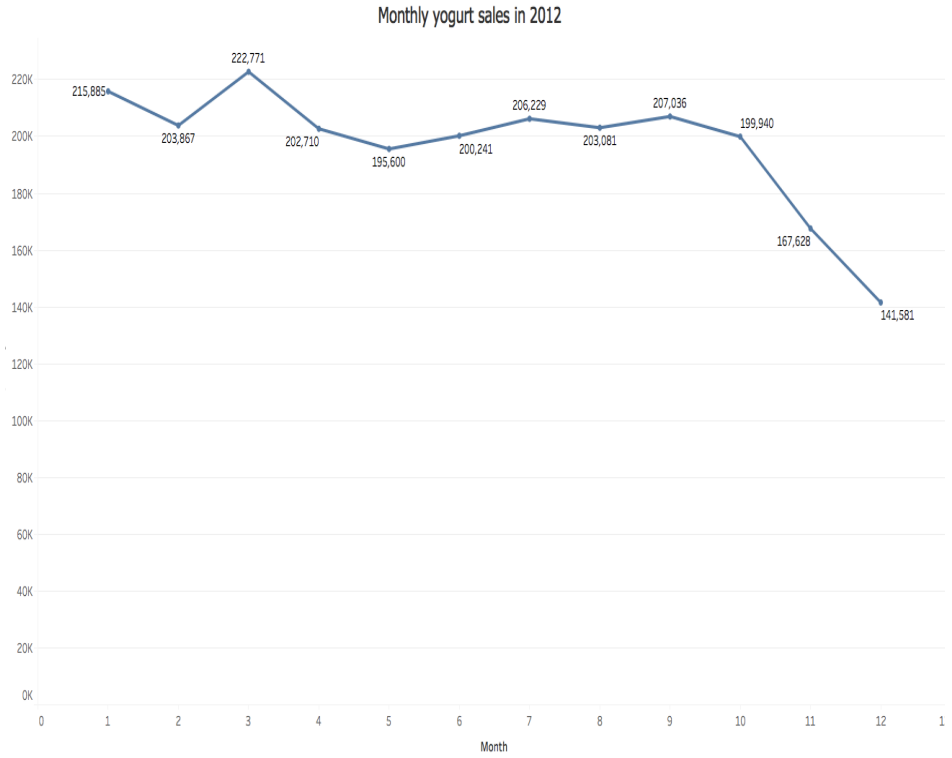
Panelist Information - Location & Total Spent



Data Exploration:

Channel Type & Monthly Sales



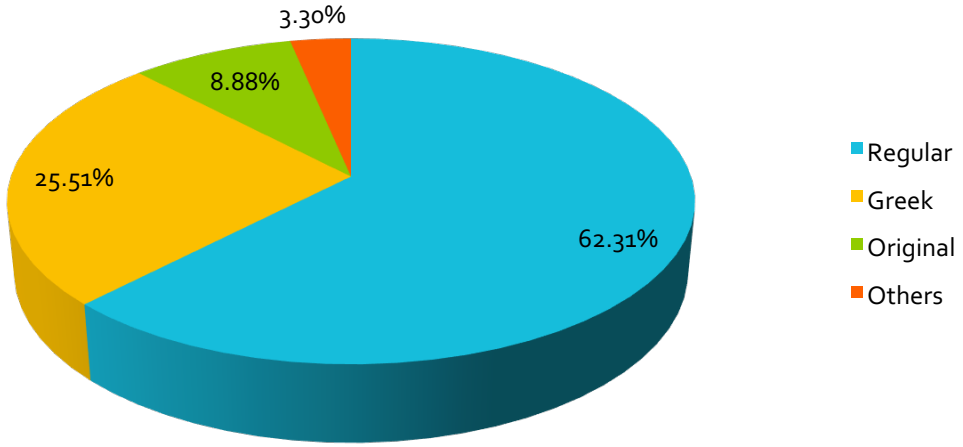


Data Exploration:

Market Share

Style	quantity	total_price_paid
REGULAR	1474492	1460485.15
GREEK	508043	776263.81
ORIGINAL	210194	175497.98
ALL NATURAL GREEK	71145	154835.29
ORGANIC	62971	120492.18
GREEK ORGANIC	24609	58702.17
CUSTARD	8206	5253.02
SWISS	4972	5260.65
SWISS PREMIUM	1230	1587.94
LEBEN	266	153.48
KEFIR	166	449.99
REAL	102	41.30
PREMIUM ORGANIC	49	190.91
PREMIUM	46	177.61
BULGARIAN	38	162.28
GELATIN	23	8.65
ORIGINAL RUSSIAN	5	12.20
FRENCH	4	1.99
GREEK KEFIR	4	4.98
PREMIUM BULGARIAN	3	8.48

Market share of different yogurt style

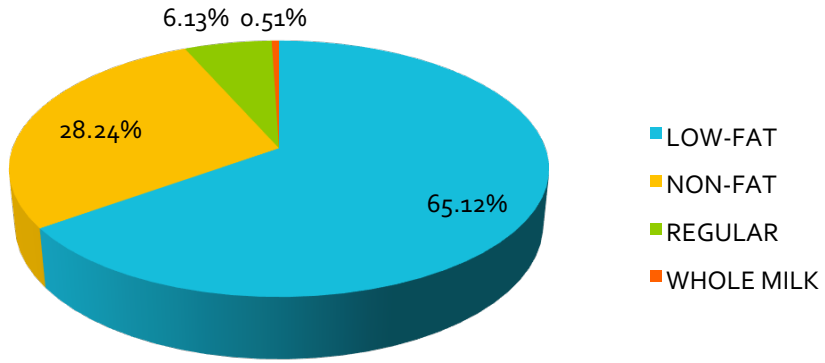


Data Exploration:

Market Share

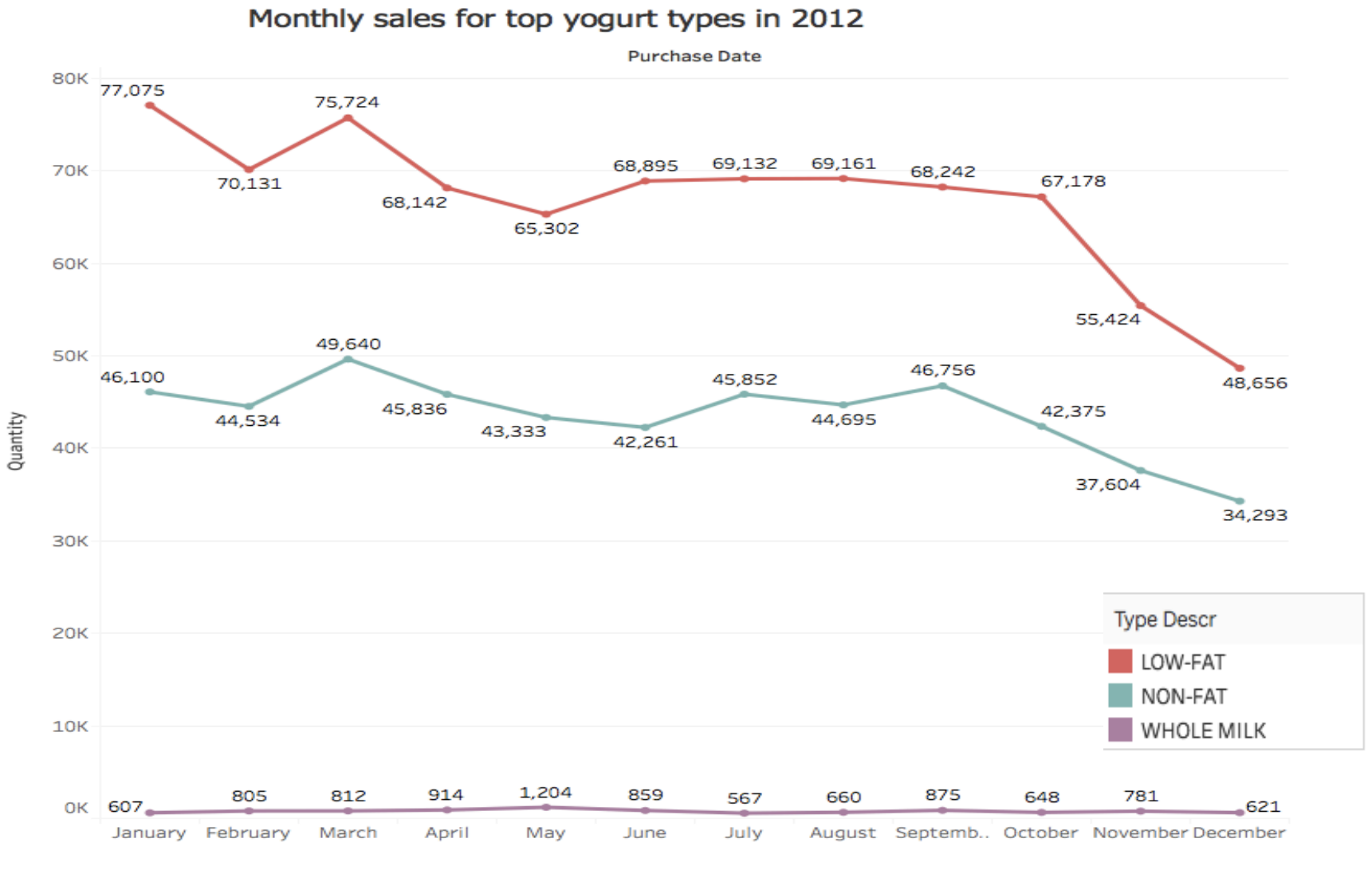
Type	quantity	total_price_paid
LOW-FAT	803062	882917.42
LIGHT NON-FAT	623193	603398.52
NON-FAT	523279	724188.92
REGULAR	139491	209160.42
FAT FREE	85237	153662.88
LITE NON-FAT	55355	23816.72
LIGHT FAT FREE	46865	26891.83
NON-FAT NATURAL	29602	45532.73
LOW-FAT LACTOSE FREE	13458	10052.13
NATURAL	11109	25404.86
WHOLE MILK	9353	14100.72
LOW-FAT NATURAL	5151	12960.06
LITE NON-FAT NATURAL	4605	2118.27
NON-FAT LACTOSE REDUCED	4495	5330.79
LIGHT	3754	8508.13
WHOLE MILK NATURAL	2232	2060.54
LITE FAT FREE	1547	623.45
FAT FREE NATURAL	964	3134.53
NON-FAT LACTOSE FREE	767	976.37
LOW FAT	635	1036.33
NON-FAT SKIM	598	1221.21
GOAT MILK	557	1162.49

Market Share of different yogurt type



Data Exploration:

Monthly Sale

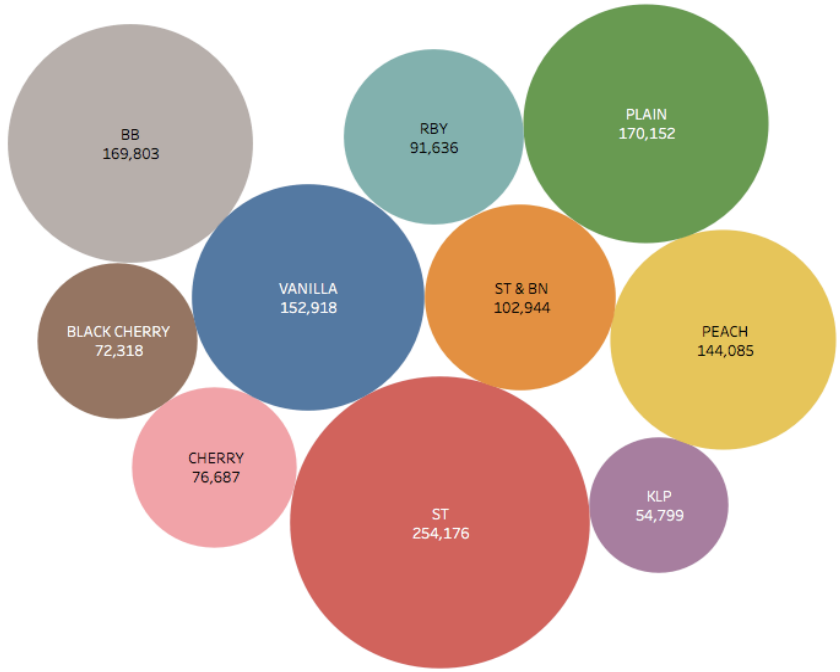


Data Exploration:

Sales of Flavor

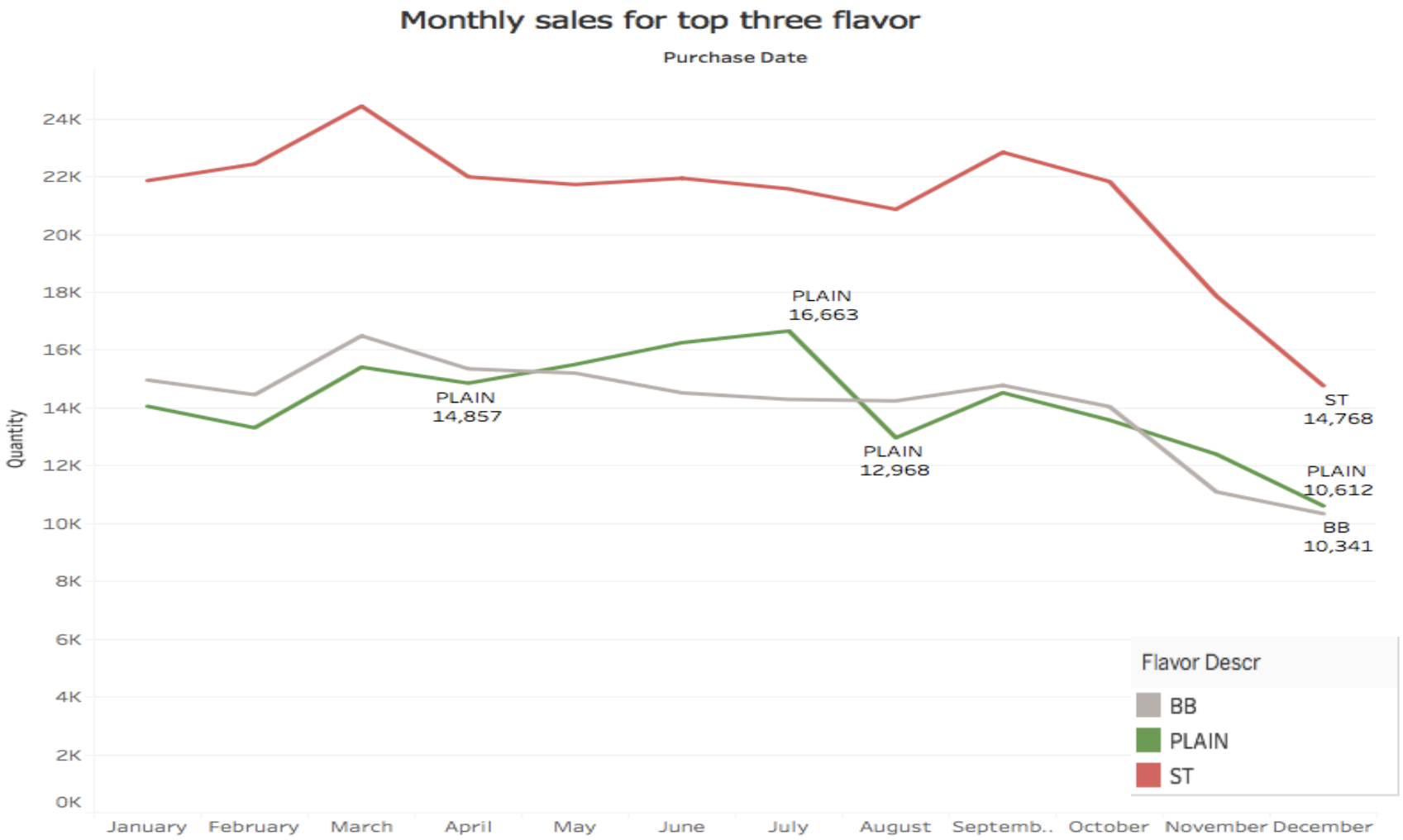
Flavor	quantity	total_price_paid
ST	254176	268684.92
PLAIN	170152	387093.45
BB	169803	165215.59
VANILLA	152918	238313.68
PEACH	144085	150574.14
ST & BN	102944	76327.51
RBY	91636	78295.90
CHERRY	76687	71509.64
BLACK CHERRY	72318	70053.97
KLP	54799	38426.09
H-P	48376	34319.41
ORANGE CREME	40448	27511.36
MIXED BERRY	34821	30777.94
BLACKBERRY	32860	18749.18
KEY LIME	31984	32284.67
FRENCH VANILLA	30243	35124.93
PINEAPPLE	28711	31817.90
VERY CHERRY	25772	15692.10
BOSTON CREAM PIE	25114	14788.23
CHOCOLATE MOUSSE	22285	13557.54
BB PATCH	21675	14384.86
HONEY	21421	31535.62

Top 10 sales flavor



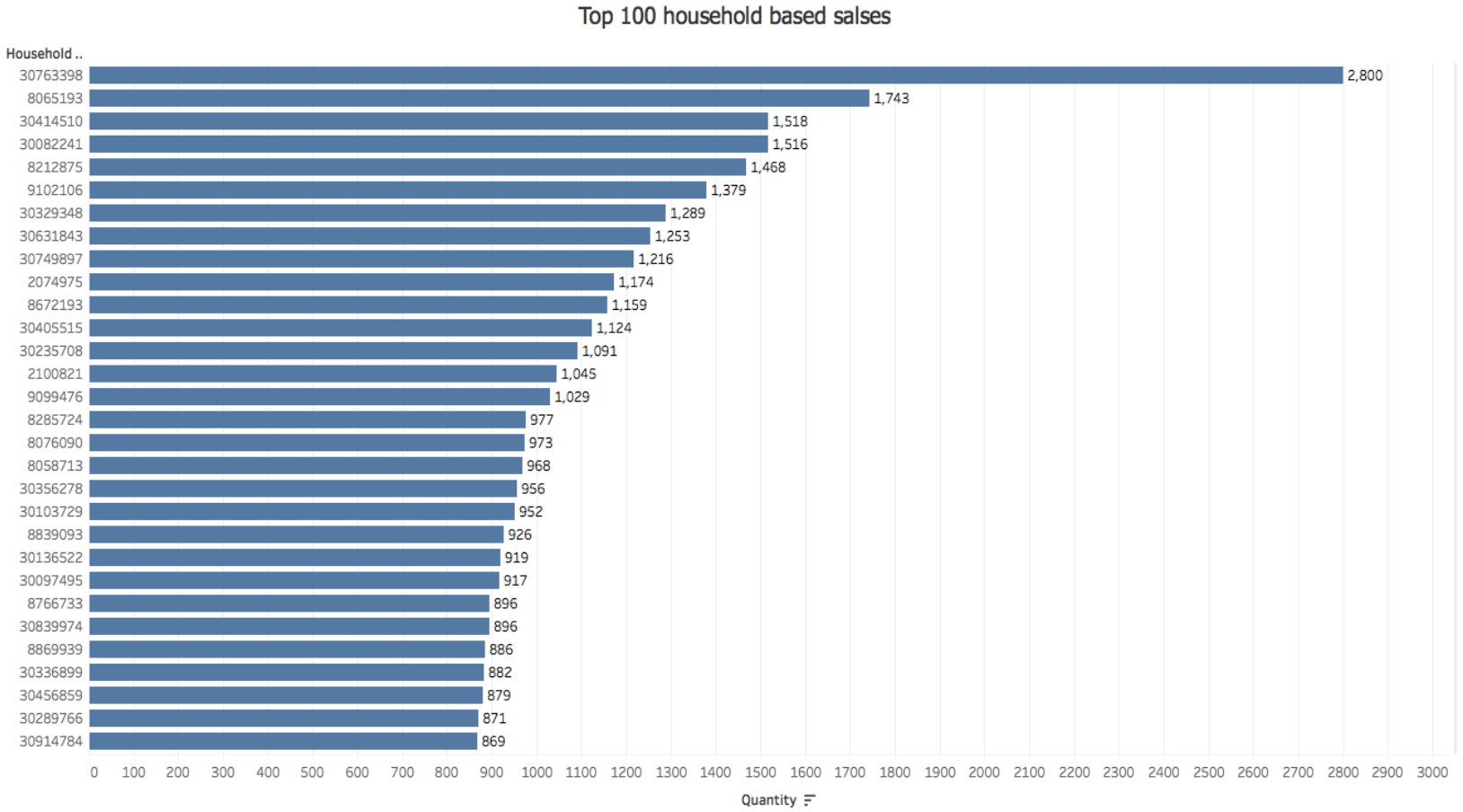
Data Exploration:

Monthly Sales



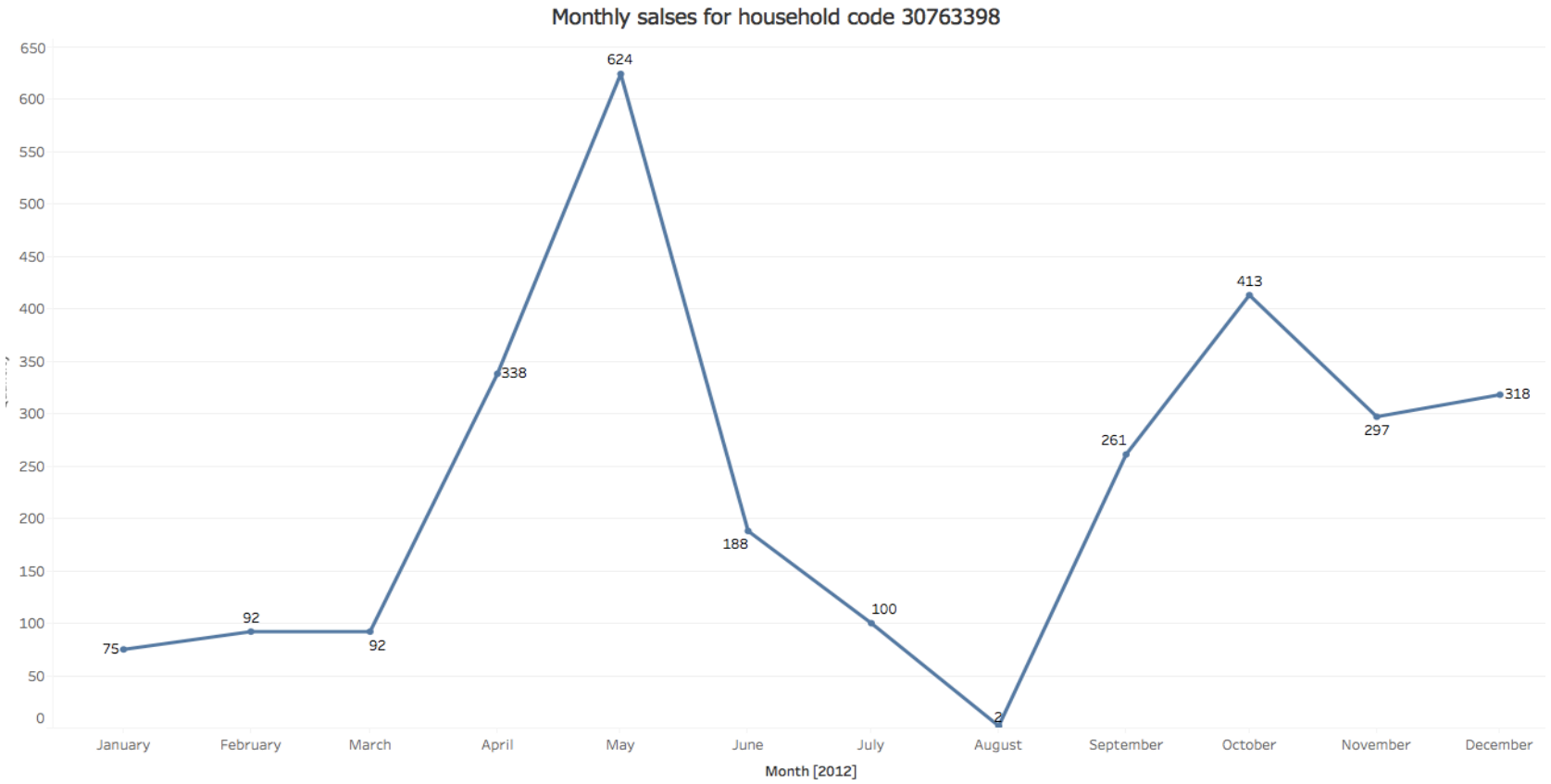
Data Exploration:

Household Sales



Data Exploration:

Household Sales



Favorite Type	Favorite Style	Favorite Flavor
Natural	Regular	ST

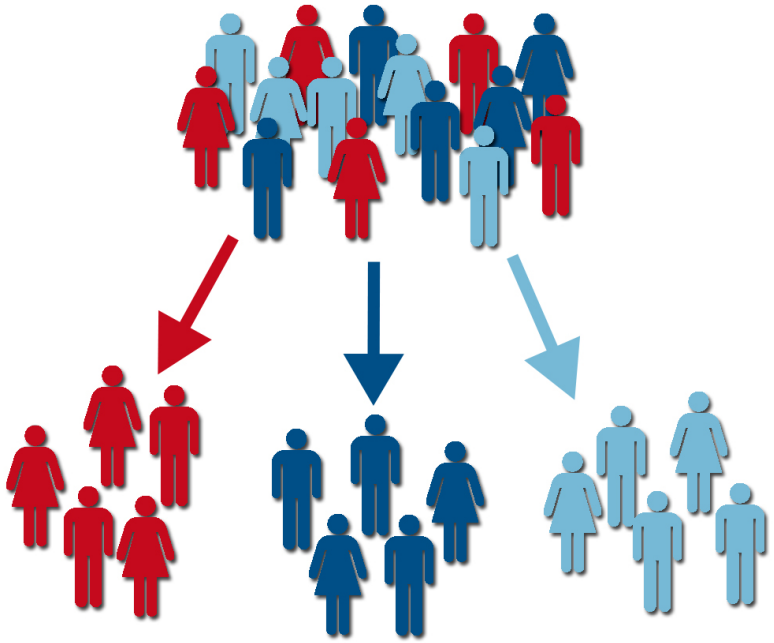
Customer Analysis –

Panelist Segmentation

LCA is selected to conduct the cluster analysis, most of the variables are categorical, 60,000+ rows data

Training and Holdout 70/30 split

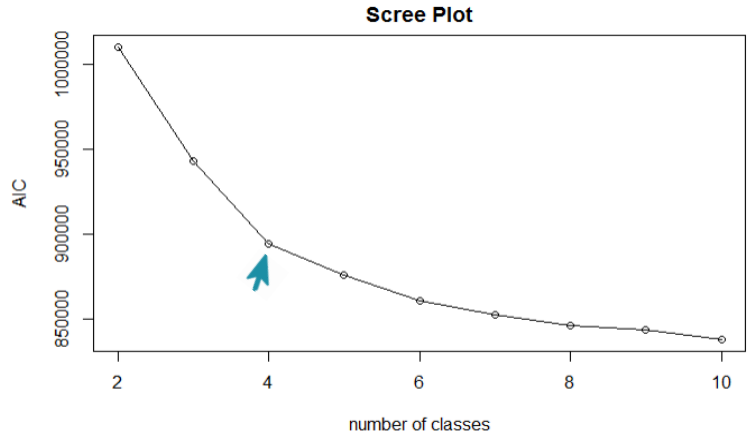
- ✓Household Income
- ✓Household Size
- ✓Type of Residence
- ✓Household Composition
- ✓Age and Presence of Children
- ✓Female/Male Head Age
- ✓Female/Male Head Employment

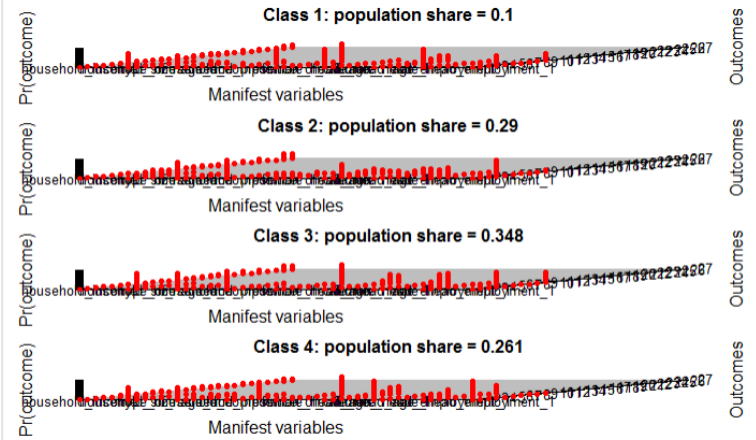


Customer Analysis:

Panelist Segmentation Result

4 clusters are selected due to scree plot and training and holdout comparison.






```
Estimated class population shares
 0.3484 0.29 0.1001 0.2615

Predicted class memberships (by modal posterior prob.)
 0.348 0.2904 0.1001 0.2615

=========================================================
Fit for 4 latent classes:
=========================================================
number of observations: 42379
number of estimated parameters: 367
residual degrees of freedom: 42012
maximum log-likelihood: -446754.9

AIC(4): 894243.7
BIC(4): 897419.9
G^2(4): 161999.8 (Likelihood ratio/deviance statistic)
X^2(4): 15886093195 (Chi-square goodness of fit)
```

```
Estimated class population shares
 0.2786 0.2597 0.1029 0.3589

Predicted class memberships (by modal posterior prob.)
 0.2774 0.2597 0.1029 0.3601

=========================================================
Fit for 4 latent classes:
=========================================================
number of observations: 18159
number of estimated parameters: 367
residual degrees of freedom: 17792
maximum log-likelihood: -190888.4

AIC(4): 382510.9
BIC(4): 385376
G^2(4): 80278.96 (Likelihood ratio/deviance statistic)
X^2(4): 8426141319 (Chi-square goodness of fit)
```

1st Cluster :

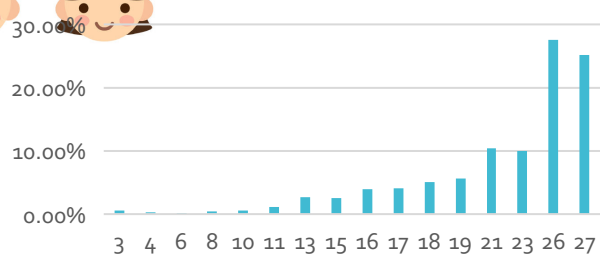
Customer Analysis:

Panelist Segmentation Interpretation

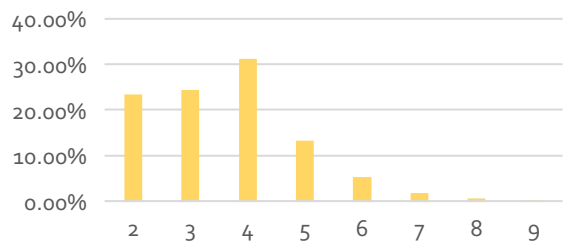
Class 1:

High household income group, 2-4 members household, one family house residence type, married couple at 44-54 years old with full-time job, most cases no children, or have kids 13-17 years old.

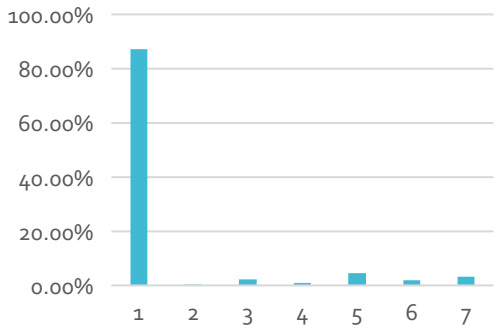
Household Income



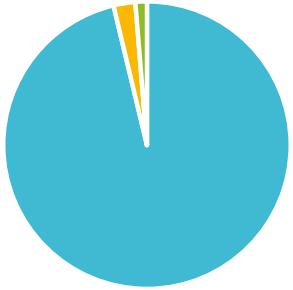
Household Size



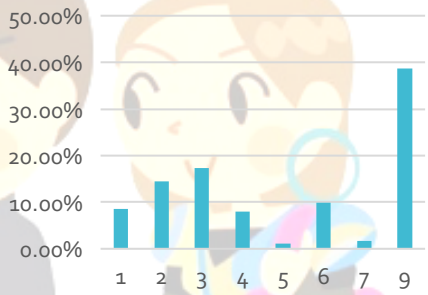
Type of Residence



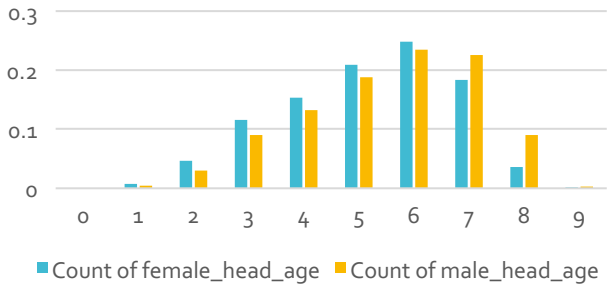
Household Composition



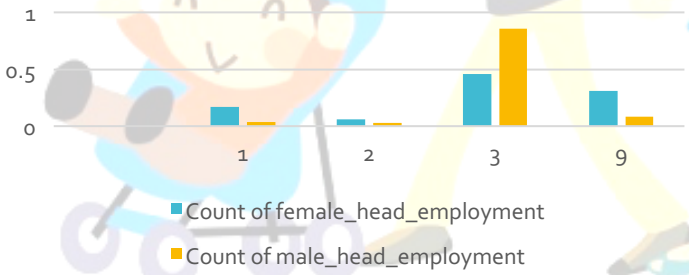
Age and Presence of Children



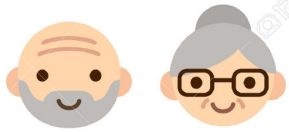
Female/Male Head Age



Female/Male Head Employment



2nd Cluster :



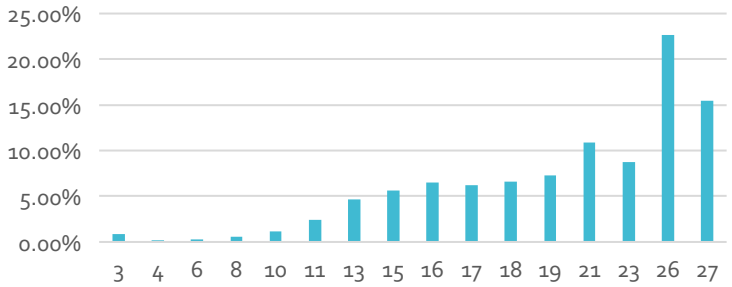
Customer Analysis:

Panelist Segmentation Interpretation

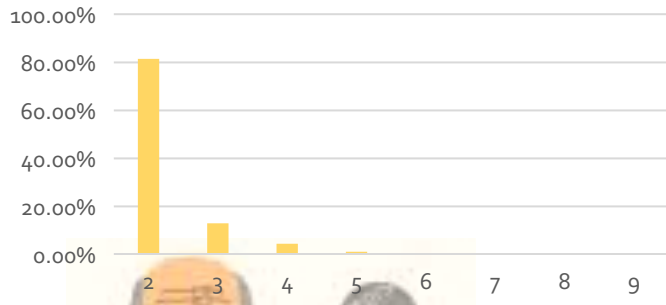
Class 2:

Household annual income 70k+, 2 members household, one family house, most cases no children, married couple, 55+ years old, the couple either works full time or retired.

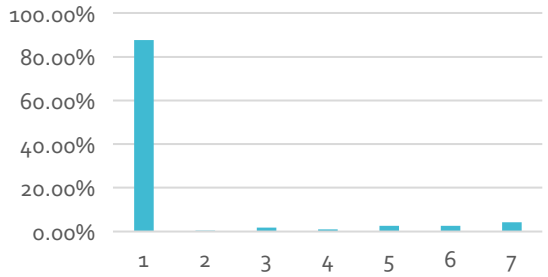
Household Income



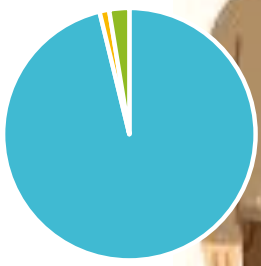
Household Size



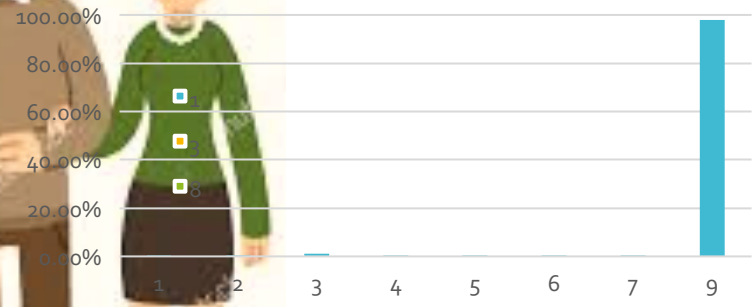
Type of Residence



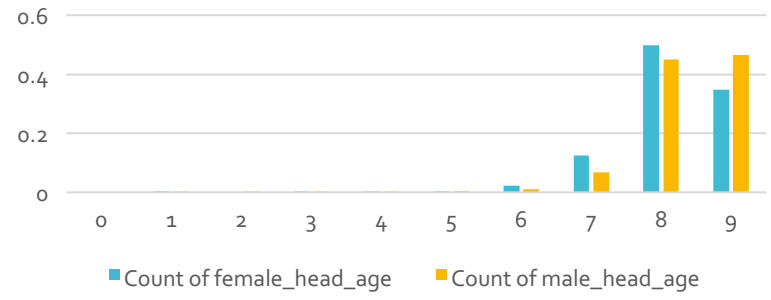
Household Composition



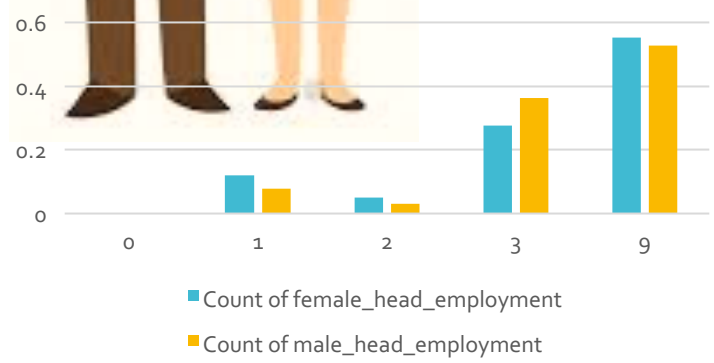
Age and Presence of Children



Female/Male Head Age



Female/Male Head Employment



3rd Cluster :

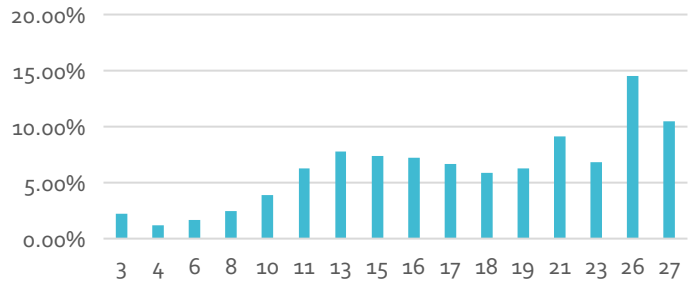
Customer Analysis:

Panelist Segmentation Interpretation

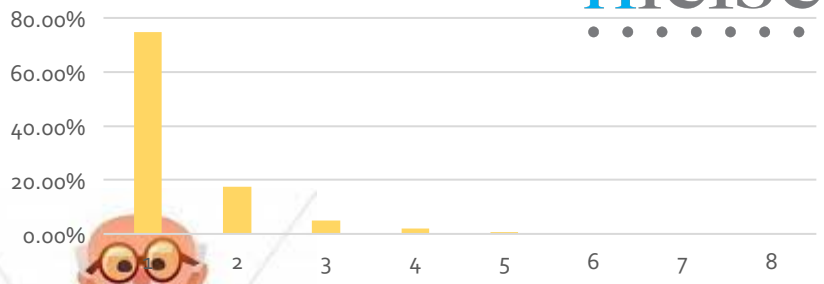
Class 3:

Household annual income mostly above 70k, but more spread out in distribution, most cases household size is 1, lives in one single family house, male living alone with no children at 55+ years old, with a full time job or retired.

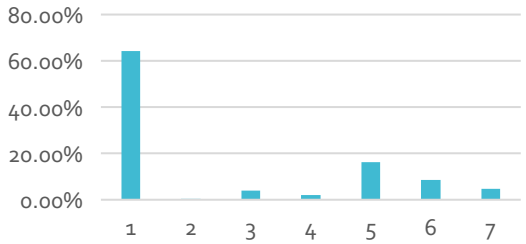
Household Income



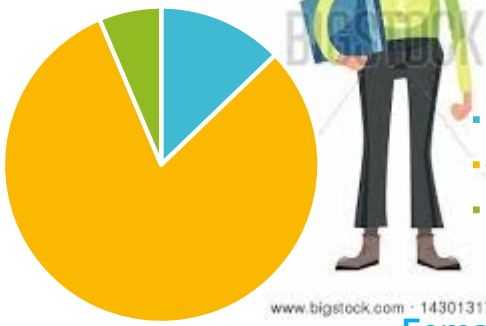
Household Size



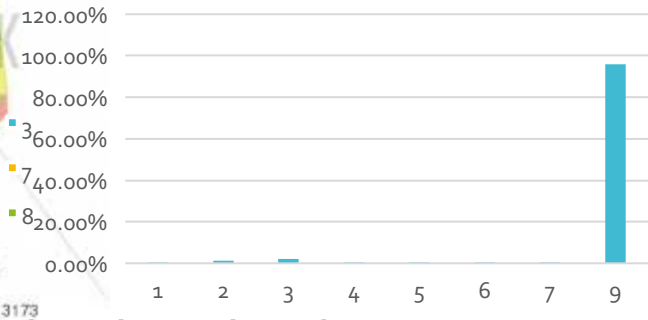
Type of Residence



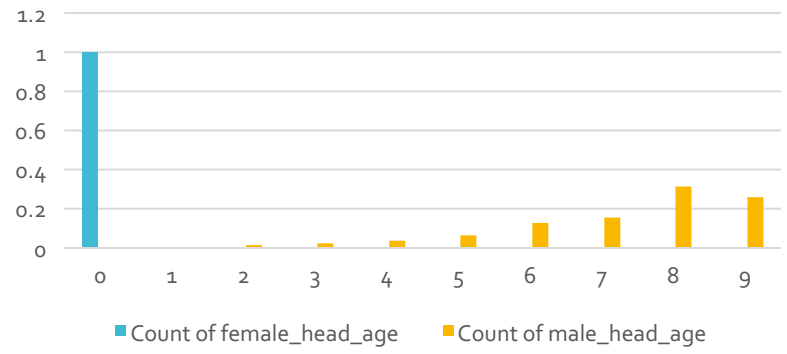
Household Composition



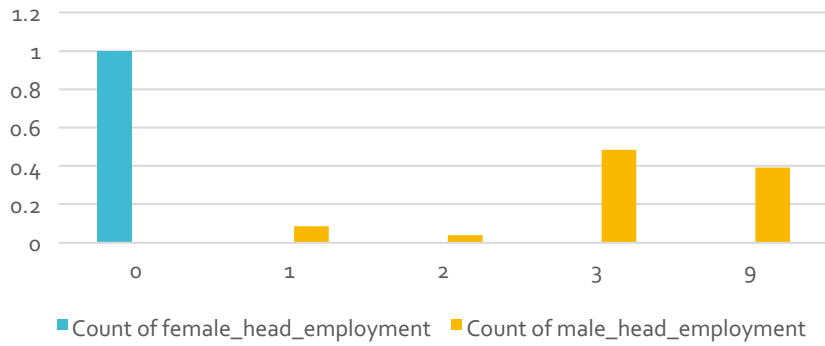
Age and Presence of Children



Female/Male Head Age



Female/Male Head Employment



4th Cluster :

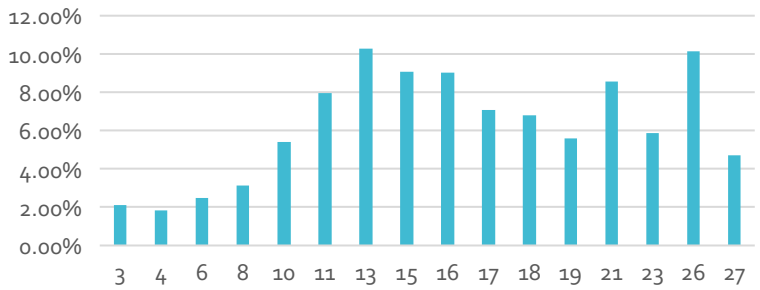
Customer Analysis:

Panelist Segmentation Interpretation

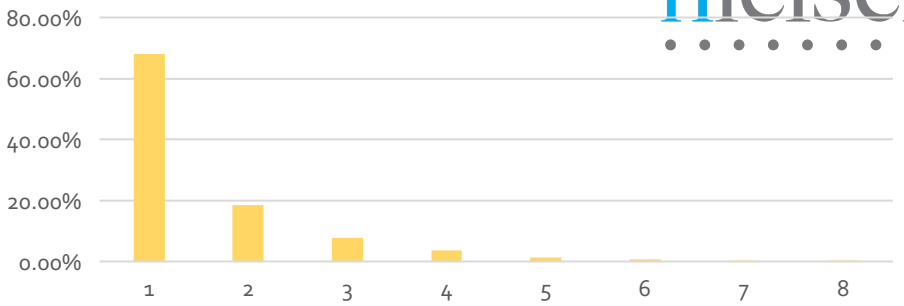
Class 4:

Household income level is very spread out, household size is mostly 1 lives in single family house, female living alone with no children, 55+ years old, working full time or retired.

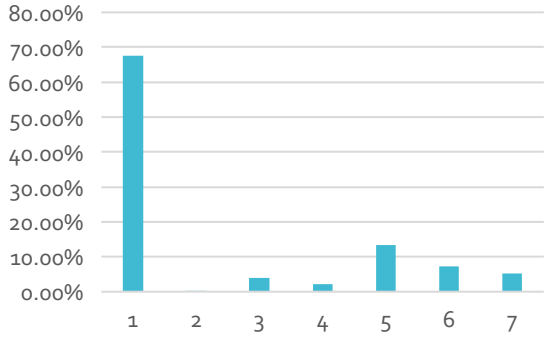
Household Income



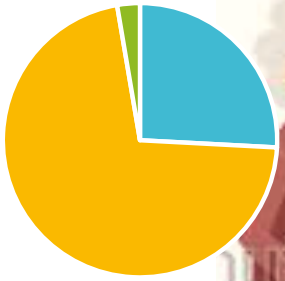
Household Size



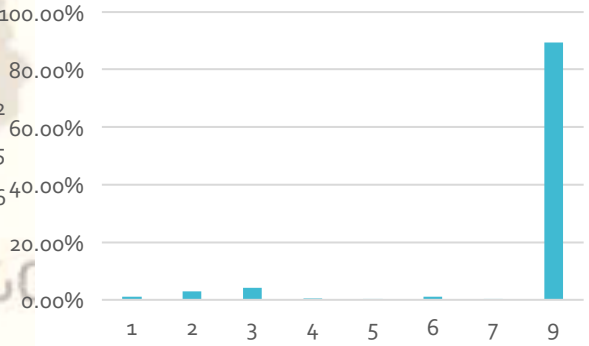
Type of Residence



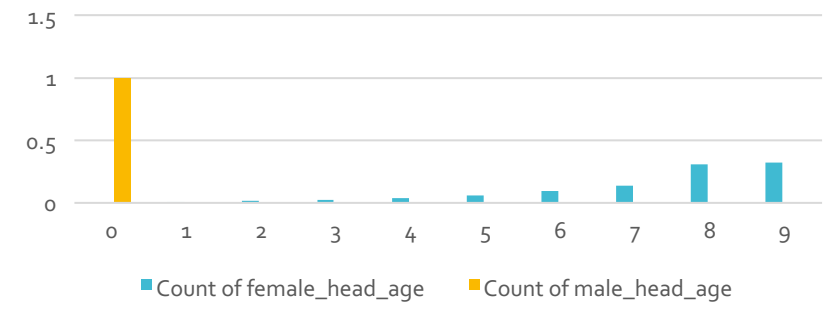
Household Composition



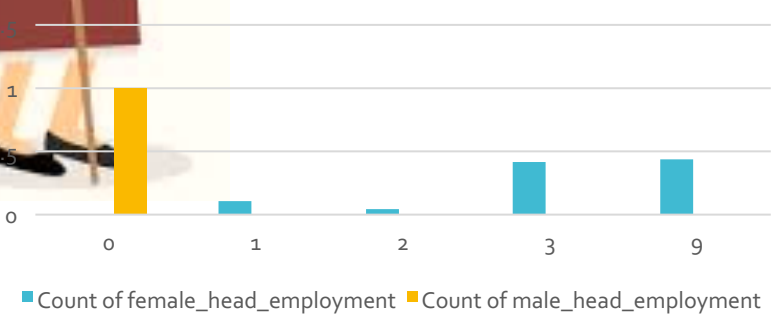
Age and Presence of Children



Female/Male Head Age



Female/Male Head Employment



Customer Analysis:

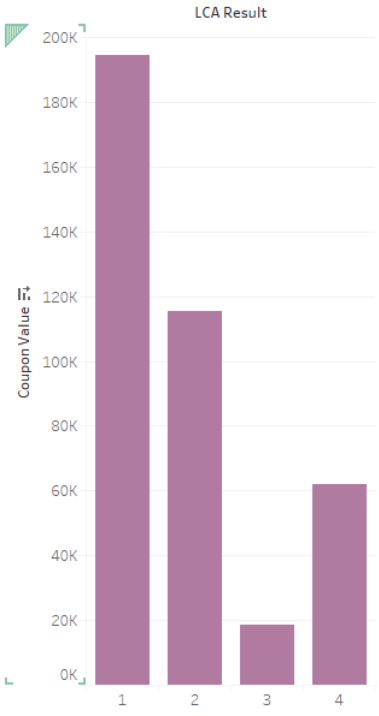
Panelist Segmentation Interpretation

LCA Result	
1	17,317
2	21,313
3	6,112
4	15,796

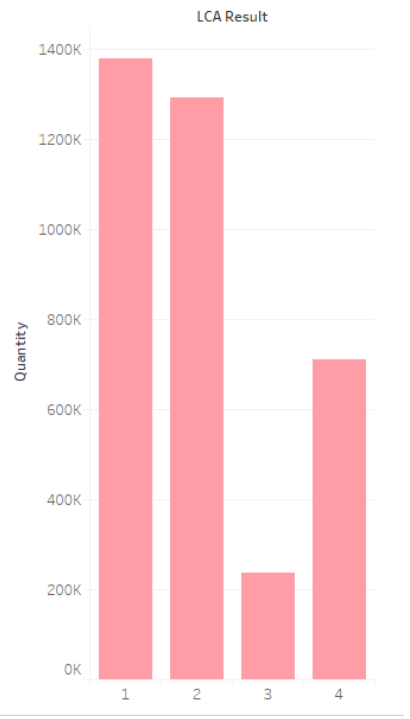
1st Cluster :

2nd Cluster:

Coupon Value for 4 Clusters



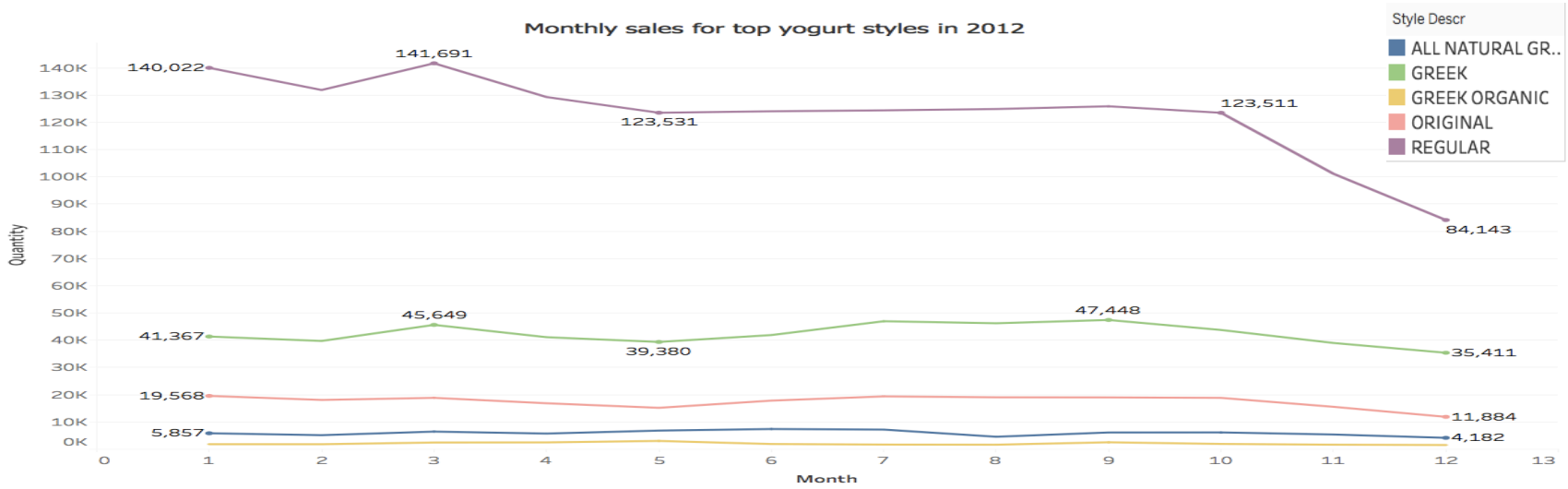
Quantity Purchased for 4 Clusters



Total Price Paid & Spent for 4 Clusters

Sales Analysis:

Monthly Sales



Overall Quantity Across All Products:

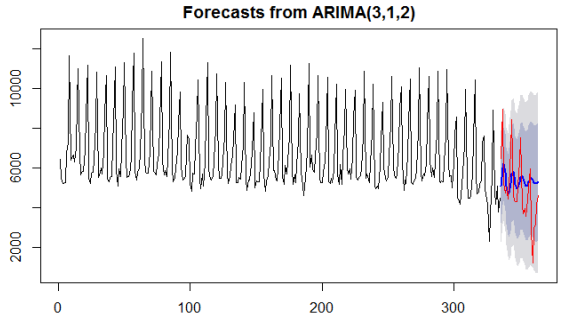
nielsen

Sales Analysis:

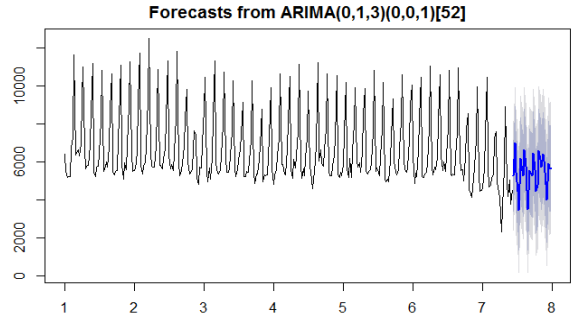
Time Series Modeling of Quantity Purchased

Training Dataset:
Daily Quantity Information from Jan to Nov, 2012, conduct weekly aggregation

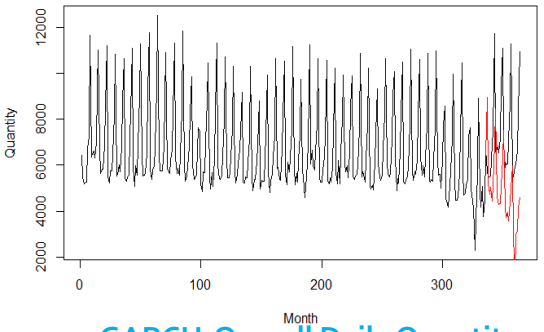
Holdout Dataset:
Daily Quantity Information for Dec, 2012 (29 days), conduct weekly aggregation



ARIMA-Overall Daily Quantity

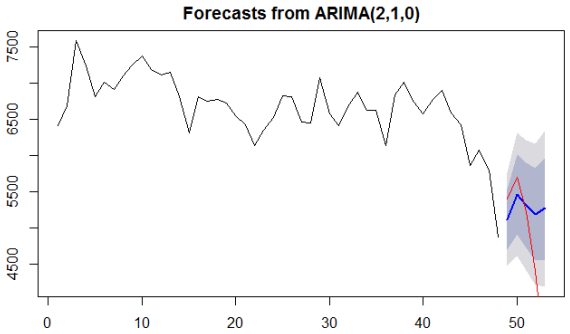


SARIMA-Overall Daily Quantity

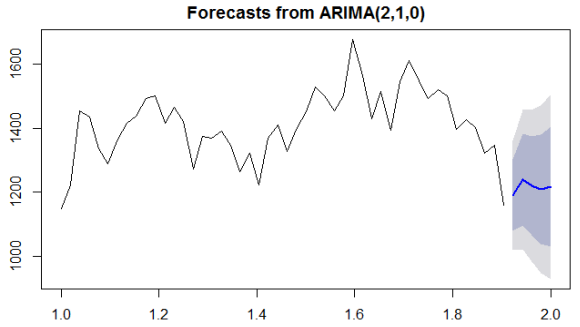


GARCH-Overall Daily Quantity

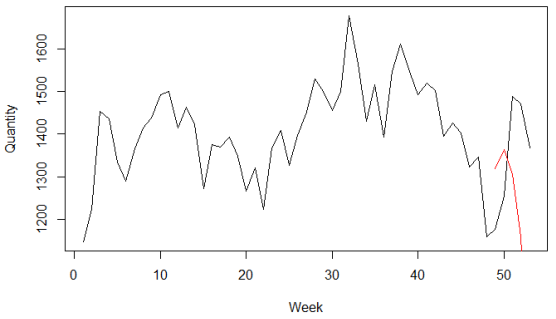
	ARIMA	SARIMA	GARCH
SMAPE	0.2970149	0.3108862	0.3709301



ARIMA-Overall Weekly Quantity



SARIMA-Overall Weekly Quantity



GARCH-Greek Weekly Quantity

	ARIMA	SARIMA	GARCH
SMAPE	0.1567993	0.1567993	0.4064851

Overall Quantity of Greek Yogurt:

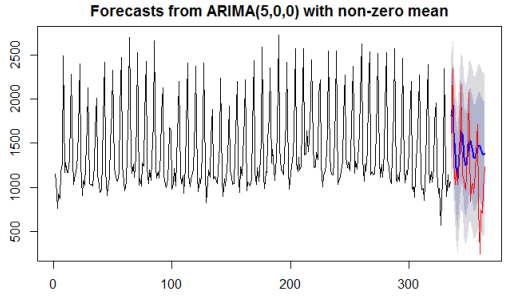
nielsen

Sales Analysis:

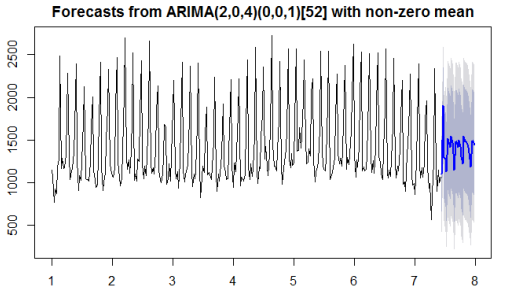
Time Series Modeling of Quantity Purchased

Training Dataset:
Daily Quantity Information from Jan to Nov, 2012, conduct weekly aggregation

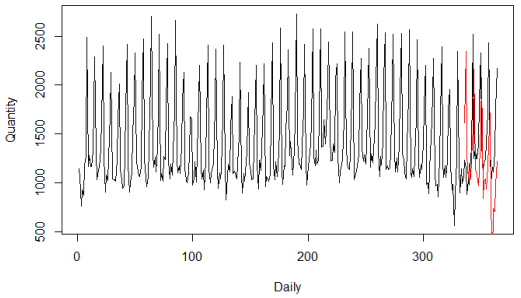
Holdout Dataset:
Daily Quantity Information for Dec, 2012 (29 days), conduct weekly aggregation



ARIMA-Greek Daily Quantity

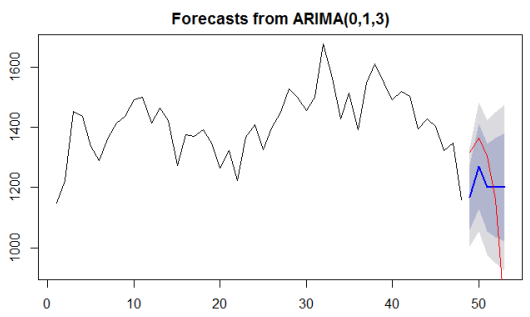


SARIMA-Greek Daily Quantity

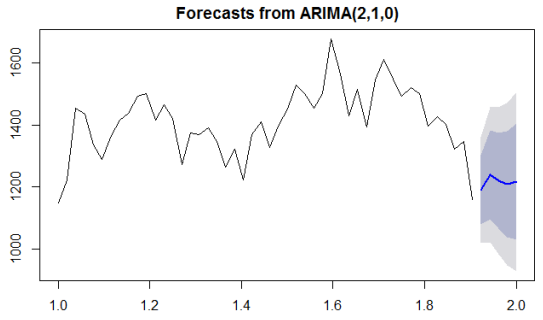


GARCH-Greek Daily Quantity

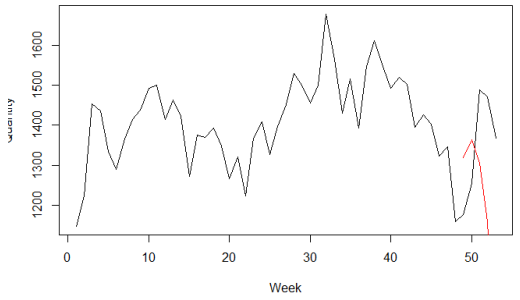
	ARIMA	SARIMA	GARCH
SMAPE	0.2771752	0.280436	0.4651712



ARIMA-Greek Weekly Quantity



SARIMA-Greek Weekly Quantity



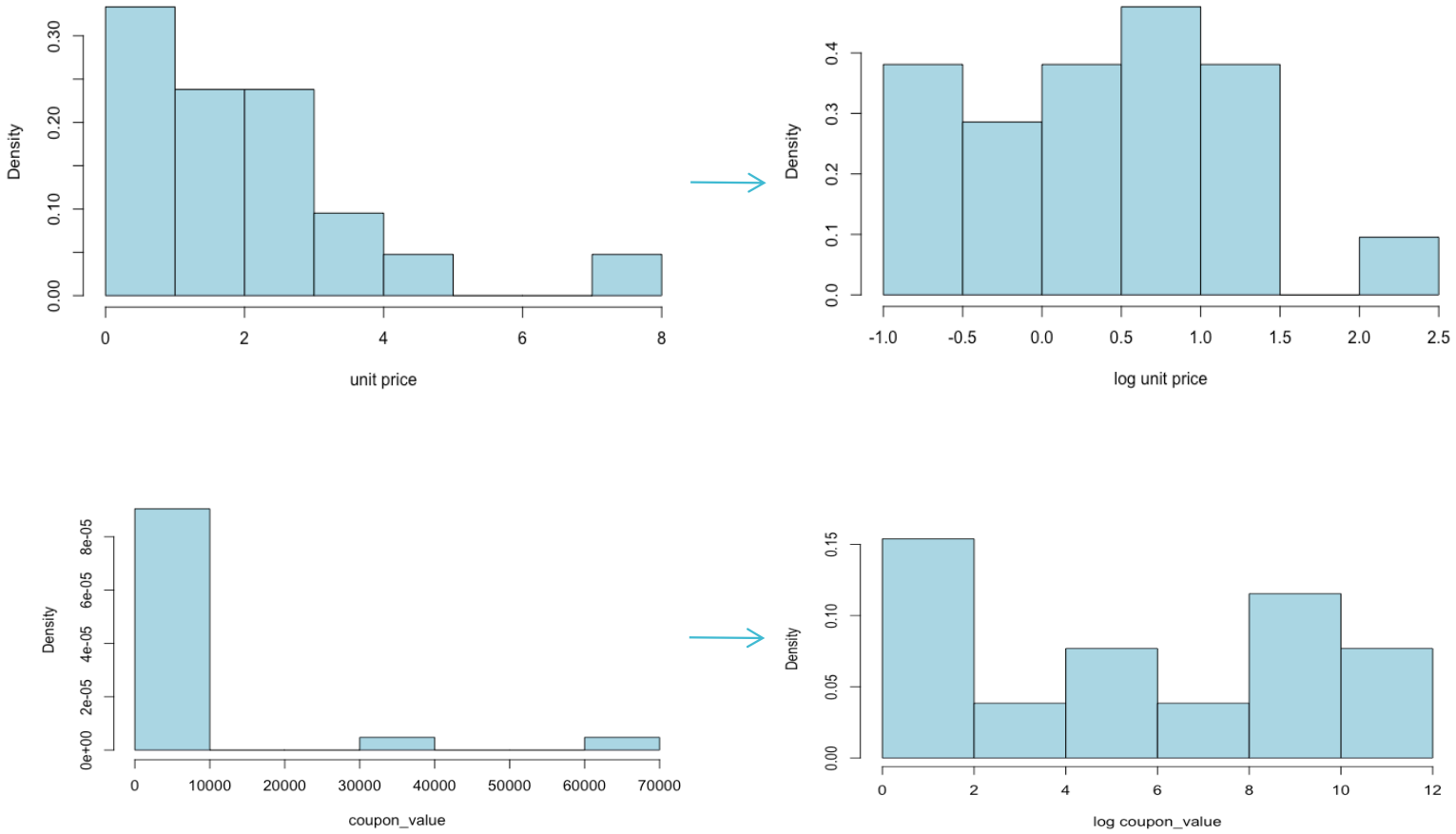
GARCH-Overall Weekly Quantity

	ARIMA	SARIMA	GARCH
SMAPE	0.1470364	0.1491714	0.2226936

Sales Analysis:

Sales Driver Analysis & Price Elasticity Analysis – Aggregate data by Yogurt Style

Marketing Mixed Models - Data Transformation



Sales Analysis:

Marketing Mixed Models – Aggregate data by Yogurt Style

Modeling and Result Analysis

m1 <- lm(log_quantity ~ log_unit_price+log_coupon_value,data=style_aggregated)

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.65704    0.42643  15.611 6.60e-12 ***
log_unit_price  -0.88641    0.46622  -1.901   0.0734 .
log_coupon_value 0.50053    0.04909  10.195 6.63e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.748 on 18 degrees of freedom
Multiple R-squared:  0.8587,    Adjusted R-squared:  0.8431
F-statistic: 54.72 on 2 and 18 DF,  p-value: 2.238e-08
```

m2 <- lm(log_quantity ~ log_unit_price+log_coupon_value+deal_flag_uc,data=style_aggregated)

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.4335 -1.2230 -0.2494  1.0806  2.9811

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.345e+00  4.297e-01  14.767 3.97e-11 ***
log_unit_price  -7.894e-01  4.378e-01  -1.803   0.0891 .
log_coupon_value 4.535e-01  5.194e-02   8.731 1.09e-07 ***
deal_flag_uc     1.427e-05  7.429e-06   1.921   0.0717 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.63 on 17 degrees of freedom
Multiple R-squared:  0.8839,    Adjusted R-squared:  0.8635
F-statistic: 43.16 on 3 and 17 DF,  p-value: 3.646e-08
```

Sales Analysis:

MCMC Models – Aggregate data by Yogurt Style

Modeling and Result Analysis

yogurt.mcmc <- MCMCregress(log_quantity1 ~ log_unit_price1 + log_coupon_value1,data=style_aggregated)
yogurt.mcmc2 <- MCMCregress(log_quantity1~ log_unit_price1 + log_coupon_value1+yogurt.agg$deal_flag_uc,data=style_aggregated)

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

	Mean	SD	Naive SE	Time-series SE
(Intercept)	6.344e+00	4.630e-01	4.630e-03	4.630e-03
log_unit_price	-7.884e-01	4.730e-01	4.730e-03	4.730e-03
log_coupon_value	4.531e-01	5.400e-02	5.400e-04	5.284e-04
deal_flag_uc	1.433e-05	7.802e-06	7.802e-08	7.802e-08
sigma2	3.006e+00	1.199e+00	1.199e-02	1.480e-02

2. Quantiles for each variable:

	2.5%	25%	50%	75%	97.5%
(Intercept)	5.444e+00	6.046e+00	6.343e+00	6.637e+00	7.280e+00
log_unit_price	-1.719e+00	-1.097e+00	-7.868e-01	-4.788e-01	1.357e-01
log_coupon_value	3.456e-01	4.180e-01	4.536e-01	4.888e-01	5.596e-01
deal_flag_uc	-1.091e-06	9.299e-06	1.438e-05	1.938e-05	3.003e-05
sigma2	1.490e+00	2.192e+00	2.757e+00	3.520e+00	5.964e+00

Sales Analysis:

Price elasticity– Aggregate data by Yogurt Style

Modeling and Result Analysis

```
> vif(yogurt.mcmc2)
   (Intercept)  log_unit_price log_coupon_value    deal_flag_uc        sigma2
      1.470939       1.275909       1.295205       1.479532       1.000202
```

The VIF test value for each variable is close to 1, which means the multicollinearity is very low among these variables. Based on the above analysis, we can accept the regression result and construct the multi-linear model of sales as follows:

log(Sales) = 6.344 -0.7884 * log(unit_price) + 0.4531* log(coupon_value) + 1.433e-05* deal_flag_uc

With model established, we can analysis the **Price Elasticity(PE)** predict the reactions of sales quantity to price. Price elasticity is defined as **%ΔQ/%ΔP**, which indicates the percent change in quantity divided by the percent change in price.

PE = (ΔQ/Q) / (ΔP/P) = (ΔQ/ΔP) * (P/Q)

Where P is the mean price of the data and Q is mean sales quantity

```
PE.unit_price<-
as.numeric(mean(yogurt.mcmc2[,2])*log(mean(YOGURT_purchase$unit_price))/log(mean(YOGURT_purchase$total_price_paid)))
PE.unit_price = -0.3696906
```

The PE indicates that 10% decrease in price will increase the sales by **3.7%**, and vice verse.

Sales Analysis:

Mixed Models– Greek Yogurt

Marketing Mixed Model for the GREEK yogurt

ml1 <- lm(log_quantity1 ~ log_unit_price1+log_coupon_value1)

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.075631 -0.033679 -0.004767  0.032913  0.105838

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.20608    0.64289  14.320 1.69e-07 ***
log_unit_price1  -0.73628    0.72076  -1.022  0.33368
log_coupon_value1  0.22546    0.06127   3.680  0.00508 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0559 on 9 degrees of freedom
Multiple R-squared:  0.6755,    Adjusted R-squared:  0.6034
F-statistic: 9.367 on 2 and 9 DF,  p-value: 0.006317
```

Sales Analysis:

Mixed Models– Greek Yogurt

Modeling and Result Analysis

ml2 <- lm(log_quantity1 ~ log_unit_price1 +log_coupon_value1+GREEK.agg$deal_flag_uc,data=GREEK.agg)

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             9.308e+00  5.009e-01  18.581 7.26e-08 ***
log_unit_price1         7.257e-02  6.389e-01   0.114   0.9124
log_coupon_value1       1.201e-01  6.223e-02   1.929   0.0898 .
GREEK.agg$deal_flag_uc 4.903e-05  1.865e-05   2.629   0.0302 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04343 on 8 degrees of freedom
Multiple R-squared:  0.8259,    Adjusted R-squared:  0.7606
F-statistic: 12.65 on 3 and 8 DF,  p-value: 0.002096
```

Similarly the VIF test value for each variable is close to 1, which means the multicollinearity is very low among these variables. So we can accept the regression result and construct the multi-linear model of sales as follows:

log(Sales) = 9.308 -0.07257 * log(unit_price) + 0.1201* log(coupon_value) + 4.903e-05* deal_flag_uc

The PE indicates that 10% decrease in price will increase the sales by 0.02%, and vice verse.

Sales Analysis:

Optimal Pricing and profit Prediction – Greek Yogurt

Modeling and Result Analysis

- We want to get higher profit rather than just higher sales quantity. So, how to set the optimal price for the new Greek yogurt to get the maximum profit based on the regression model above?
- To simplify the question, we can let the deal_flag_uc = 0, the unit_price = 1.531316 (mean value), and the coupon = 2503.361 (mean value).

log(Sales) = 9.308 -0.07257 * log(price) + 0.1201* log(2503.361)

= 10.24788 - 0.07257*log(price)

- Assume the marginal cost(C) per unit of yogurt is 0.25*1.531316. We can calculate the profit (Y) by the following formula:

Y = (price – C) * Sales Quantity

= (price-0.382829) * exp(10.24788 - 0.07257*log(price))

= (exp(10.24788)*x-10804.45)/1.075268*log(x)

Optimal price is **$1.9** and the maximum profit is **$29,419.72** per month

Analysis –

Optimal Coupon value and profit Prediction – Greek Yogurt

Modeling and Result Analysis

- To analyze the influence of coupon value, we can set the deal_flag_uc = 0, the unit_price = 1.531316 (mean value)

log(Sales) = 9.308 -0.07257 * log(1.531316) + 0.1201* log(coupon)

= 9.277076 + 0.1201* log(coupon)

Sales = exp(9.277076 + 0.1201* log(coupon))

Assume the marginal cost(C) per unit of yogurt is **0.25*1.531316**. We can calculate the profit (Y) by the following formula:

Y = (1.531316 – C) * Sales Quantity

= (1.531316-0.382829) * exp(9.277076 + 0.1201* log(coupon))

= 1.148496*exp(9.277076 + 0.1201* log(x))

Optimal coupon is about **$3,000** and the maximum profit is **$32,115.5** per month

Analysis –

Optimal promotion amount and profit Prediction – Greek Yogurt

Modeling and Result Analysis

To analyze the influence of promotion activity, we can set the unit_price = 1.531316 (mean value), and the coupon = 2503.361 (mean value).

log(Sales) = 9.308 -0.07257 * log(1.531316) + 0.1201* log(2503.361)+ 4.903e-05*deal

= 10.21691 + 4.903e-05*deal

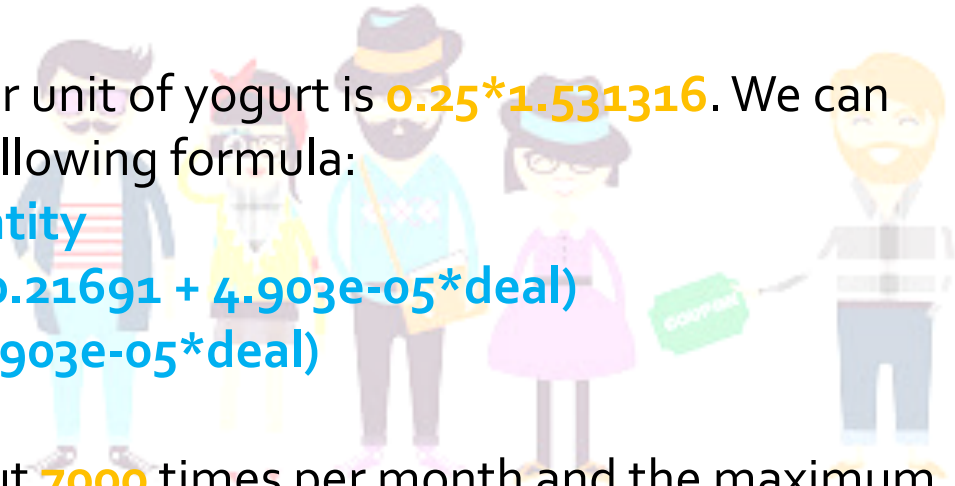
Sales = exp(10.21691 + 4.903e-05*deal)

Assume the marginal cost(C) per unit of yogurt is **0.25*1.531316**. We can calculate the profit (Y) by the following formula:

Y = (1.531316 – C) * Sales Quantity

= (1.531316-0.382829) * exp(10.21691 + 4.903e-05*deal)

= 1.148496*exp(10.21691 + 4.903e-05*deal)



Optimal promotion time is about **7000** times per month and the maximum profit is **$44,292.62** per month

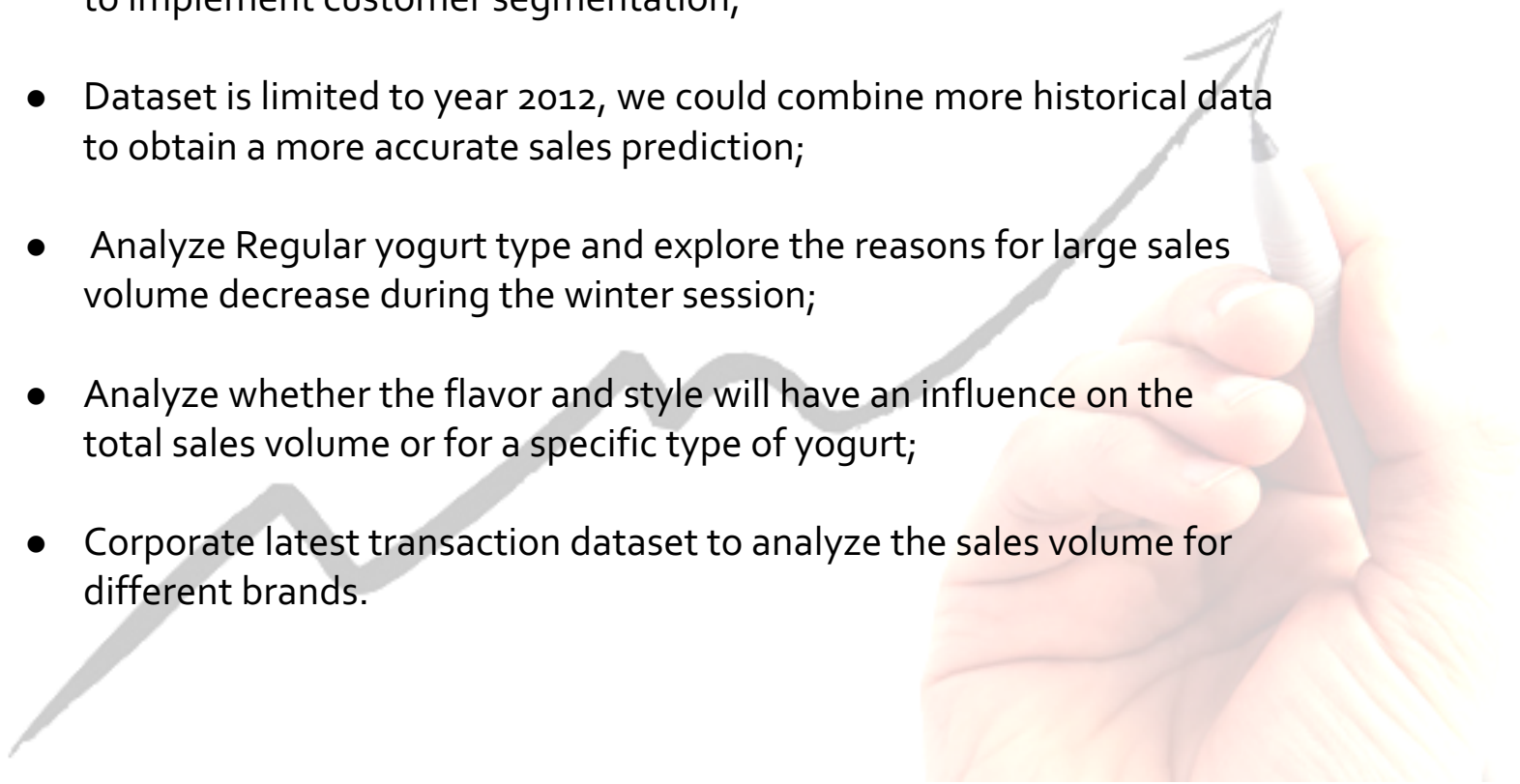
Summary of Findings

- Cluster 1 and 2 has the largest yogurt need. More coupon should be distributed to Cluster 2.
- ARIMA has the best prediction performance for both overall yogurt sales amount and Greek yogurt sales volume.
- Monthly total yogurt sales volume is comparatively during spring and summer, it peeks at March, but will decrease significantly during the winter season.
- Generally speaking, yogurt price does not have a significant influence on the sales of yogurt, instead coupon and promotion has a relatively significant impact on the sales quantity.
- Specifically, coupon has a larger influence on the total yogurt sales, while promotion has a larger impact on the Greek yogurt sales.

Price	Coupon	Promotion	Max Profit
1.9	$2,503	0	$29,419.72
1.53	$3,000	0	$32,115.51
1.53	$2,503	7000	$44,292.62

Further improvements

- Panelist might not be representative of the real customer group, young people group is missing in the Panelist;
- Corporate other variables such as product unit price, coupon amount to implement customer segmentation;
- Dataset is limited to year 2012, we could combine more historical data to obtain a more accurate sales prediction;
- Analyze Regular yogurt type and explore the reasons for large sales volume decrease during the winter session;
- Analyze whether the flavor and style will have an influence on the total sales volume or for a specific type of yogurt;
- Corporate latest transaction dataset to analyze the sales volume for different brands.



Q&A

Thank you