James Lee
Jet.com Case Analysis
JL985@Cornell.edu
5/12/17

USING K-MEANS CLUSTERING ALGORITHM TO ANALYZE PRICING COMPETITIVENESS OF VENDORS ON A SUBCATEGORICAL LEVEL

Abstract: Case study presents three different vendors, each with its own unique prices for various products. Initial analysis reveals that different vendors have competitive advantages in different categories of products. Further Investigation reveals the presence of sub-clusters of products within each category. I investigate the use case of K-Means clustering algorithm using R to further group the data into subcategories. Result shows that Jet can further optimize cost savings by choosing vendors based on the algorithm's assignment.

Data format:

Case presents price data of roughly 500 products. Some examples of the data are shown below.

Merchant	Product SKU ID	Category	Price	Jet Sales
Jasmine's	J0001	Electronics	4.46	157
Alex's	J0001	Electronics	3.79	157
Leo's	J0001	Electronics	3.65	157

Measuring Vendor Competitiveness:

For each vendor and product, we measure its percentage price deviation from the average price. Essentially, this tells us, on average, how much % cheaper or more expensive vendor's prices are from its peers' prices.

$$For\ each\ Merchant,\quad \frac{1}{\#\ of\ Products}\sum_{i=1}^{\#\ of\ Products}\frac{Merchant's\ Price\ for\ Product\ i - Average\ of\ All\ Merchants'\ Prices\ for\ Product\ i}{Average\ of All\ Merchants'\ Prices\ for\ Product\ i}$$

We then scale these measures using sales * price to essentially credit the vendor if it offers cheap prices in products that are responsible for higher proportion of Jet's costs. Weighted average calculation is done within R using ddply package.

Result: (-x% indicates that the vendor offers, on average, x% cheaper prices)

Merchant	Average Price Deviation	Rank
Alex's Store	-2.4%	2
Jasmine's Shop	-4.4%	1
Leo's Bodega	6.9%	3

Merchant	Weighted Average Price Deviation	Rank
Alex's Store	.45%	3
Jasmine's Shop	7.2%	2
Leo's Bodega	-5.0%	1

Tables show us that while Jasmine's offers cheapest prices on an arithmetic average basis, Leo's offers the best prices on a weighted average basis. In fact, their results are dramatically different based on the approach.

Assessing Vendors' Categorical Competitive Advantages:

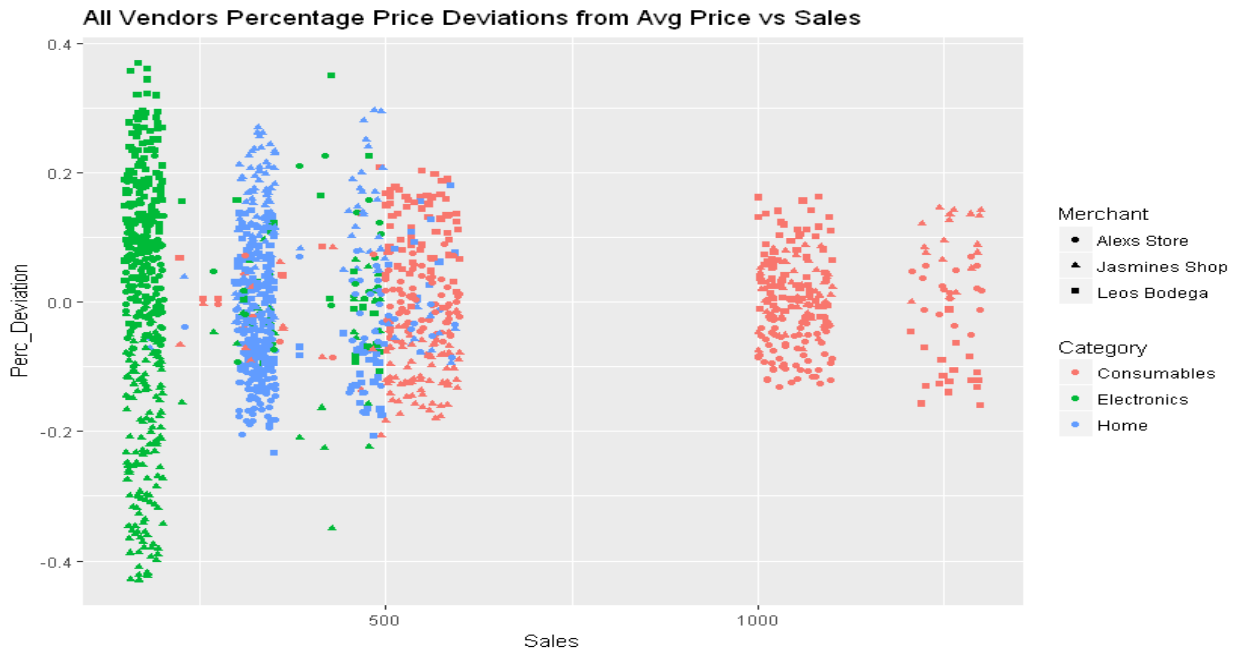
We can segregate the data into different categories and run the same analysis to investigate categorical competitive advantages of the vendors.

Category	Most Competitive Merchant	Mean	Weighted Mean
Consumables	Leo's Bodega	6.3%	-3.4%
Electronic	Jasmine's Shop	-19.0%	-4.8%
Home	Leo's Bodega	-.73%	-11.6%

If Jet can choose vendors selectively based on category, using the above output would not be a bad idea. However, further analysis shows that we can do better.

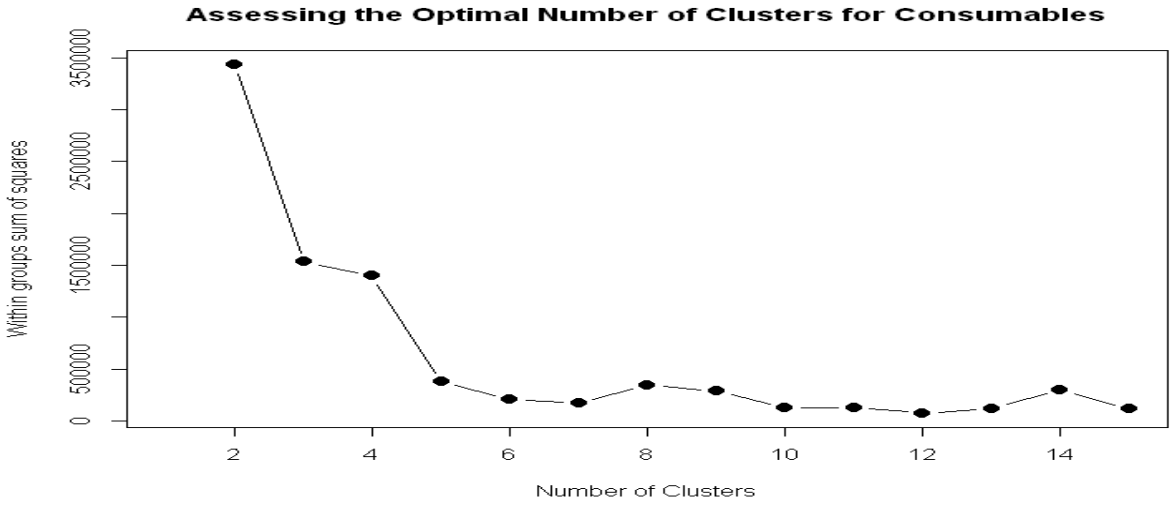
Applying K-Means Clustering Algorithm to Identify Subcategories:

Below scatterplot of price deviations vs sales, reveals that there are visually apparent clusters of products within the categories.



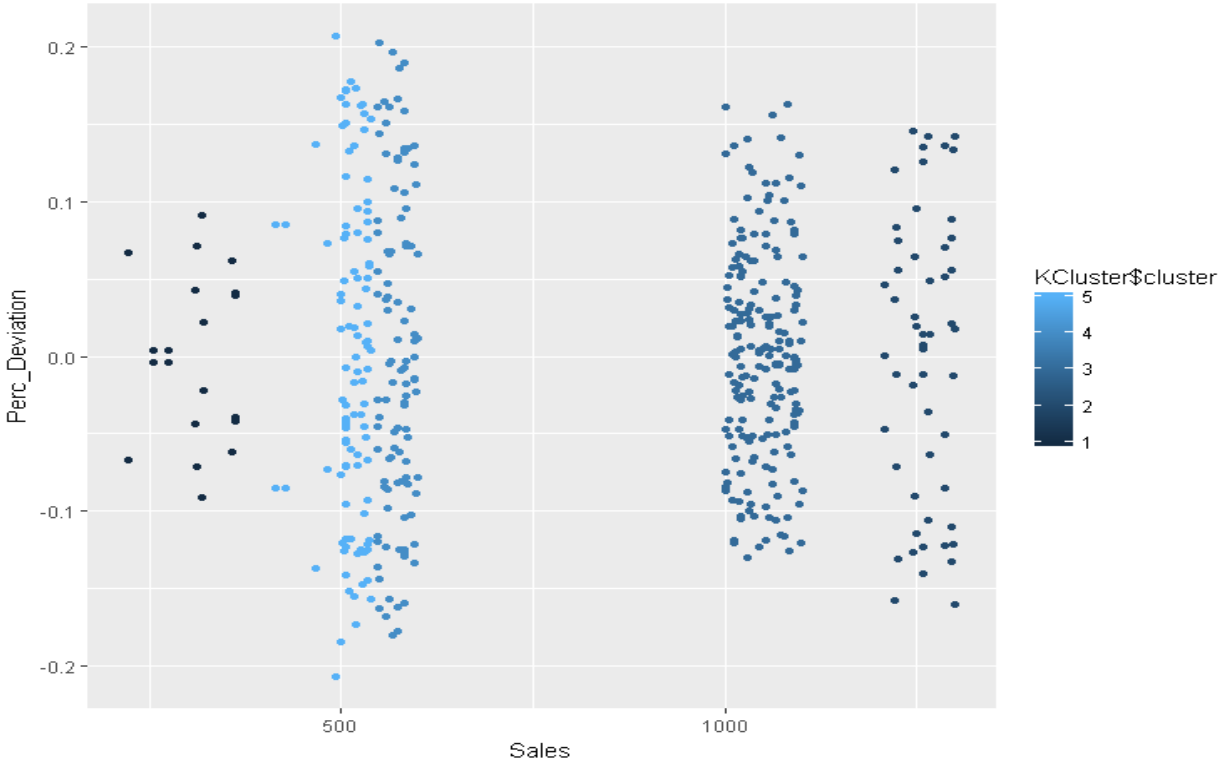
We can assign these observations into subsets using K-Means algorithm, which is a machine learning approach that attempts to assign observations based on some criterion. In this case, we will be using the level of Sales as a criterion.

For this paper, we will talk about the Consumables category (red dots) as it has the most interesting patterns. First, we determine the number of clusters to use by plotting the sum of squares, a measure of modeling error, vs different numbers of clusters as shown below. Tighter the fit, lower the sum of squares, will be. However, simply fitting more clusters will almost always create a better fit. Therefore, we choose the number of clusters where it is apparent that the marginal benefit of increasing the clusters doesn't merit the risk of overfitting the data.



As shown, marginal benefit of adding clusters dramatically falls off after 5 clusters. Therefore, we use 5 clusters.

Consumables data grouped into five clusters,



We segregate the data into clusters based on the algorithm assignments and run our original analysis again.

Category	Cluster Center	Most Competitive	Least Competitive
Consumables	318	Alex's	Jasmine's
Consumables	514	Jasmine's	Leo's
Consumables	572	Jasmine's	Leo's
Consumables	1048	Alex's	Jasmine's
Consumables	1260	Leo's	Jasmine's

Result is rather surprising. While it originally seemed that Leo's offers superior prices in Consumables, it only excels in one of the five clusters within Consumables. We can see that Jet can lose out on a lot of cost savings if it merely chooses vendors on a category basis. Below is a summary of the results from the K-Means Clustering analysis. My recommendation for this case is that Jet would choose the most competitive vendor in each of these clusters.

Category	Sales Cluster	Most Competitive Vendor	Least Competitive Vendor
Electronics	<=200	Jasmine's	Leo's
Electronics	>200	Leo's	Alex's
Home	<=350	Alex's	Jasmine's
Home	>350	Leo's	Jasmine's
Consumables	<500	Alex's	Leo's
Consumables	500-600	Jasmine's	Leo's
Consumables	1000-1100	Alex's	Jasmine's
Consumables	>1100	Leo's	Jasmine's