Case Analysis Studies
Of
Diffusion Models on E-Commerce Transaction Data

By
Taariq Lewis
A.B. Economics-Philosophy
Columbia University, 1996

By
Bryan Long
B.S. Electrical Engineering
Northeastern University, 2002

Submitted to the MIT Sloan School of Management

In partial fulfillment of the requirements for the degree of

Masters of Business Administration

At the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2009

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
JUN 10 2009
LIBRARIES

ARCHIVES

© 2009 Taariq Lewis and Bryan Long. All rights reserved

The authors hereby grant to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.

Signature of Author:_____
Taariq Lewis
MIT Sloan School of Management
May 8, 2009

Signature of Author:_____
Bryan Long
MIT Sloan School of Management
May 8, 2009

Certified By:_____
Vivek Farias
Operations Management/System Dynamics
Thesis Advisor

Accepted By:_____
Debbie Berechman
Executive Director, MBA Program
MIT Sloan School of Management

Case Analysis Studies
Of
Diffusion Models on E-Commerce Transaction Data

By

Taariq Lewis and Bryan Long

Submitted to the MIT Sloan School of Management on May 8, 2009

In partial fulfillment of the requirements for the degree of

Masters of Business Administration

Abstract:
As online merchants compete in the growing e-commerce markets for customers, attention to data generated from merchant and customer website interactions continues to drive ongoing online analytical innovation. However, successful online sales forecasting arising from historical transaction data still proves elusive for many online retailers. Although there are numerous software and statistical models used in online retail, not many practitioners claim success creating accurate online inventory management or marketing effectiveness forecast models. Thus, online retailers with both online and offline strategies express frustration that although they are able to predict sales in their offline properties, even with substantial online data, they are not as successful with their online-stores.

This paper attempts to test two analytical approaches to determine whether reliable forecasting can be developed using already established statistical models. Firstly, we use the original Bass Model of Diffusion and modify it for analysis of online retail data. Then, we test the model's forecasting effectiveness to extrapolate expected sales in the following year. As a second method, we use statistical cluster analysis to categorize groups of products into distinct product performance groups. We then analyze those groups for distinct characteristics and then test whether we can forecast new product performance based on the identified group characteristics. We partnered with a medium-sized online retail e-commerce firm with both online and offline retail channels to provide us with online transaction data.

Using a modified Bass Diffusion Model, we were able to fit a sales forecast curve to a sample of products. We then used k-means cluster analysis to partition products into similar groups of sales transaction-behavior, over the period of 1 year. For each group, we tried to identify characteristics which we could use to forecast new product launch behavior. However, lack of accurate, characteristic mapping of products made it difficult to establish confidence in cluster forecasting for some groups with similar curves. With more accurate characteristic mapping of products, we're hopeful that cluster analysis can reasonably forecast new product performance in online retail catalogs.

Thesis Supervisor: Vivek Farias
Title: Assistant Professor, MIT Sloan School of Management

Table of Contents

Table of Figures

Table of Tables

Introduction

We think it will be helpful to provide background information on the industry we chose. For the purposes of this thesis, we make the following assumptions about the marketplace and provide background information on the current offerings. These offerings are geared towards online retail and are centered on the study and research of analysis of transaction data. This is by no means an exhaustive list and excludes customized solutions and other niche offerings that may be offered currently.

What is Online Retail

Online Retail may be defined as all business to consumer transactions that are conducted via the Internet or through some electronic medium that substitutes for a consumer to visit a physical store or use of a telephone to place an order transaction for a product or service.

According to the US Online retail sales growth through 2010 was expected to show growth at anywhere from 10% to 15% with total online retail sales projected to reach $165 Billion by 2009.[1]

Figure 1 shows the historical growth of online retail in the United States in 13 categories ranging from Computer Hardware to Office Products.

As merchants compete in the growing e-commerce markets for customers, attention to data generated from merchant and customer website interactions are important to helping these companies understand their business competitive positioning. Out of attending to this need for data-driven intelligence, new industries have sprung up with companies that provide various analytical services to measure effectiveness.

What is Online Marketing

Online marketing consists of the various methods that are used by merchants to call consumers to act on an e-commerce site. Online marketing may consist of web advertising, email advertising, online merchandizing and product promotions. Online marketing can also consist of using offline or traditional media to push customers into the online store. For example, traditional retailers may use catalogs or on-air advertising to encourage customers to purchase online.

[1] Source: Jupiter Research, Inc., New York, NY, unpublished data (copyright)

What is E-Commerce Analytics

A by-product of internet traffic and e-commerce transaction on any retail website is the storage of data, via weblogs or databases, of each e-commerce transaction that occurs. E-Commerce analytics consist of the various methods and tools that are used to measure customer interaction with an online property. E-commerce transaction can thus include any type of customer interaction with an online store.

Companies provide various analytical services to measure effectiveness in such areas as:

- Site Content
    - Analyzing website logs, tags, and cookies to understand how consumers interact with website content including links, images, videos, etc. This can be useful for improving site usability and effectiveness. In addition, web-masters and designers use web-log analysis to verify web-link integrity and site quality.

- Path Analysis
    - Analyzing website logs, tags, cookies and back-links to track customers as they traverse to and then through an online store. Path analysis can help online properties understand how customers interact with the site. It can be used to answer such question as: "How do customers get to the site?" Or "Where are customers going on the site and how does that determine the consumer online shopping experience?" Path analysis also includes site-wide tracking to add another level of understanding.

- Online advertising
    - Analysis of advertising effectiveness across online store and website. Click-through rates and conversion rates are used to measure return on advertising dollars spent. It can be used between online and offline advertising campaigns to understand trends.

- Business Rules
    - Data-driven website content updates that are based on threshold triggers that are pre-defined and usually static. Based on a user's entry into a website, dynamic content will be delivered to the online shopper. Thus, customers who click a link through an advertising campaign may see a different layout than an online shopper arriving via an Internet search engine.

- Merchandizing
    - Data-driven promotions to drive consumer behavior to meet certain e-commerce goals. Online merchants may use A/B testing, or "split testing". These tests help merchants test a small sample of online consumer behavior against a baseline measurement.[2]

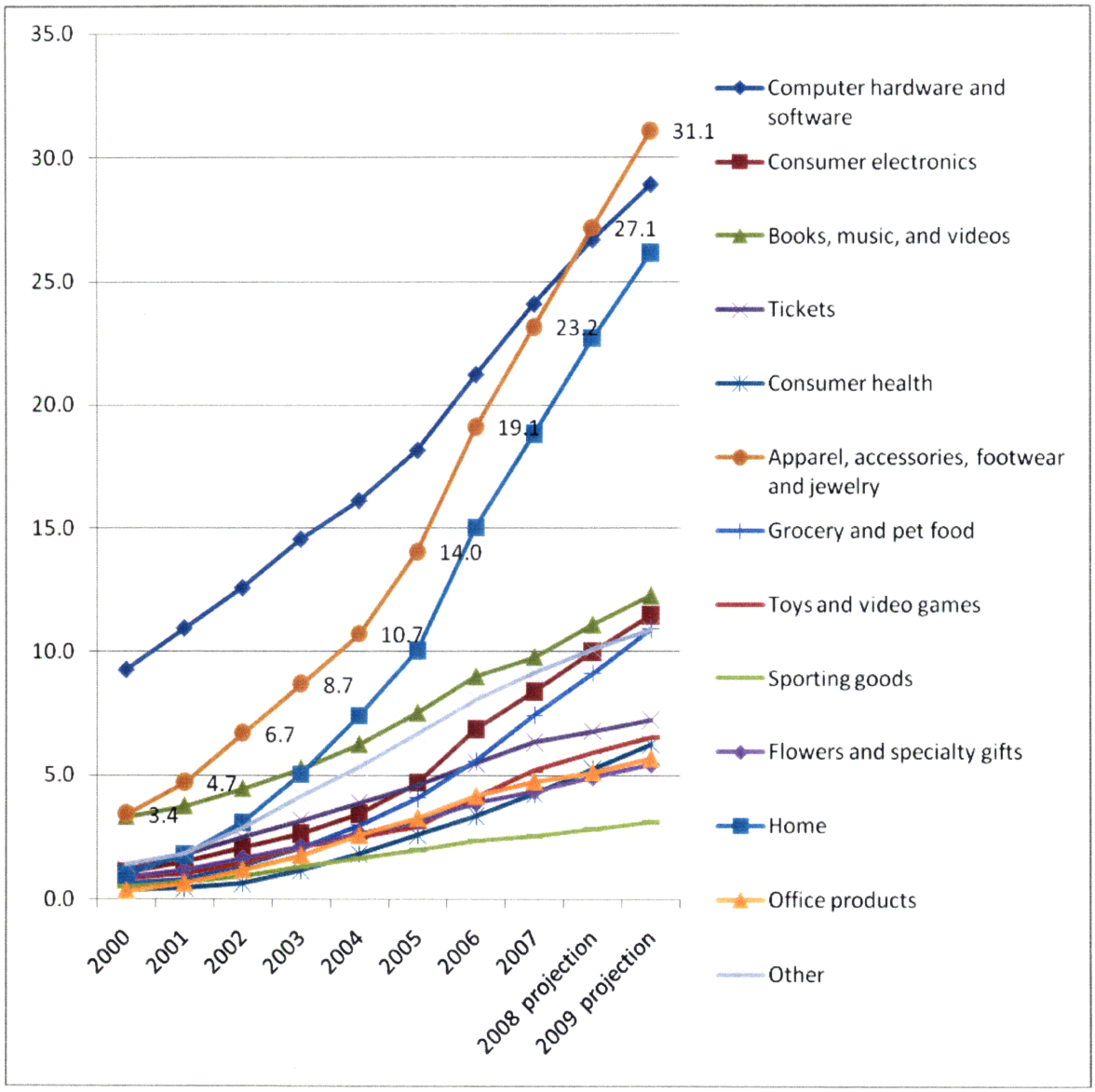

Figure 1: E-Commerce trends by product[3]

[2] Lewis, 2007

Diffusion Models:

Diffusion models are equations that are typically used to describe and model the process by which innovation "is communicated through certain channels over time among members of a social system." Diffusion models were first introduced into marketing-science during the 1960s.(Bass, 1969)

What is the history of the tools?

In the late 1970s, Mahajan and Muller (Muller, 1979) described diffusion models as an attempt to describe a theoretical framework that could accurately model how innovation spreads through a population of adopters. "The purpose of the diffusion model is to depict successive increases in the number of adopters and predict the continued development of a diffusion process already in progress (Christopher J. Easingwood, 1983)." These models grew out of a need to attempt to forecast the progress of innovation[4]. The view was that once an innovation was introduced into a market, the innovation would purchased by a set of early adopters. These early adopters would use the product, and then by word of mouth, would inform others of the values and benefits. The "diffusion" of the value, benefits and other product information would influence the uptake of the innovation of product over time. One of the popular equations used to model these phenomena was known as the Bass Model.

[3] U.S. Census Bureau, Statistical Abstract of the United States: 2009 (128th Edition) Washington, DC, 2008; http://www.census.gov/statab/www/

[4] Easingwood Mahajan and Muller 1983

The Bass Model

The Bass Model describes the diffusion of innovation by taking of certain inputs:

1. N(t) Cumulative number of adopters at time, t
2. N-bar is the ceiling on the potential number of adopters
3. A is the coefficient of innovation
4. B is the coefficient of imitation
5. F(t) – N(t)/N-bar is the ratio of adopters who have adopted the innovation by time (t)

The equation forms that describe the diffusion curve of the above inputs are as follows:

$$\frac{dN(t)}{dt} = a[\bar{N} - N(t)] + B \times \bar{N} \times N(t) \times [\bar{N} - N(t)]_5$$

Or

$$\frac{dF(t)}{dt} = [a + bF(t)] \times [1 - F(t)]$$

Original Hypothesis

The Bass Model is typically applied to novel products at a level of the manufacturer. It has been used to predict market penetration and adoption rates as a whole. Using a similar notion we hypothesized that it could be applied to certain products that individual stores might sell. If a product was novel or differentiated from any other offering in the market it would be possible to apply a diffusion model to the information. This would allow the model to predict the total sales and rate of dispersion of a product based on limited data collected in the first weeks of availability. We looked to apply our hypothesis to an ecommerce site due to the availability of the data and the fact that all transactions will occur on the site and will be relatively straight forward to collect and analyze. Over the years, researchers have developed a number of applications of the Bass model to various hypotheses in marketing and inventory management (Teck-Hua Ho, 2002).

[5] Bass 1969

The original hypothesis was:

> *Given access to all ecommerce sales data for a given store over a discrete time period and a discrete catalog of products, one can identify and then apply a diffusion algorithm to predict the rate of sales and total cumulative sales for a new product.*

Online Retail Partner:

In order to test the diffusion model and later hypothesis against the data, we partnered with an online-retail or e-commerce company in the United States to be our research sponsor. Our original objective in finding a partner was that they had a large amount of transactions online and that they introduced several new products per sales cycle. Our partner is an online retail company with a large, established company with both an online and offline retail operation. It was not ideal for the company to have 'off-line' sales that were not captured in our data. However, we felt that given a large portion of sales are captured in the data and that it is difficult to get any sizable company to revel their sales data it was an acceptable data set to move forward with.

Why do we use this data?

The objective of this case-analysis is to work with transaction data from an e-commerce sponsor company. Sales transaction data must contain information about product sales over time and be granular enough to look at daily sales volume for any given SKU as well as the average sales price for the day. We felt that this information should be readily available from any sponsor company and we would be able to filter and prepare the data for our analysis in a timely manner. Using basic data that would be available across multiple sponsors allows us the flexibility to add more information from additional partners if additional sponsor companies could be found.

How is the data prepared?

We obtained the data through the following data sources:

1. E-Commerce software transaction history: We interacted with the e-commerce software vendor of our sponsor company. We requested an XML file consisting of all recorded transactions. Data was cleaned by the vendor of any personally identifiable information present. Each record was based on a single order and had information such as date of transaction, items ordered with quantity and price, unique customer number, use of gift card, shipping zip code, and whether the customer signed up for future contact from the company.

2. Google Analytics: We also secured access to Google Analytics to help verify the quality of the data submitted by the e-commerce transaction software. This data was available in a more generic format based on dates, volume, and price of the items. Google Analytics contained very little information on the individual customer's profile of each transaction. One advantage to using Google Analytics was the fact that the data could be exported in a more desirable format and there was no middle process manipulating the information (in order to clean the customer information).

We obtained and processed the data in order to obtain the fields we felt would be beneficial to our identification and analysis of possible diffusion patterns. This was completed in various ways, but included creating customized software, written in the JAVA language, to parse and reformat the XML data. Google Analytics data was concise enough to import the data directly into Microsoft Excel for analyzing and visualizing the information. Our goal was to make a robust process that could be repeated with more partners or future data sets.

Diffusion Model Case-Analysis:

Our first exercise with the attached data was to observe the cumulative sales trends of individual products in 2008 and 2009. We selected a number of products that exhibited sales trends over the year and attempted to chart 2008 and 2009 cumulative sales. The objective was to observe behavior of product sales from year to year.

We selected a product that had full 2008 transaction data and partial 2009 data. Our first product selected was product PB, a durable goods product. A durable good product is a product that a customer may purchase and use repeatedly across a certain period of time.

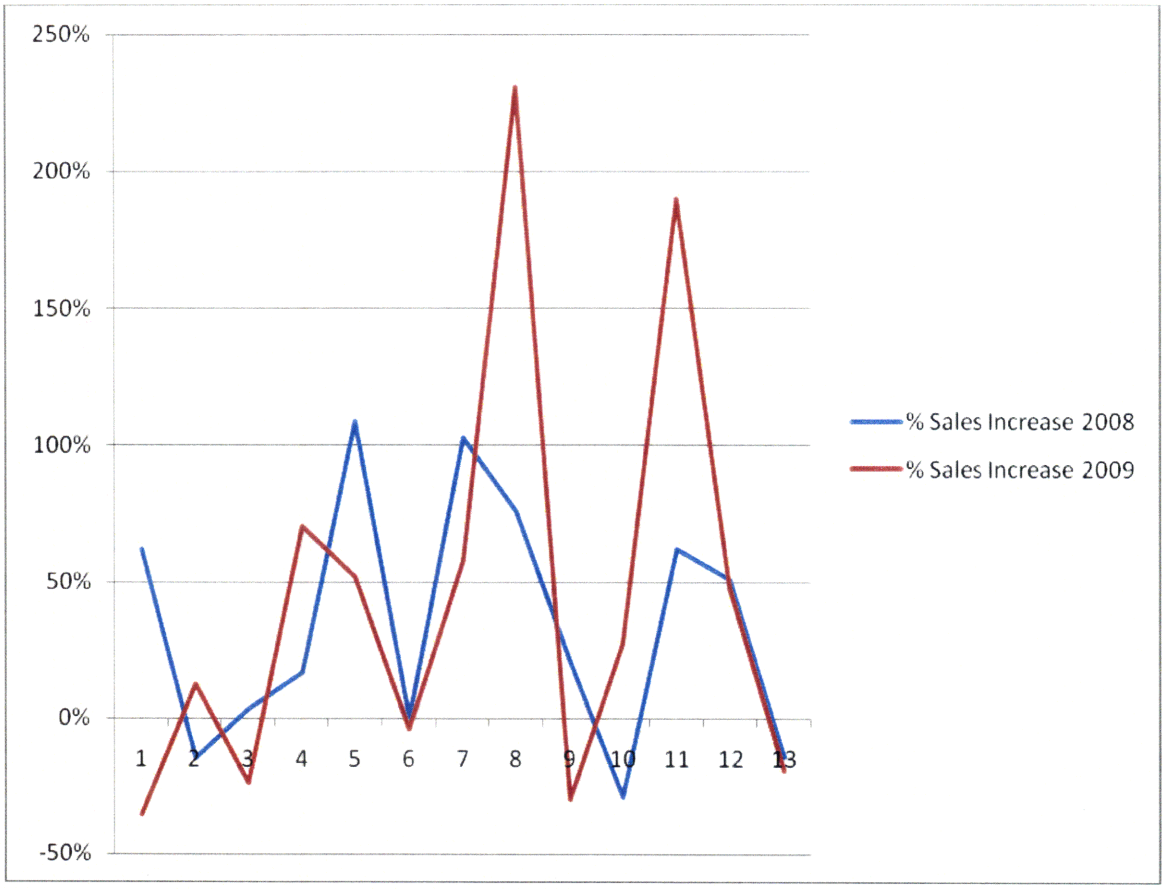

Figure 2: % Sales increases PB Product, 2008 and 2009

Figure 2 shows the % increase in sales for product, PB. This product is priced at $12.95. The chart shows the changes in sales during 2008 and 2009. The data points used included the weekly sales over the first 14 weeks of each year. Our objective was to determine whether sales in 2008 could be modeled and then effectively extrapolated to a full forecast for the 2009 year. If we could successfully fit our diffusion model to the cumulative sales changes, we would successfully model the diffusion pattern of sales over the year.

From the rates of change of sales, above, it appears that sales increases and decreases occurred during very similar times of the year in 2008 and 2009. However, the magnitude of the increase in sales was greater in 2009.

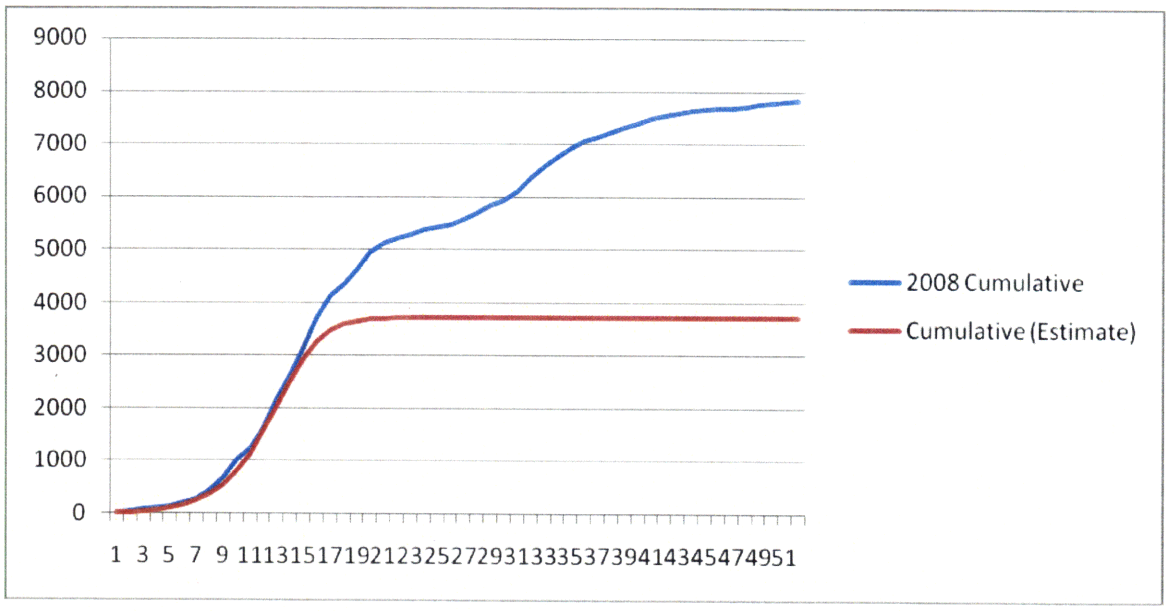

Figure 3: PB product 14-week Cumulative Sales 2008 actual and 2009 estimated

The 2008 Cumulative curve in **Figure 3** represents cumulative sales for Product PB. The Cumulative 2009 estimate was used with a modification to the traditional Bass Model. The Diffusion Model that describes the Cumulative Estimate is as follows:

$$\frac{\bar{N} - N(t)}{\bar{N}} \times B \times N(t)$$

Where:

1. $\bar{N}$ is the total estimated cumulative 2009 sales of product PB
2. $N(t)$ is the estimated 2009 sales at time t.
3. B is the coefficient of imitation

Model Issues:

The model equation was fitted against a sample of 10 products in a similar way as product PB. A number of products showed similar Bass curve type distributions of cumulative sales. However, there were some issues with diffusion forecasts past 14 weeks for each of the products tested. In order for our work to be useful we felt that we would need to predict sales at or before week 14 of the year. This was based on our particular sponsors ordering behavior for the year. Any prediction or notification after such time would not be timely and would probably be useless to the sponsor. Using data at week 12 to try to fit a diffusion curve with our equation yielded undesirable behavior. The model appeared to be fitting the curve well for

the first 16 to 18 weeks, but was only a 'good' fit beyond week 18 on one of the ten products that we identified as potential candidates. We had narrowed down the products from over 1500 to 10 in an effort to pick the most likely candidates that would work with our model and having only a 10 percent success rate was dismal. Our conclusion was that there was a strong seasonality effect on our products in the early spring and the winter months. The tapering off of sales was due to the end of the year and not necessarily from saturation of the market as would be required for the Bass Model to be effective. This observation influenced us to analyze the life-cycle of these products more closely. We determined that given the data set that we had (about 14 months of transactions) we could not see a full cycle of any product. Our partner's typical successful product could stay on their site for several years. We could not use the Bass Diffusion Model because our 'saturation' of customers was not occurring. We noticed strong seasonality patterns in all of the products that we looked at, so we decided to examine the possibility of clustering products together based on their sales trajectory over a given year. In addition we noticed that products tend to exhibit similar patterns year after year, **Figure 2.**

In addition to observations previously mentioned, we felt that cluster analysis could be performed across the catalog of products, new and old, and would be more applicable to an individual retailer versus a diffusion model. Our initial goal was to provide better sales predictions for new products. We felt that if we could cluster products and identify characteristics among clusters we would be able to provide not only estimated sales information, but provide expected demand on a week to week basis. We felt that with the appropriate information we would be able to assign a new product into a cluster and be able to estimate total sales over a year's time within the first 6-10 weeks of launch. This objective was one of our original hopes with the diffusion model. In addition after spending time interviewing with employees at the sponsor company it was clear that this type of information would be more useful to them, specifically their inventory control group.

Cluster Analysis:

The objective here was to understand how pervasive are distribution similarity across different groups of products.

Our team collected all weekly sales transaction records, for each product SKU, from Google Analytics and our online-retailer's e-commerce software provider. Then, weekly sales were normalized across the total of 52 weeks so that the cumulative sales patterns of various products could be examined as groups. In addition this would allow us to group a product that might sell only a few hundred items and one that sells thousands of units together. Our goal with normalization was that we didn't want the volume of

sales to be a critical factor in clustering. Sorting by volume is something that is not only easily done, but is usually an obvious piece of information to an ecommerce retailer.

Data was then prepared into a training and test partitions using a 60% of the available observation for the *training* sample and a 40% of the observations for our *test* sample. The partition size of the training partition consisted of 973 products. The partition size for the test partition consisted of 650 products. Once successfully, tested, we would use the 2009 catalog sales data to forecast and validate the particular clusters. 2009 catalog data will be blind data against which we will validate each cluster.

What is clustering?

We used a clustering methodology as a grouping method to quickly identify clusters of similar groups of products according to some particular measurement. According to Shmueli and Patel, clustering is an appropriate tool for "market structure analysis: identifying groups of similar products according to competitive measures of similarity (Galit Shmueli, 2007)."

We used a non-hierarchical approach to clustering knows as K-Means. According to Shmueli and Patel, k-means approach to clustering is "divides a sample into a predetermined number *k* of nonoverlapping clusters so that clusters are as homogenous as possible with respect to the measurements used."[6] The process of computing the clusters is an iterative one that starts form the *k* initial clusters. Clustering of large data-sets of 100 records, or more, is also reported to be most effective using a k-means clustering approach(Galit Shmueli, 2007). According to the SAS JMP software module the k-means process can be described as follows:

> "The k-means approach to clustering performs an iterative alternating fitting process to form the number of specified clusters. The k-means method first selects a set n-points called cluster seeds as a first guess of the means of the clusters. Each observation is assigned to the nearest seed to form a set of temporary clusters. The seeds are then replaced by the cluster means, the points are reassigned, and the process continues until no further changes occur in the clusters. When the clustering process is finished, you see tables showing brief summaries of the clusters. The k-means approach is a special case of a general approach called the EM algorithm, where E stands for Expectation (the cluster means in this case) and the M stands for maximization, which means assigning points to closest clusters in this case."[7]

[6] *Ibid* 12

[7] JMP Help Files, SAS Institute 2008

Pre-determining the number of *k* clusters into which we partition the sample was done manually. We began with a large number for k, approximately 15. We ran each clustering process on JMP and observed the results. We iteratively reduced *k* until there were no ending clusters with less than 30 members. The reduction of *k* through this process resulted in a k of size 9.

Iterative Clustering using K-Means: 2008 Old Products

Cluster Summary

750 iterations .

Cluster	Count	Max Distance
1	109	1.27820843
2	228	1.12688535
3	40	0.74904074
4	192	1.30152767
5	64	1.38080534
6	98	1.18299809
7	45	0.66571951
8	146	1.46231675
9	51	1.3066909

Table 1: Cluster Summary 2008 Full-Year Products

Table 1 describes the initial output of a K-Means cluster analysis of the normalized cumulative sales transactions from January 1, 2008 through December 31, 2008. Each cluster is labeled by a number with a count of the number of products in that cluster. The *Max Distance* measures the maximum distance from the center of the cluster to the furthest row in that cluster.

Products that exhibited no sales for the entire year were excluded from the analysis. Products that exhibited no sales during the first 10 weeks were excluded from this current sample set and placed into a separate sample set labeled "New Products." Due to the time lag in the initial sales, these products would skew other clusters of full-year products or be placed into a separate cluster group. We felt that these products could yield useful information on the launch of a new product in the future. In addition, the sales pattern for the following year would not exhibit the similarities as was evident in **Figure 2** for an old product. Products with no sales from week 10 would be new products introduced during the year. Our sponsorship partner informed us that products are added to the catalog during the year and we assumed those would be captured as "New Products" exhibiting zero sales for extended periods of time. A separate cluster analysis was executed on these products and will be described forthwith.

2008 Full-year Sales: Parallel Coordinate Plots of Clusters

Cluster 1: 149 Members

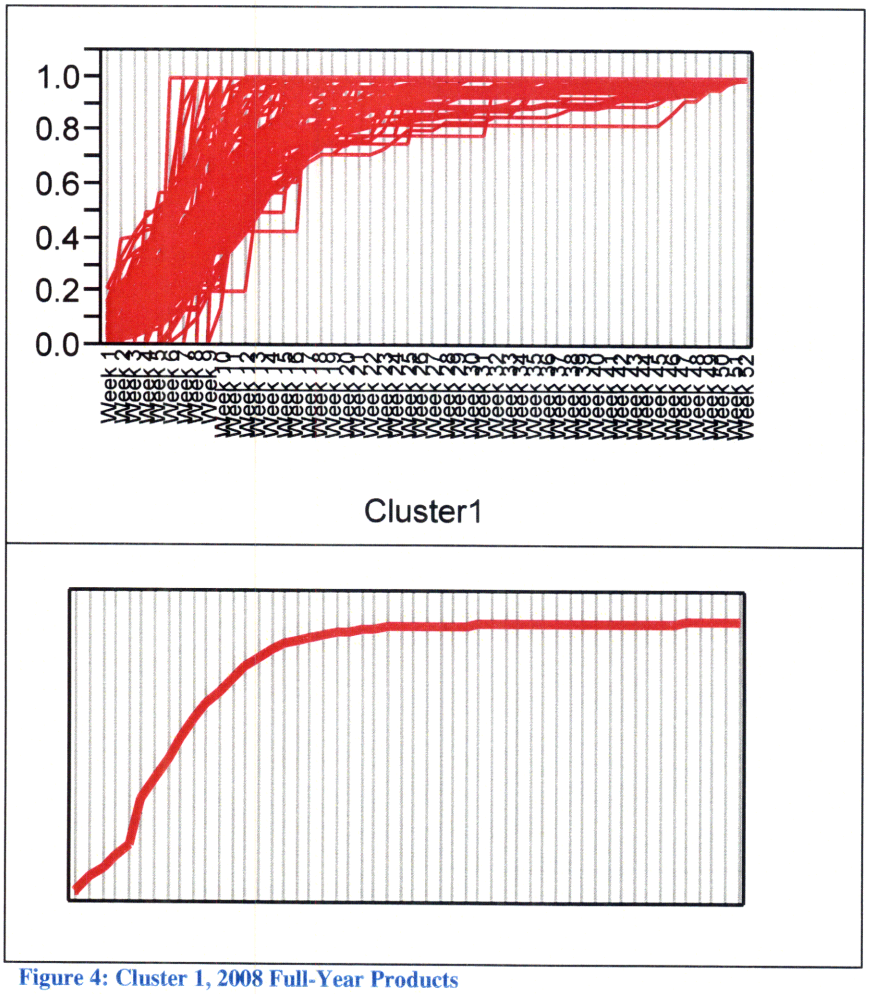

Figure 4: Cluster 1, 2008 Full-Year Products

Cluster 1, represented in **Figure 4**, is the 4th largest cluster with 149 members. The products contained in this cluster have a seasonal sales pattern that launches early in the year, during week 1 and then decrease from week 13 through week 20. These products may possibly consist of "winter season" products that are launched and marketed at the beginning of the year.

Cluster 1	Minimum Distance	Maximum Distance	Range
Mean	0.0402573	0.999341	0.95908361
Standard Deviation	0.00228725	0.27432622	

Table 2: Cluster 1 Statistics – 2008 Full Year

Cluster 2: 228 Members

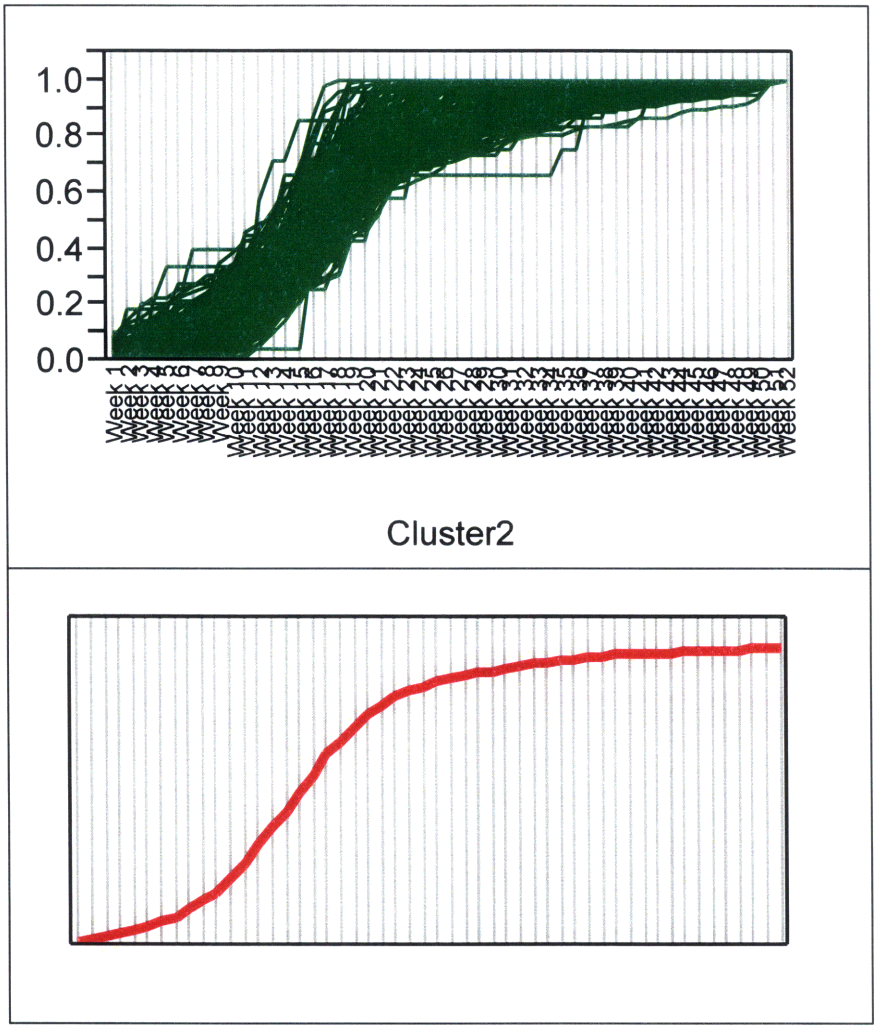

Figure 5: Cluster 2, 2008 Full-Year Products

Figure 5 represents the weekly change of cumulative annual sales of Cluster 2 members and the subsequent cluster mean. Cluster 2 is the largest cluster of the set with 228 products as members.

The distribution of sales appears to follow an S-curve pattern of a Bass model of diffusion of innovation. Sales, for products in this cluster, start slowly, but increase at a higher rate. The inflection points appear to be for sales increases, weeks 4 through week 10. Weeks 20 through weeks 30 appear to be that time of year that sales decrease with little sales occurring during the end of year holidays. There is no strong holiday sales effect with this current cluster.

Cluster 2	Minimum Distance	Maximum Distance	Range
Mean	0.00905558	0.99909879	0.99004321
Standard Deviation	0.00208439	0.12109352	

Table 3: Cluster 3 Statistics – 2008 Full Year

Cluster 3: 40 Members

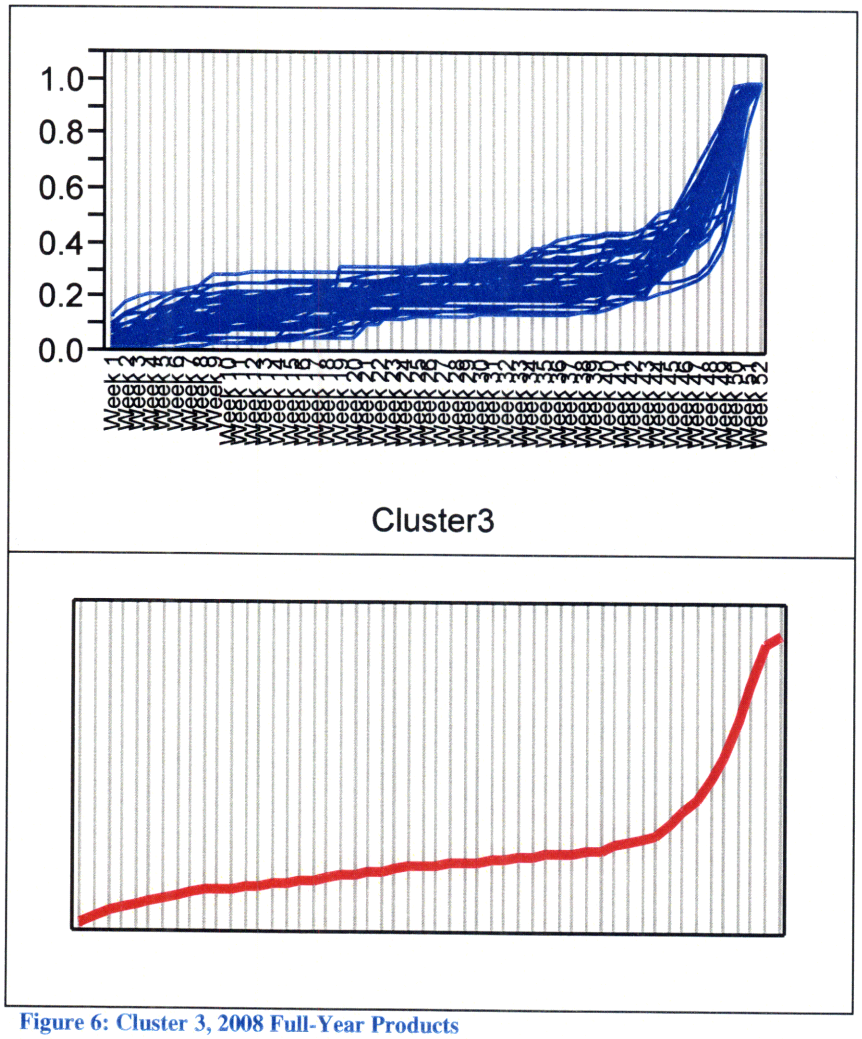

Figure 6: Cluster 3, 2008 Full-Year Products

Figure 6 consists of the smallest cluster group, Cluster 3, with only 40 members.

Cluster 3 products appear to exhibit a very gentle slope of increasing sales during the first 40 weeks of the year. The cumulative sales then show a sharp increase during the last 10 to 15 weeks of the year. Thus, it appears that Cluster 3 products are most popular during the holiday season and may be holiday gift purchases that are in the catalog for the full year.

Inspection of products in this cluster shows that the products are small, relative to other products in the catalog. One particular product appears to be a mini "gift" version of a larger regular product, from another cluster.

Cluster 3	Minimum Distance	Maximum Distance	Range
Mean	0.03503712	0.97201371	0.93697659
Standard Deviation	0.03115272	0.11422886	

Table 4: Cluster 3 Statistics – 2008 Full Year

Cluster 4: 192 Members

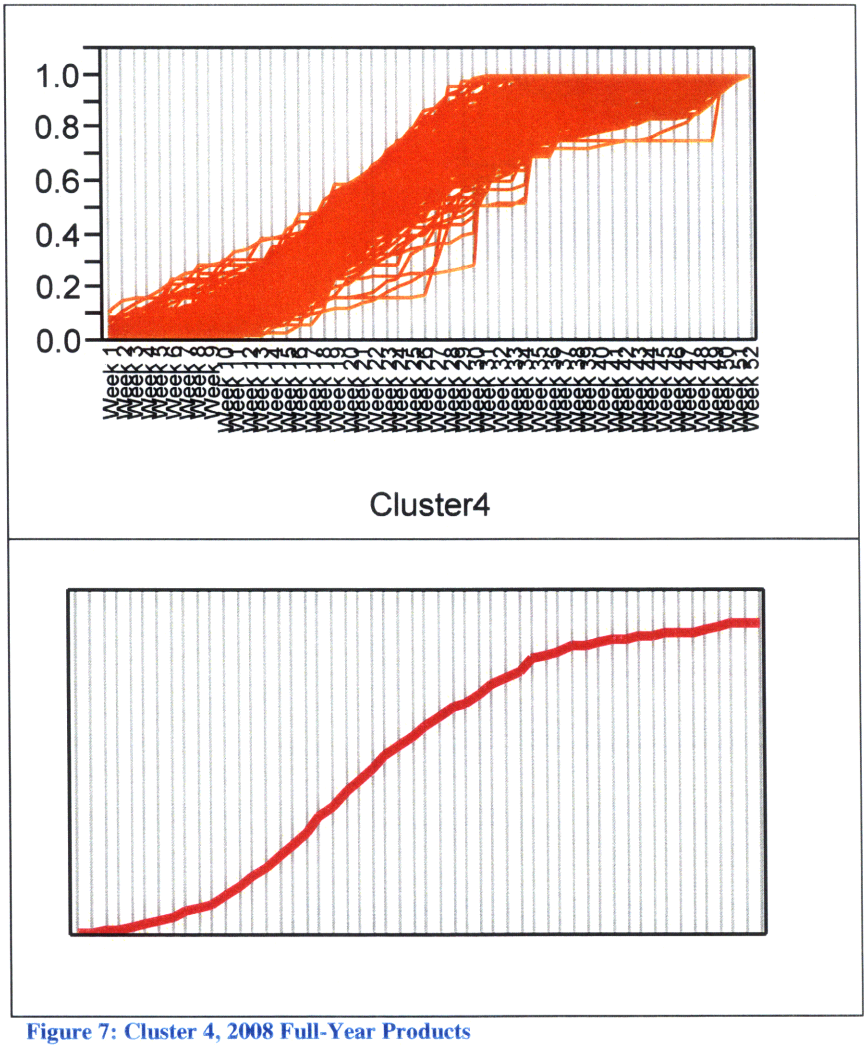

Figure 7: Cluster 4, 2008 Full-Year Products

Figure 7 consists of the coordinate plot of the 2nd largest cluster of products in the 2008 product catalog. These products exhibit a similar path of seasonality as that of Cluster 2. However, the increase in sales is

a less steep curve that increases from week 10 through week 35. Sales in Cluster 4 also continue to increase during the holiday season more steeply than in Cluster 2. The concentration of products around the mean appears to be very high.

Cluster 4	Minimum Distance	Maximum Distance	Range
Mean	0.00789099	0.99843246	0.99054147
Standard Deviation	0.00284518	0.11235816	

Table 5: Cluster 4 Statistics – 2008 Full Year

Cluster 5: 64 Members

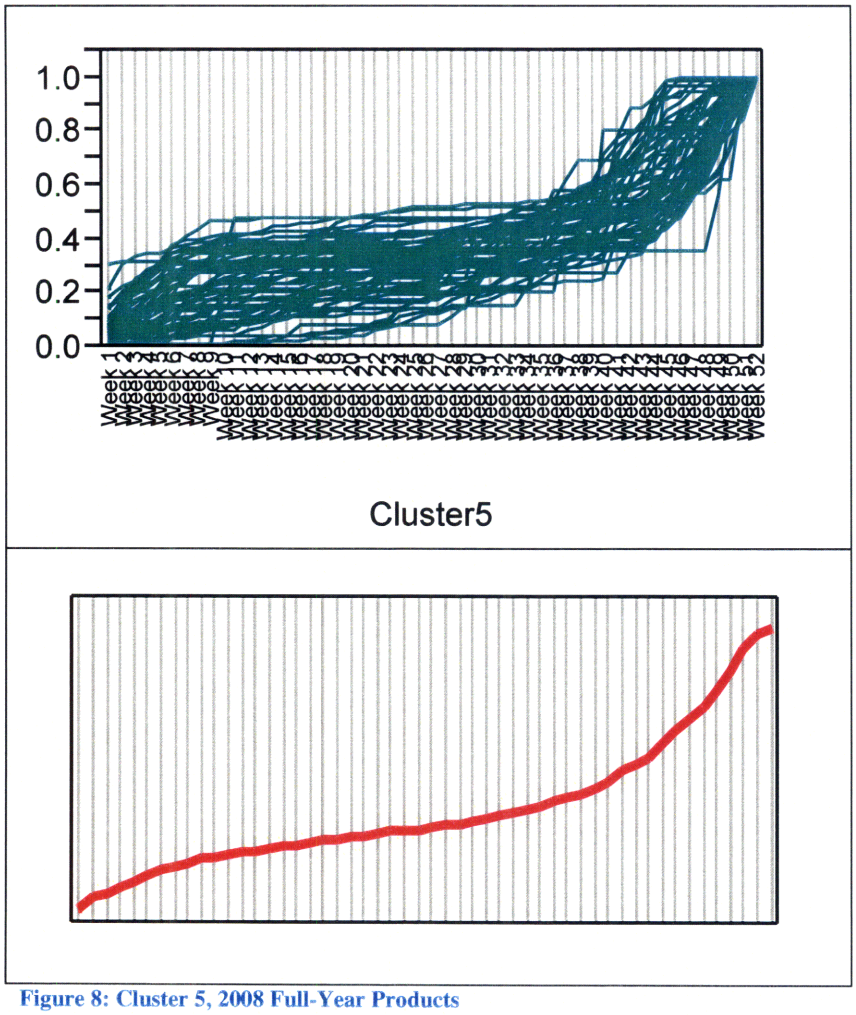

Figure 8: Cluster 5, 2008 Full-Year Products

Figure 8 shows the Cluster 5 products that exhibit a similar sales pattern to those products listed in Cluster 3. However, these products annual sales distribution appear to be higher than Cluster 3. Also, similar to Cluster 3, the products in Cluster 5 have a high sensitivity to increasing end of year holiday sales. The increasing slope at the end of the year for Cluster 3 begins at week 28, around the middle of the year. This could mean that these products have mid-year seasonal market trends that reflect increasing customer demand at that time.

Inspection of the Cluster 5 table also shows a wide band of products, compared to the previous clusters which appear to be organized around together bands. This band width indicates a wider degree of variability in product sales rates over the year.

Cluster 5	Minimum Distance	Maximum Distance	Range
Mean	0.05377043	0.97982749	0.92605706
Standard Deviation	0.03071681	0.15690702	

Table 6: Cluster 5 Statistics - 2008 Full Year

Cluster 6: 98 Members

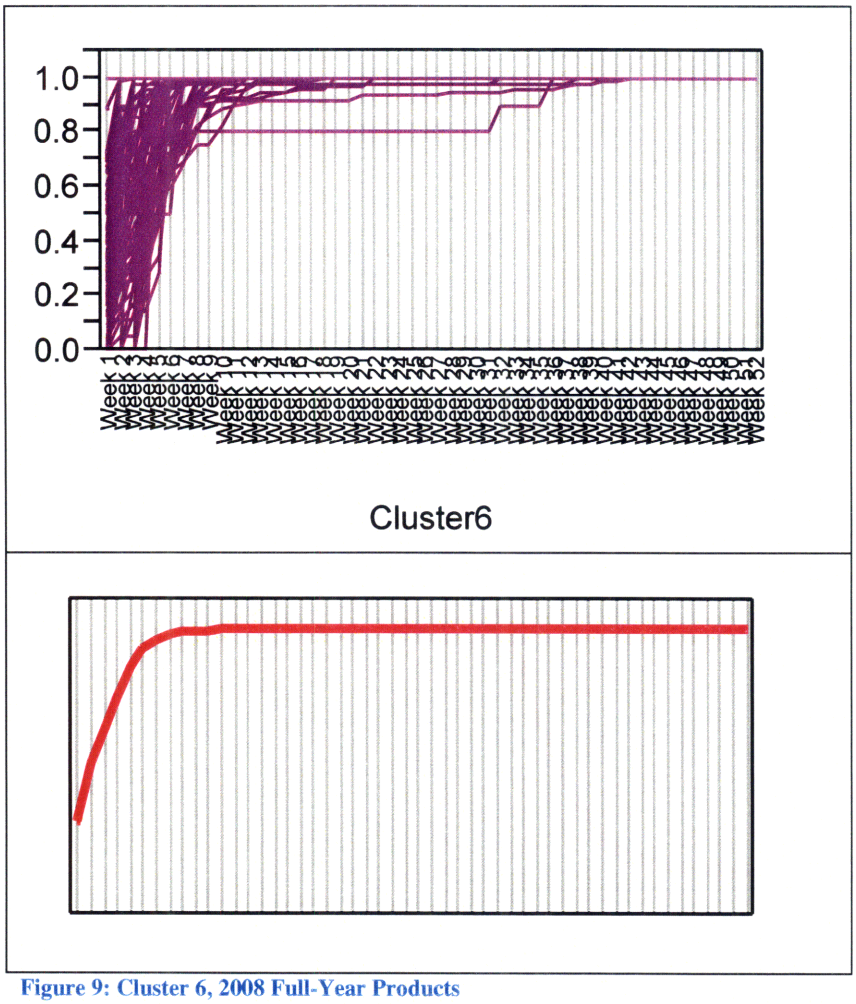

Figure 9: Cluster 6, 2008 Full-Year Products

Figure 9 Products are grouped into Cluster 6. These products experience very high sales growth in the first 8 weeks of 2008. Then, the products quickly taper off. This may be that these products sold out of inventory early in the year and were not re-ordered by the online retailer. We believe that if the retailer re-ordered, then sales would pick up once the products were replenished.

Our online-retailer sponsor company informed us that some products had very long lead times. As such, those products were usually difficult to re-order once launched.

Cluster 6	Minimum Distance	Maximum Distance	Range
Mean	0.32482524	1	0.67517476
Standard Deviation	0	0.34982632	

Table 7: Cluster 6 Statistics – 2008 Full Year

Cluster 7: 45 Members

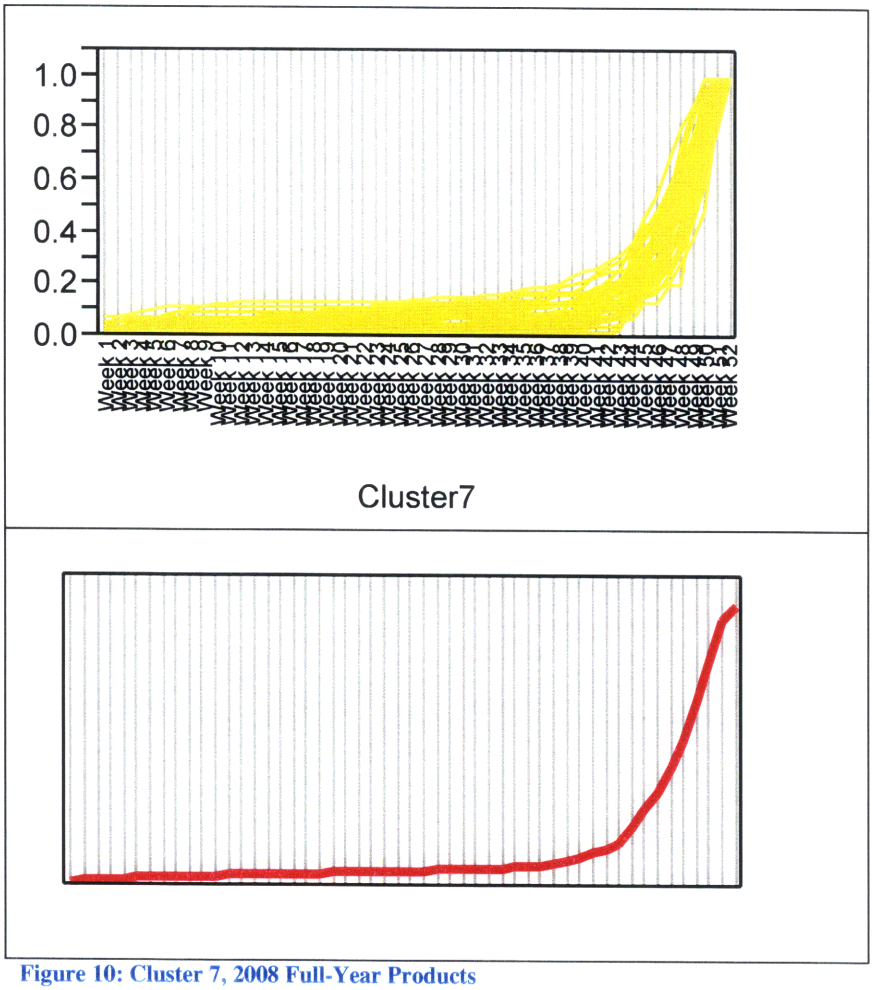

Figure 10: Cluster 7, 2008 Full-Year Products

Cluster 7 Products, in **Figure 10**, make up the 2nd smallest cluster with 45 members from the test partition of products. These products appear to have the most sales growth during the last 10 weeks of the year, from week 41 through week 52. Product sales are very low during the first 40 weeks of the year. Inspection of the products in this segment indicates that the products are holiday products which strong holiday characteristics in their color, size and function. Additionally, some of these products are not related to the online-retailers traditional business.

Cluster 7	Minimum Distance	Maximum Distance	Range
Mean	0.01171485	0.94454987	0.93283502
Standard Deviation	0.01773075	0.13158724	

Table 8: Cluster 7 Statistics – 2008 Full Year

Cluster 8: 146 Members

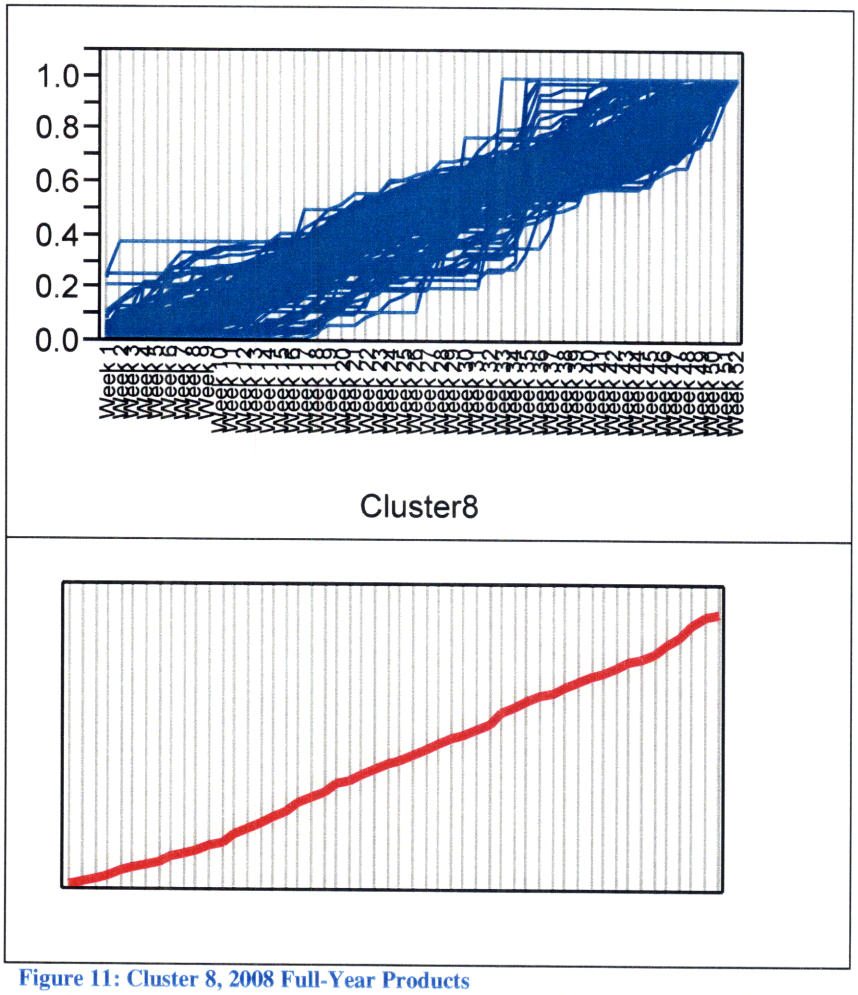

Figure 11: Cluster 8, 2008 Full-Year Products

Cluster 8, described in **Figure 11**, is a very unique cluster. The cluster mean appears as a linear growth line from week 1 with steadily increasing growth through week 52. Similar to Cluster 5, there is a wide band of product cumulative sales around the cluster mean. The maximum distance to the cluster center is the highest of all clusters, measuring at 1.4.

It appears that products in Cluster 8 are purchased throughout the year at a steady rate. This could mean that these products are necessary products for the customers of our online-retail sponsor and do not respond well to advertising or marketing efforts during the year

Cluster 8	Minimum Distance	Maximum Distance	Range
Mean	0.01963221	0.9900194	0.97038719
Standard Deviation	0.01575374	0.11085825	

Table 9: Cluster 8 Statistics – 2008 Full Year

Cluster 9: 51 Members

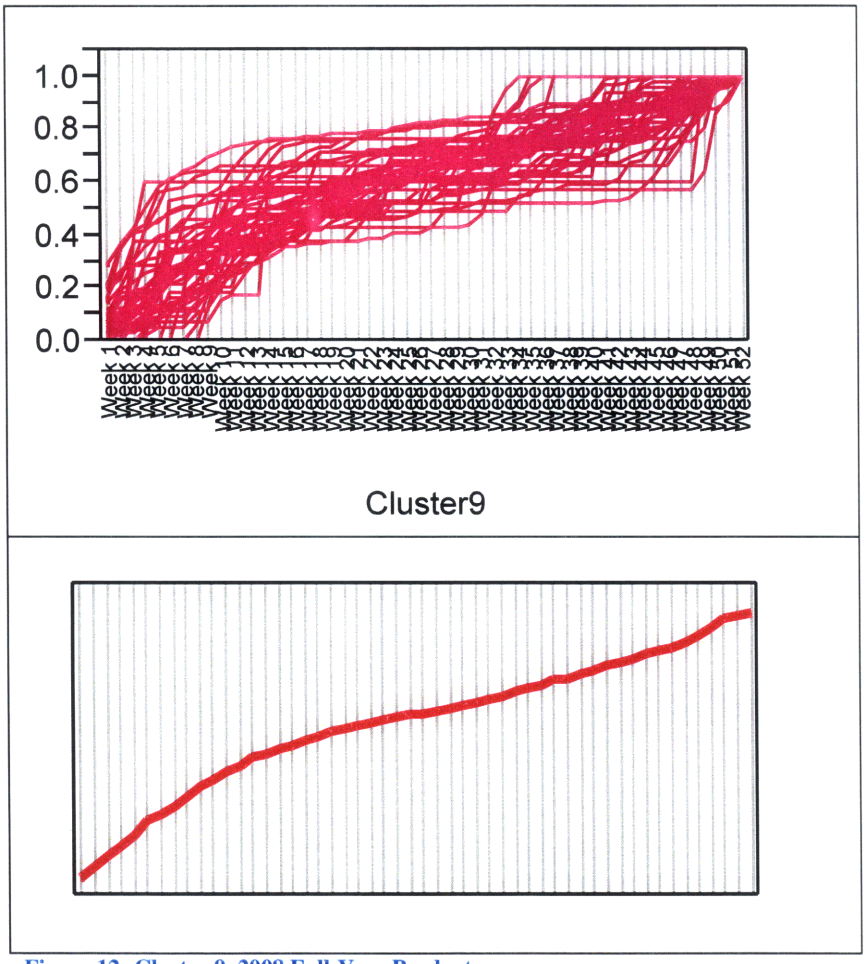

Figure 12: Cluster 9, 2008 Full-Year Products

Figure 12 consists of products members of Cluster 9 that appear to have a similar sales pattern to the products in Cluster 8. Both clusters have some of the highest maximum distances from the center of the cluster. Additionally, Cluster 8 and Cluster 9 have the weakest seasonality affects of all the products identified in the 2008 sample.

However, Cluster 9 does appear to have the highest week 1 through week 40 annual cumulative sales, before the holiday season pick-up. Thus, these products appear to outsell Cluster 3, Cluster 5 and Cluster 7 during the first 40 weeks of the year and under-perform during the holiday season's expected uptick for holiday product shopping.

Cluster 9	Minimum Distance	Maximum Distance	Range
Mean	0.05987761	0.99256428	0.93268667
Standard Deviation	0.01742875	0.16037638	

Table 10: Cluster 9 Statistics – 2008 Full Year

Cluster Means: 2008 Full-year products

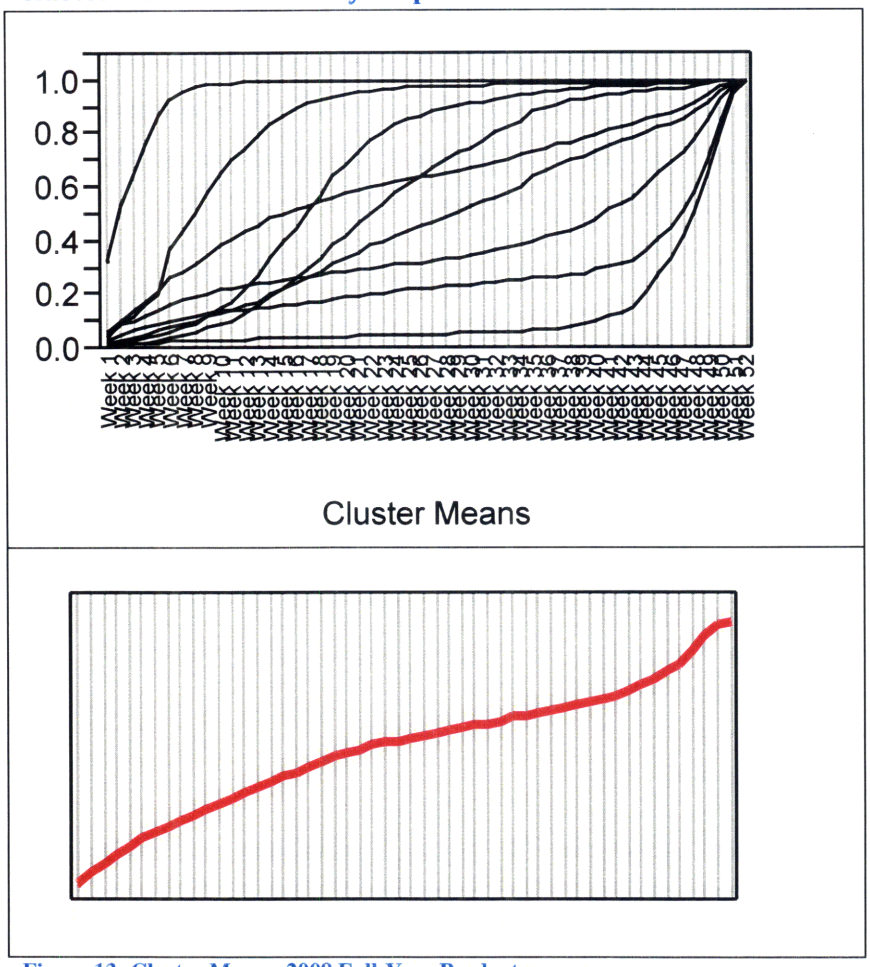

Figure 13: Cluster Means, 2008 Full-Year Products

Figure 13 represents all the cluster means and the mean cumulative sales curve of all the 9 clusters. Interestingly, if the number of clusters is optimal, then the mean cumulative sales of all full-year products appear to follow a mostly steady increase from week 1, with a slight uptick during the holiday season.

Iterative Clustering using K-Means: 2008 New Products

In order to separate new products from full-year products, we created a separate sample of all products that did not have any sales before week 10. This sample consisted of a total of 886 product IDs. We created a *Test* partition and *Validation* partition using a 60% split for the test partition and 40% for the validation partition. Below is the Cluster analysis for these products.

Cluster Summary

500 iterations

Cluster	Count	Max Distance
1	204	1.35286399
2	59	1.1853648
3	87	1.58974037
4	37	1.0214895
5	3	0.22222222
6	3	0.34836141
7	33	0.86640183
8	40	1.57331841
9	65	1.50851761

Table 11: Cluster Summary 2008 New Products

Table 11 represents the cluster summary output from JMP k-means processing of "new products" launched in 2008. The cluster output for these products differs from the original cluster output in the following ways:

1. There are 2 very small clusters with 3 products. These clusters may not provide any
2. There are 2 clusters with maximum distances of 1.5, greater than the maximum distance of 1.4 for the all-year products in the previous cluster analysis.

Cluster 1 (2008 New Products): 204 Members

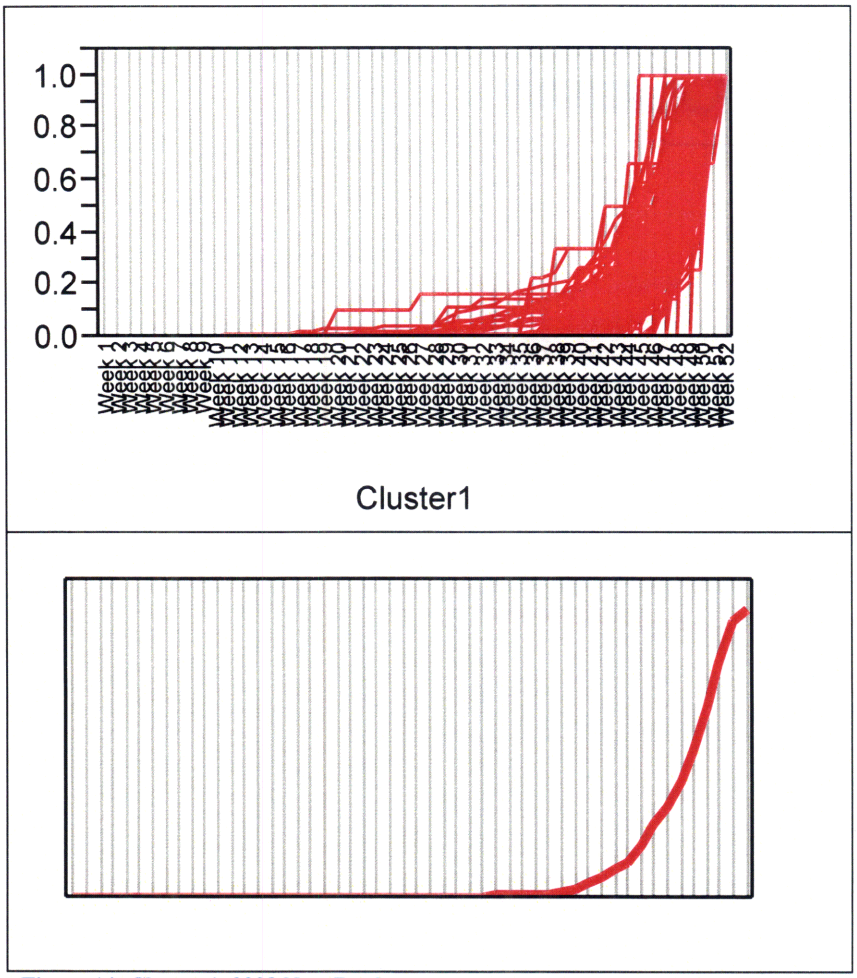

Figure 14: Cluster 1, 2008 New-Products

Figure 14 shows the cluster behavior and cluster-mean for Cluster 1 of the 2008 New Products. This Cluster performs similarly to our "holiday" Cluster, Cluster 7 in the 2008 Full Year products *k*-means analysis. Products generate very few sales until 37th week or during the last 3 – 4 months of the year. These products may be items that our online partner purchased for holiday sales, midyear, and launched online by week 40.

Cluster 1 (New Products)	Minimum Distance	Maximum Distance	Range
Mean	0	0.96042351	0.96042351
Standard Deviation	0	0.19253205	

Table 12: Cluster 1 Statistics – 2008 New Products

Cluster 2 (2008 New Products): 59 Members

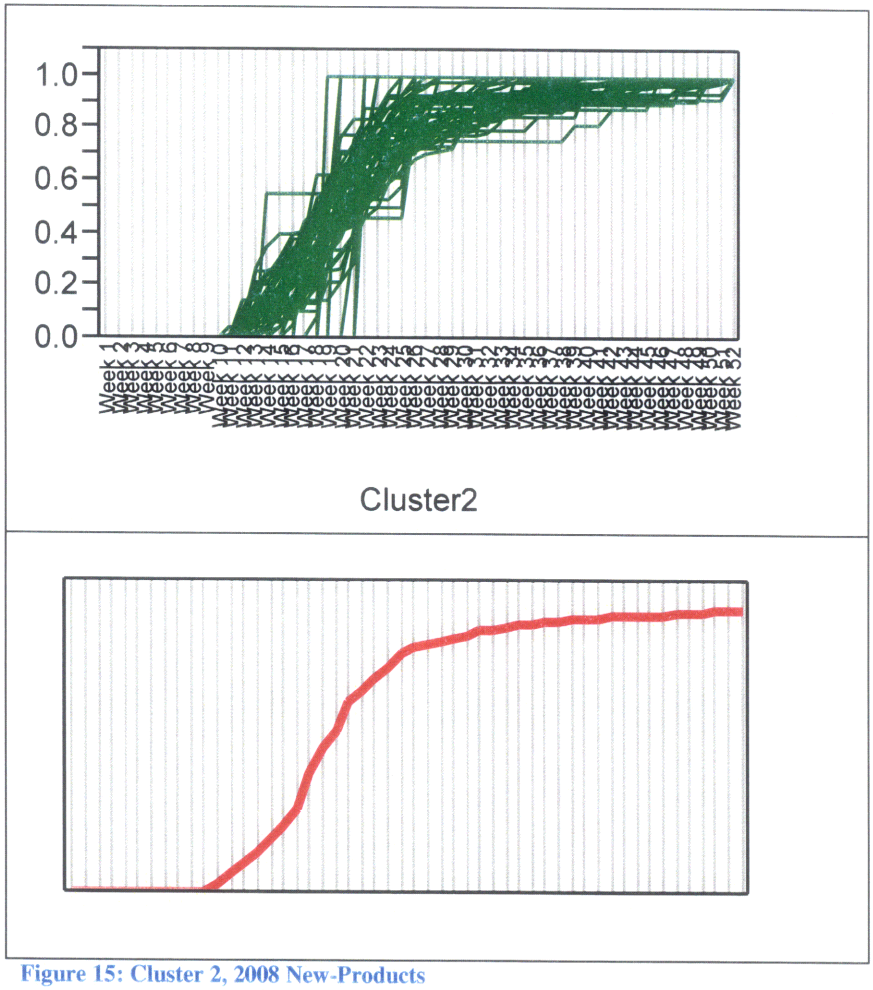

Figure 15: Cluster 2, 2008 New-Products

In Cluster 2, **Figure 15**, products are launched during week 10 of 2008. The cumulative sales of these products appear to move in a mid-year diffusion curve that tapers off between the 25th through the 30th weeks. These products appear to be specific, seasonal products that are promoted and sold during the spring season. However, these products could also be re-stocks of prior products that were stocked out during the first 10 weeks of the year.

Cluster 2 (New Products)	Minimum Distance	Maximum Distance	Range
Mean	0	0.99610672	0.99610672
Standard Deviation	0	0.20199672	

Table 13: Cluster 2 Statistics – 2008 New Products

Cluster 3 (2008 New Products): 87 Members

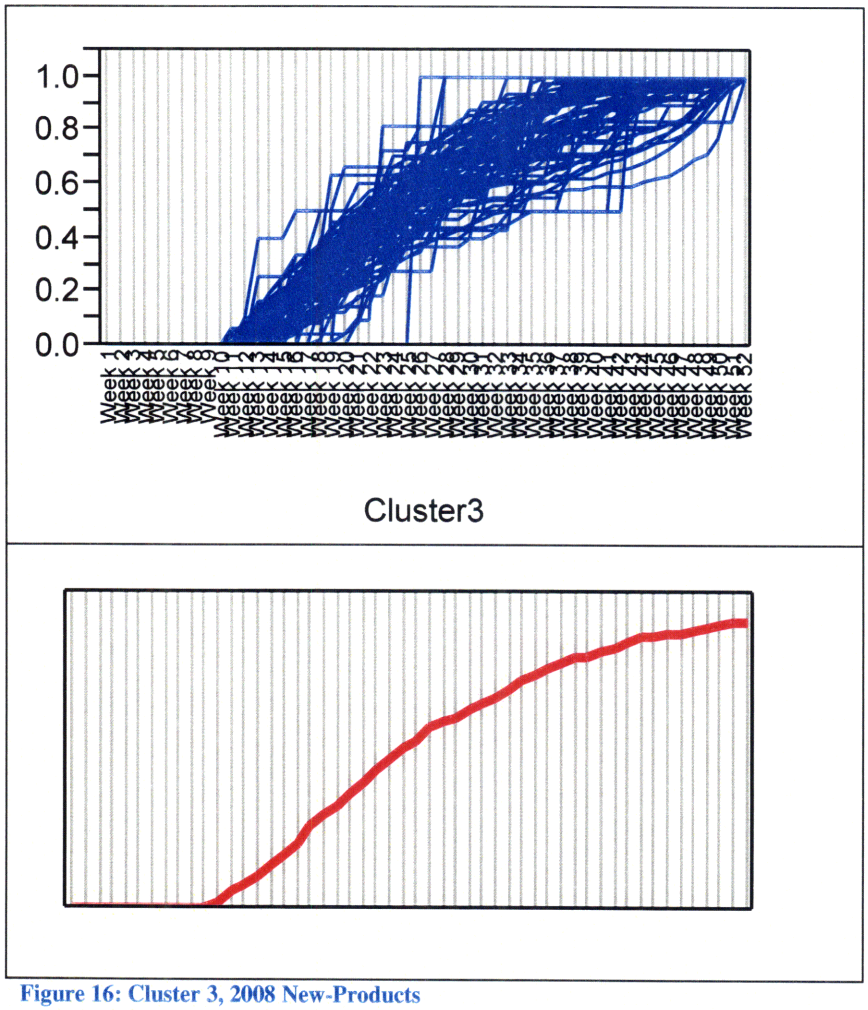

Figure 16: Cluster 3, 2008 New-Products

Figure 16, showing members of Cluster 3 of 2008 New Products is another cluster with large distances from the cluster center. The shape of the cluster-mean curve appears to curve very slightly with sales growth almost linear in shape. This product does not appear to have strong holiday sales sensitivity. Cumulative sales appear to decrease from week 40 through week 52.

Cluster 3 (New Products)	Minimum Distance	Maximum Distance	Range
Mean	0	0.99513836	0.99513836
Standard Deviation	0	0.13655775	

Table 14: Cluster 3 Statistics – 2008 New Products

Cluster 4 (2008 New Products): 37 Members

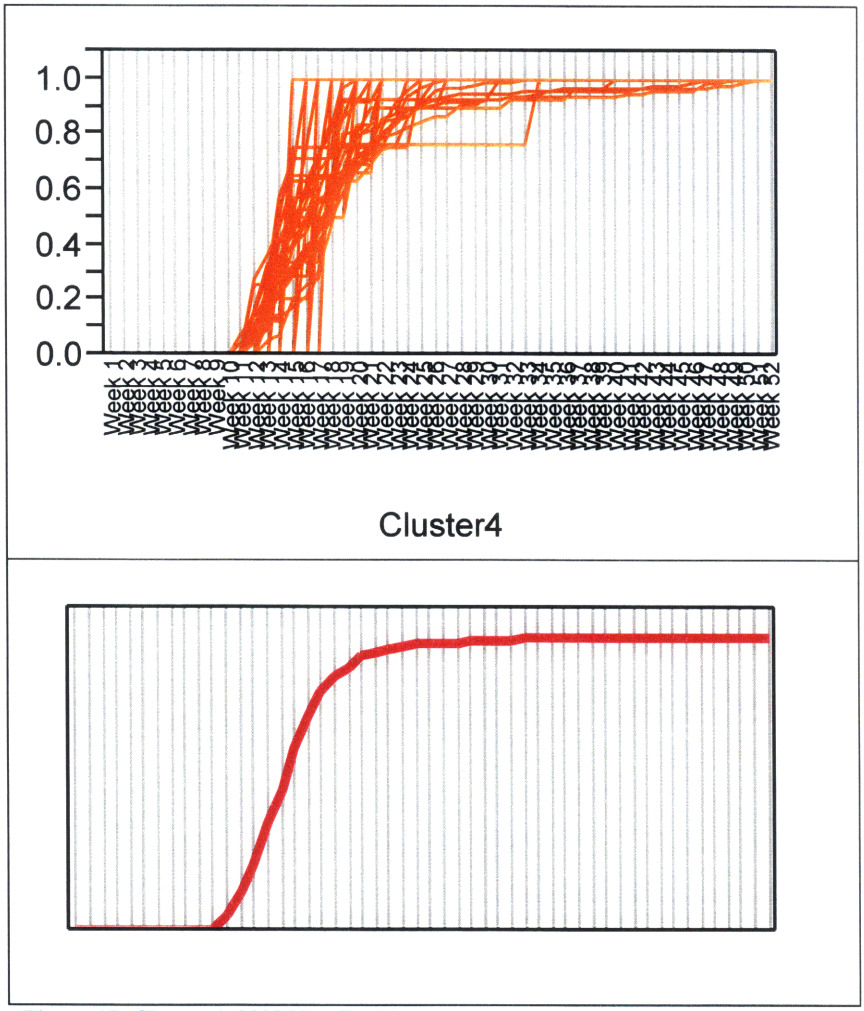

Figure 17: Cluster 4, 2008 New-Products

Cluster 4 for 2008 New Products appears to be similar as Cluster 2, but a smaller group with only 37 members. This product is launched mid-year, with sharp sales from week 10 quickly ending by week 20th. Again, it is possible that these products were highly popular and were out of stock within 10 weeks of their launch. Our online retail partner informed us that due to lengthy lead times, some popular products would run out of stock during the middle of the year.

Cluster 4 (New Products)	Minimum Distance	Maximum Distance	Range
Mean	0	0.99995676	0.99995676
Standard Deviation	0	0.28497944	

Table 15: Cluster 4 Statistics – 2008 New Products

Cluster 5 (2008 New Products): 3 Members

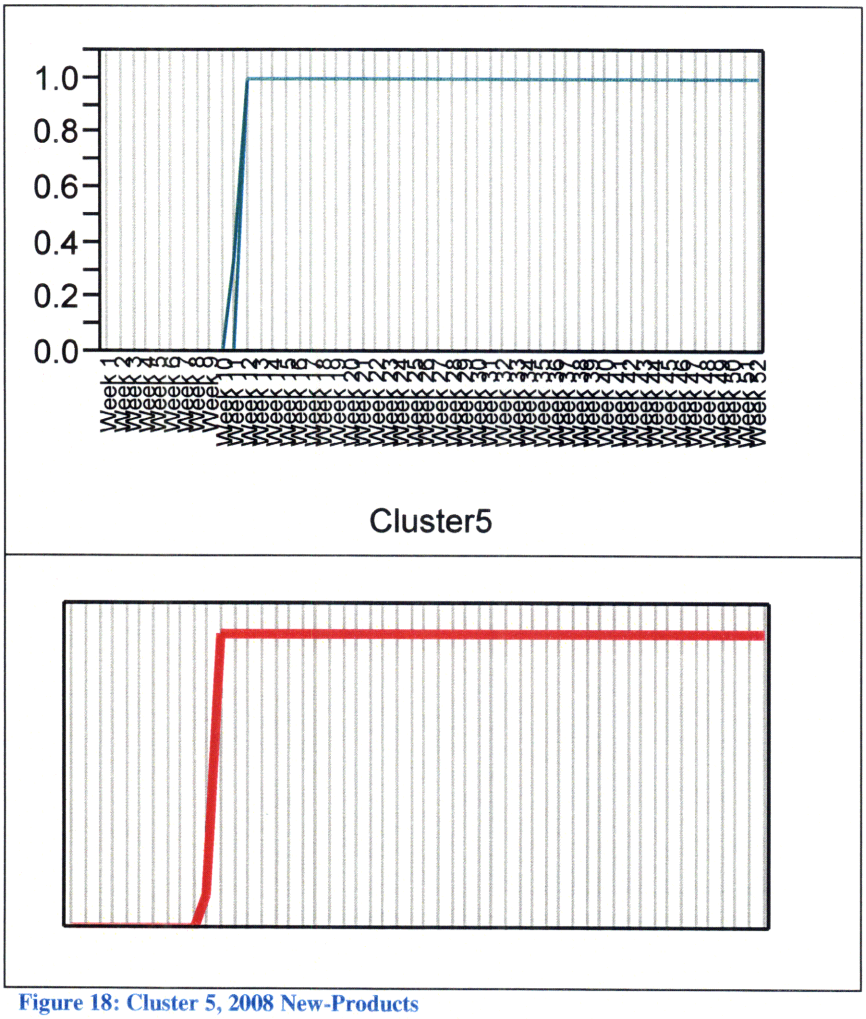

Figure 18: Cluster 5, 2008 New-Products

Cluster 5, in **Figure 18**, is a small cluster of only 3 products that were launched at the same time and quickly sold out within 3 – 4 weeks. This cluster's small number of members may indicate that no conclusive points can be further ascertained from this group.

Cluster 5 (New Products)	Minimum Distance	Maximum Distance	Range
Mean	0	1	1
Standard Deviation	0	0.19245009	

Table 16: Cluster 5 Statistics – 2008 New Products

Cluster 6 (2008 New Products): 3 Members

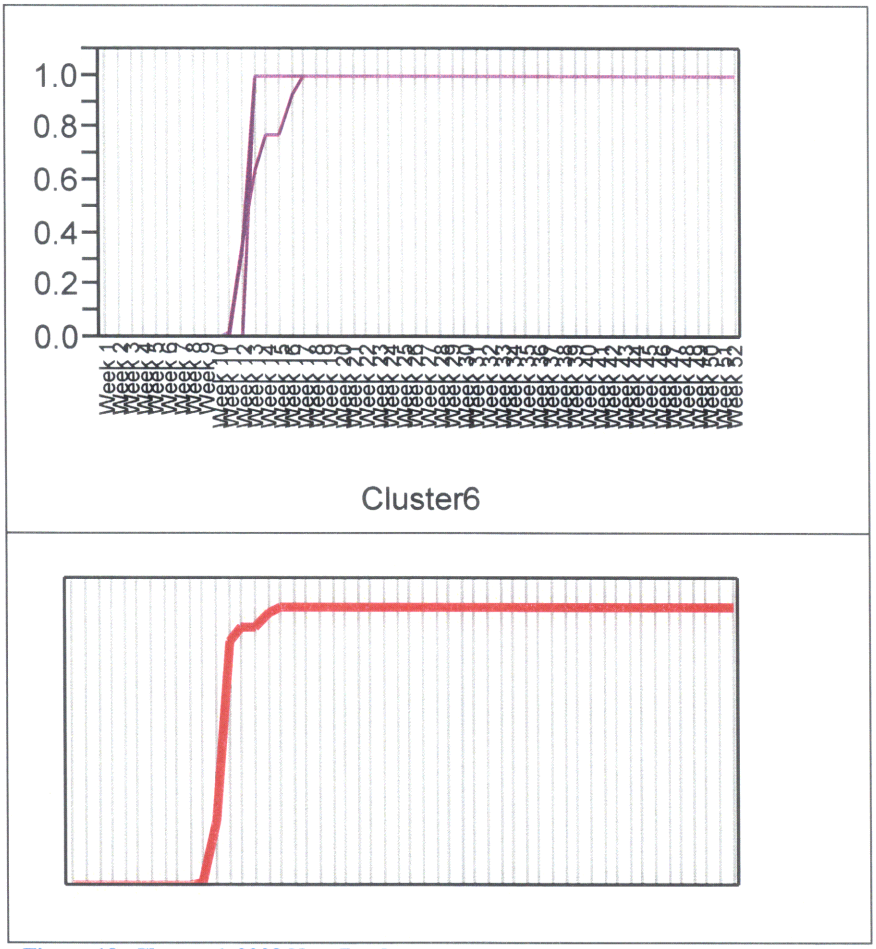

Figure 19: Cluster 6, 2008 New-Products

Cluster 6 is very similar to Cluster 5. It is also is a small cluster of only 3 products that were launched at the same time and quickly sold out within 2 – 3 weeks.

Cluster 6 (New Products)	Minimum Distance	Maximum Distance	Range
Mean	0	1	1
Standard Deviation	0	0.209119	

Table 17: Cluster 6 Statistics – 2008 New Products

Cluster 7 (2008 New Products): 33 Members

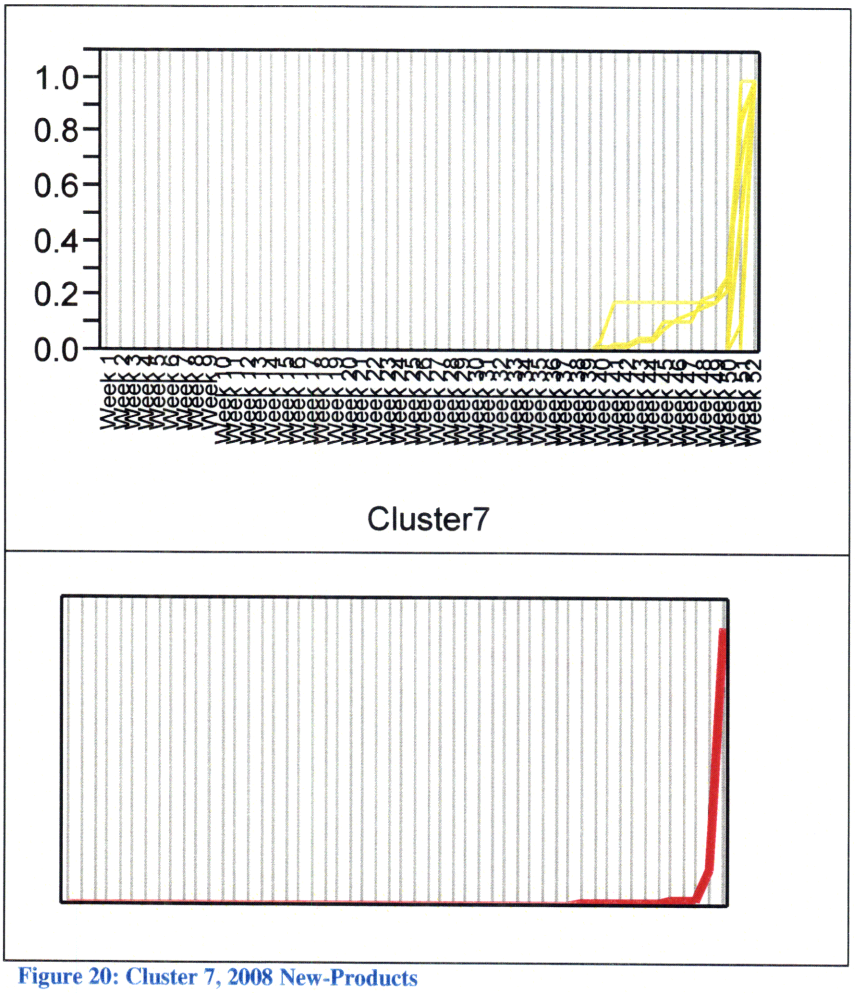

Figure 20: Cluster 7, 2008 New-Products

Cluster 7, **Figure 20**, consists of holiday products launched during the holiday season. These 33 members experienced rapid sales growth during product launches that started around the last 10 weeks of the year.

Cluster 7 (New Products)	Minimum Distance	Maximum Distance	Range
Mean	0	0.13464322	0.13464322
Standard Deviation	0	0.29487465	

Table 18: Cluster 7 Statistics – 2008 New Products

Cluster 8 (2008 New Products): 40 Members

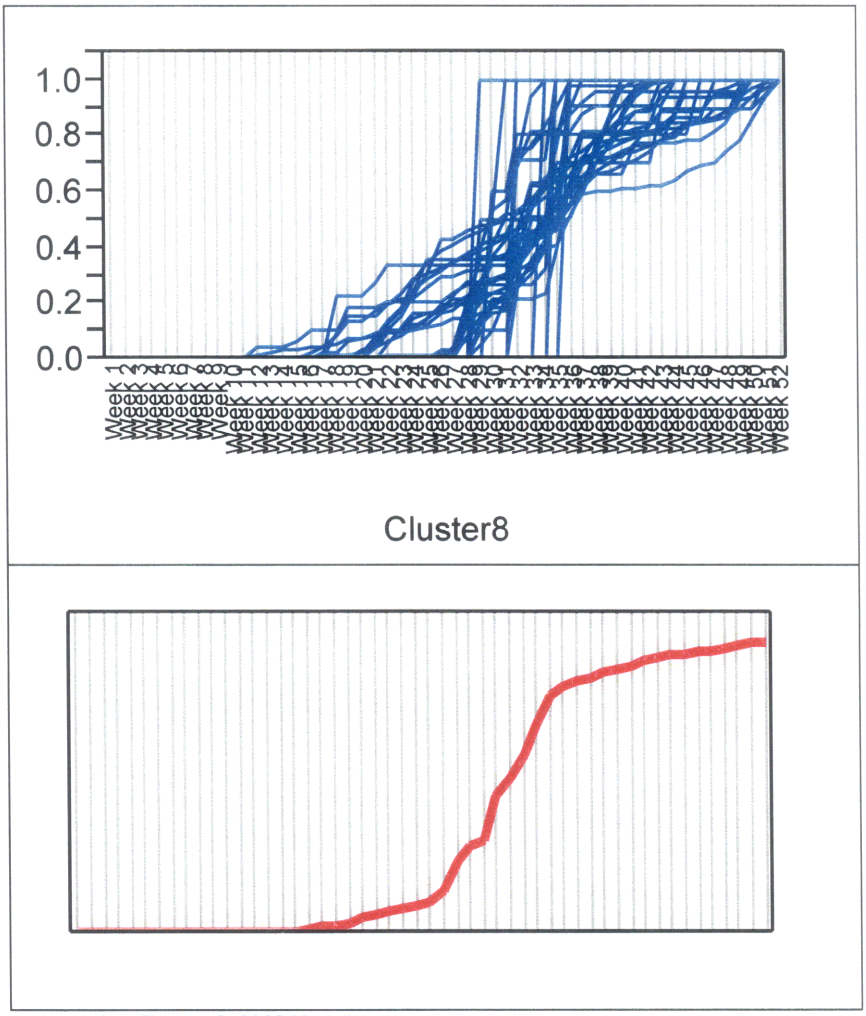

Figure 21: Cluster 8, 2008 New-Products

Cluster 8 in **Figure 21** of the 2008 New Products appears to consist of a number of products launched from week 10 through week 35. The cluster mean curve of this group shows quick sales increases through week 37 and then rapidly decreasing sales through the holiday season.

It is possible that these products were seasonal for the summer, ending at fall and then winter.

Cluster 8 (New Products)	Minimum Distance	Maximum Distance	Range
Mean	0	0.9957251	0.9957251
Standard Deviation	0	0.28290941	

Table 19: Cluster 8 Statistics – 2008 New Products

Cluster 9 (2008 New Products): 65 Members

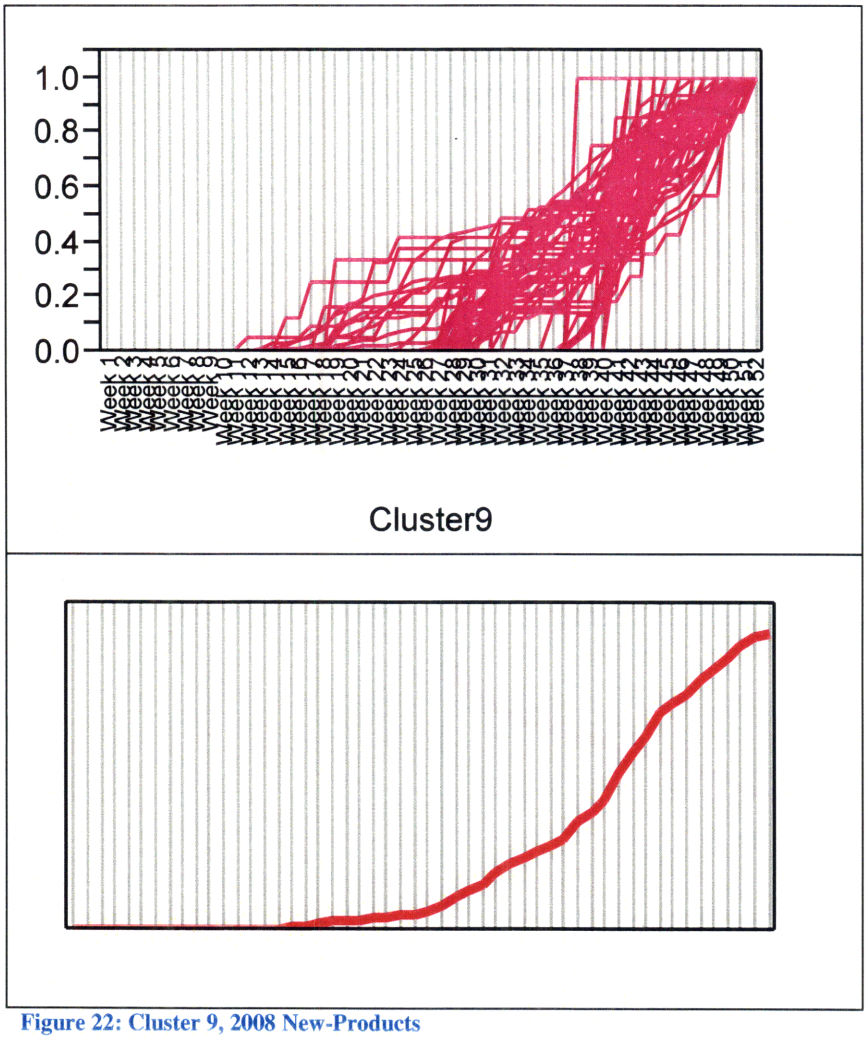

Figure 22: Cluster 9, 2008 New-Products

Cluster 9 in, **Figure 22,** consists of a small group of 65 members. The products appears to be launched anywhere from week 10 through week 40. The dominant trend is for the products to increase in sales from week 40 through the end of the year. However, in the exclusion of 2009 data, we cannot infer any substantial conclusions from this chart.

Cluster 9 (New Products)	Minimum Distance	Maximum Distance	Range
Mean	0	0.98598344	0.98598344
Standard Deviation	0	0.21026167	

Table 20: Cluster 9 Statistics – 2008 New Products

Cluster Means: 2008 New Products

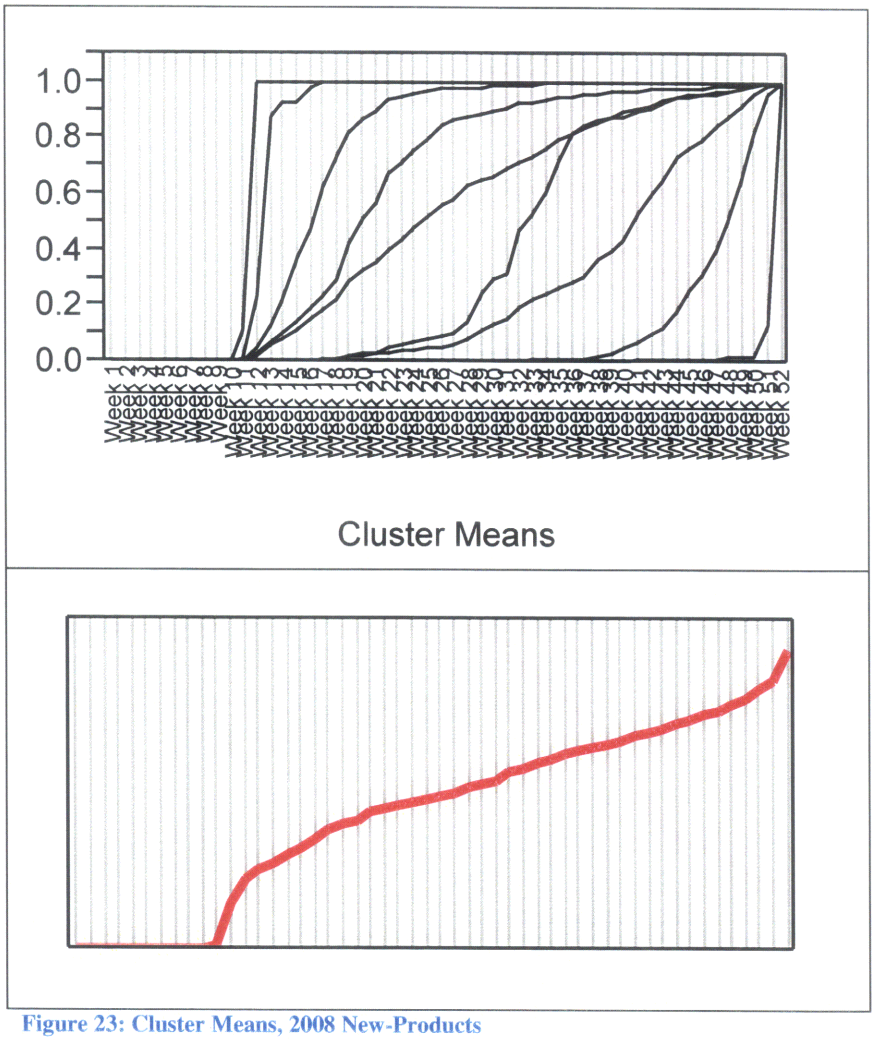

Figure 23: Cluster Means, 2008 New-Products

The Cluster Means, **Figure 23**, of the 2008 New Products showed that there are a number of clusters with products that sell strongly. However, it was not clear whether the holiday season has an impact on sales or whether these products continued to sell well into the 2009 year.

Conclusions

Exploring our two key approaches, Bass Model and Cluster Analysis, we arrived at a number of broad conclusions outlined below:

Bass Diffusion Model Analysis

Based our initial findings using a modified Bass Diffusion model we concluded that our particular retailer did not exhibit, at least over a 14 month period, trends that could be explicitly associated to traditional diffusion methods as described by Bass' original work.

Specifically, we noticed strong seasonality of cumulative sales changes over time for products in the 2008 online catalog. In addition, no saturation was found when we incorporated more recent data into our analysis. As such, cumulative sales increases appeared to decrease at the end of the sales season, not because the market has become saturated with a particular product. We feel that the longer life cycle of our retail partner's products made it necessary to look over a greater time horizon. This was impossible due to the constraints on obtaining data that could be trusted for any period reaching back more than one year. We feel that the Bass Model may be applicable to certain industries that observe a short product life cycle, but were unable to secure an appropriate research sponsor to examine the data. A potential industry that we felt the Bass Model would yield interesting insight was online retail fashion.

Clustering Analysis

The clustering analysis we performed yielded promising results. We were able to cluster almost 1000 products into nine clusters. In addition, with no industry specific knowledge, we were able to start to categorize the groupings. We look forward to sharing our insights with our retail partner and validate the clusters and find more linkages between the products in cluster. This initial clustering will also be helpful to start to look at the sales patterns within each cluster. Finally, these clusters are a starting point for having conversations with the retailer to find more relevant sets of data or have them start to log certain characteristics that come from the linkages that are apparent to the client within the clusters.

2008 Full-Year Cluster Analysis

Although we were not able to determine the optimal number of clusters for the 2008, full-year, sales data, we were able to identify 9 individual clusters. These clusters have particular characteristics that explain the groupings. The general characteristics are described in the following cluster table:

Cluster Number	Number of Members	Cluster Name	Cluster Description
1	149	Early Spring Gardening	Products with strong sales for winter gardening or early spring gardening preparation
2	228	Late Spring Gardening	Products with strong sales for mid-winter gardening or late spring gardening preparation
3	40	Normal Medium-All year and Strong Holiday	Medium all-year sales with holiday sales
4	192	Late Spring Gardening Discretionary	Products with strong, but less intense sales during late spring season. Possibly discretionary
5	64	Strong Medium-All year and weak Holiday	Products that are more essential during the entire year with strong
6	98	Winter stock out	Products that sold-out in Winter
7	45	Strong Holiday	Products that strictly showed strong holiday sales
8	146	Consistent All-Year	Mostly consumables with consistent all-year products with no high variability in demand
9	51	Strongest Medium-All year and weak Holiday	Strongest spring sales with rest of year performance and holiday sales

Table 21: 2008 Full-Year Cluster Descriptions

The cluster-names given in **Table 21** are not definitive, but an approximation, based on our sponsors system of promoting products during their particular sales season. In order to refine these categories, a full inventory of the catalog would be required. SKUs with full inventory characteristics would then be matched into their particular clusters, with the specific, common inventory characteristics acting as identifiers to the cluster.

2008 New Product Cluster Analysis

Cluster Number	Number of Members	Cluster Name	Cluster Description
1	204	Holiday Sellers	Products purchased specifically to drive holiday sales conversions
2	59	Summer Products	Products purchased in the spring and used through the August
3	87	New Consistent All-Year	New consumables products launched during the year.
4	37	Strong Spring	These products are launched in the spring, but sell out before end of summer
5	3	Excluded	We excluded these products due to the very small group member size
6	3	Excluded	We excluded these products due to the very small group member size
7	33	Late Launch	Due to the extremely late product launch and its affect on the cumulative sales figure, it is unclear whether there are any common characteristics
8	40	Spring-Summer with Autumn Sales	Products launched in late spring and summer for Autumn sales
9	65	Late Launch and slow	Products launched late during the year, but with slow sales growth appear to make up this particular group

Table 22: 2008 New Product Cluster Descriptions

Limitations

While working with any data there are always obstacles to overcome. We had no shortage of such things in our data and here we go into some of the limitations in our data set as well as some shortcomings of our model in its present state. Identified below are the key issues that we encountered with our research.

Data Integrity

- **SKU Changes** – Our sponsoring company informed us that they occasionally changed the SKU of an existing catalog product, as needed during the year. They do this for several reasons, including a minor update of the product, accounting reasons, or to minimize customer confusion. Examples of this include a second generation of a product that may have a slight modification, but does not necessarily warrant a new product for the purposes of our model. Other examples might be a new channel of distribution (or marketing) or possibly a new cost to a product. We found examples where a product that is offered in multiple colors might have 3 unique SKU's and then is migrated to one new SKU with a color indicator (from 123, 124, 125 to 200 Red, 200 Blue, 200 Green). The size of our research sample of 2008 and 2008 new products consisted of 1,643 products. Due to time constraints and catalog description constraints, we were not able to match each catalog ID to verify all possible SKU changes.

- **Fraudulent Charges** – Credit card fraud occurs on our partner's site. Although they typically correct the problem in their final numbers, neither Google Analytics nor the E-commerce software provider removes the orders from their transaction database. This is because it is typically caught while the product is being processed. Although we could account for the extreme cases (i.e. a $10,000 gift card) most fraud charges would be hard for us to detect.

- **Incomplete Catalog Details** – In order to check SKUs integrity, we reviewed the catalog SKUs and product description for all products 2,500 products in our sponsor's catalog. We secured this information from our sponsor's e-commerce software company. However, there were many products in the catalog that lacked a description or were marked with descriptions of #N/A. Such lack of descriptions made it harder to see full product descriptions for all members in a particular cluster.

Model Flexibility

During our research, we encountered a number of key issues that affected our ability to fully understand and verify our modeling approaches. Below, we identify some of the modeling issues that constrained our research efforts.

- **Marketing effects** - In or analysis we did not take into account the effect of marketing although we considered this an important measure. Our initial reasoning for this was that 1) the data was not readily available to us and 2) we felt that the sales trajectory would still be similar, but marketing would just amplify the total effect. With this in mind, we felt that normalizing the sales data would lessen the effect of the marketing campaign. However, within the time constraints of our research, we were not able to verify this initial assumption as valid and the degree to which it would affect the outcomes we were able to observe in both the diffusion analysis and the cluster analysis.

- **Promotions** – Retailers, in addition to generally marketing, may execute a number of actions to generate immediate response from customers. Usually these may be described as "promotions" and are implemented at the discretion of the retailer. Our sponsor offered several different promotions during the year to customers. These promotions included, free shipping, quantity discounts, and special bundles. We did not account for the effect of these promotions on either our diffusion or cluster analysis due to the constraints on time and access to specific promotion-history data from our sponsor.

- **Price changes** - Our retail partner does not typically manipulate the price of goods sold in their catalog. While talking to our thesis sponsor, we learned that typically price is drastically reduced in order to clear excess inventory or when they retire a product from the catalog. In addition, the price reductions typically happen on specific days determined by their chief marketing officer. We felt that our cluster analysis approach was robust enough to handle occasional price changes, as described by our thesis sponsor. However, we believe that there may be an upper bound of such discrete price changes, above which cluster analysis may not yield effective results. During our research, we were not able to test that boundary to fully account for how price changes would affect cluster behavior.

Next Steps

In general we feel that the information here can be used to better time marketing efforts of certain product clusters. We would like to see a trial where a subset of customers is marketed certain products that our clusters show to be optimally timed and compared to a control group of products that is picked by the retail partner with traditional methods. We would use metrics such as click through rate, conversion, and average order size to determine the success of our model.

One of the major issues that we ran into was the integrity and availability of relevant data. It seemed that a lot of information was contained within the company, but was not collected and stored in a meaningful way. The first area where we would like to see more information collected and analyzed is the attributes of the clusters. We feel our retail partner should observe the groupings and determine the exact attributes that are common in each cluster. This attributes might include dimensions, pricing, target segment, or other general characteristics that someone with industry knowledge would be able to quantify or categorize and store. We feel that using the clusters as a starting point to start talking about these features of the products will prove extremely helpful.

Beyond product characteristics we feel that starting to incorporate customer purchases over time would produce extremely useful data for the marketing department of our partner. Due to the limitations of the data we were unable to perform these steps over the course the thesis, but our work presented here is relevant. We feel that we can use the nine product clusters as a way of quickly finding patterns in purchasing. This will give the company a useful tool to analyze over several years a way to understand who is most likely to buy a new product given the cluster it would likely fall into based on product characteristics. This type of information could lead to tailored efforts for various customer clusters. Currently our partner uses rudimentary segmentation of their customer list to decide what promotions to send to whom.

Finally, one critical set of data to incorporate into any recommendations based on our clustering is the goal of the retail partner. Information such as margin and amount of inventory could be included in order for the retailer to optimize the model based on a goal of maximizing revenue, maintaining a certain margin, or eliminating excess inventory to reduce holding costs. This type of information would have to be included in any recommendation in order to remain relevant and give tradeoffs between different recommendations.

We feel that the timing of marketing effort and the addition of better customer segmentation will result in a revenue lift to our retail partner. Using our methods we feel that marketing will be able to present more

relevant communications to existing and potential customers that will result in higher click through, conversion, and average order sizes.

In addition to the potential mentioned above, our models can be used to predict and monitor the pattern of sales of new products. Our models can be used to monitor across thousands of SKU's on a daily basis. Using the predicted total sales of a product and the cluster that it falls into we would be able to help with order timing and inventory control. Finally, we could provide some level of revenue management by recommending adding more products to a certain cluster in order to boost revenues at certain times of the year. Revenue management can be particularly helpful for retailers that have strong seasonality to their sales.

Bibliography

Bass, F. M. (1969). A New Product Growth Model for Consumer Durables. *Management Science* , 215-227.

Christopher J. Easingwood, V. M. (1983). A Nonuniform Influence Innovation Diffusion Model of New Product Acceptance . *Marketing Science* , 273-295.

Galit Shmueli, N. P. (2007). *Data Mining for Business Intelligence.* New Jersey: Wiley.

Muller, V. M. (1979). Innovation Diffusion and New Product Growth Models in Marketing. *Journal of Marketing* , 55-68.

Teck-Hua Ho, S. S. (2002). *Managing Demand and Sales Dynamics in New Product Diffusion Under Supply Constraint.* Pennsylvania: The Wharton School.