THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF SUPPLY CHAIN AND INFORMATION SYSTEMS

IMPROVING VEHICLE UTILIZATION THROUGH THE USE OF DATA MINING

HENG LI
Spring 2011

A thesis
submitted in partial fulfillment
of the requirements
for baccalaureate degrees
in Supply Chain and Information Systems; Economics
with honors in Supply Chain and Information Systems

Reviewed and approved* by the following:

Robert Novack
Associate Professor of Supply Chain and Information Systems
Thesis Supervisor

Anup Sen
Visiting Professor of Supply Chain and Information Systems
Thesis Supervisor

John Spychalski
Professor Emeritus of Supply Chain and Information Systems
Honors Adviser

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

Transportation, or the process of transferring materials, goods, and services from one location to another, is becoming of increasing importance in logistics optimization and supply chain management. With expanding development, management, and application of database technology and information systems, enterprises can use statistical and data mining techniques to trace out their transportation patterns. The results help reveal and highlight hidden trends, changes, and correlations which may help identify opportunities for improvements in both transportation and the overall supply chain network.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENT

First and foremost I would like to thank Dr. Robert Novack and Dr. Anup Sen for not only providing me the opportunity to work with them, but also for guiding this research. Second I would like to thank Mr. Hui Liu for assisting me with some of the technical aspects of the project and showing me all the exciting aspects of using data mining software. My gratitude also goes out to Mrs. Tracie Shannon from the Center for Supply Chain Research for facilitating this project and, furthermore, to the representatives of the company for which this study was conducted for providing me with data, clarifications, and recommendations.

I would also like to take this chance to thank all my professors, classmates, and friends in State College who aided me academically, professionally, and personally throughout the years. Finally, I thank my dear parents and guardian for supporting my education in the United States and offering me a plethora of advise and support over the years.

Chapter 1

INTRODUCTION

1.1 Importance of Transportation in a Supply Chain Network

Modern logistics involves an ever-growing complex network in which transportation, warehousing, distribution, packaging, information processing and other related links are interconnected. Among all of the links in a supply chain network, transportation plays a crucial part in logistics management and is a key determinant of the operation's efficiency (Tseng, Yue, & Taylor, 2005). Transportation is, by definition, the process of transferring materials, goods, and services from one location to another. However, with logistics optimization becoming of increasing importance, the transportation process has changed. Transportation is not only the movement of goods and services between locations but also a focal point of strategic and tactical decision making for modern enterprises. Improvements made in the transportation process can help businesses optimize fill rates, shorten lead-time, increase efficiency, and reduce carbon footprints, all of which cut costs and enhance overall customer service.

More and more enterprises begin to recognize the importance of improving their transportation processes. These companies are therefore investing heavily in expanding development, management, and application of database technology and information systems. These developments allow the companies to record transactions in every business process, including the transportation transactions. The next logical step involves accessing the data and extracting valuable information to help businesses reveal and highlight hidden trends, changes, correlations, and risks, which may derive actionable business

intelligence (Wu, 2002). Whether or not an enterprise can effectively utilize information collected in the data warehouse and make proactive, knowledge-drive decisions has become a vital differentiator of corporate success (Mathew, 2005). With that being said, the purpose of this thesis is to:

1. Utilize various data analysis techniques to explore the data extracted from the data warehouse from a consumer packaged goods manufacturer;
2. Develop a data analysis methodology that this manufacturer can utilize for further transportation network analysis.

The remainder of the thesis will identify and discuss current issues in transportation and introduce the background of the consumer packaged goods industry.

1.2 Current Issues in Transportation

Before exploring methods to create improvements in transportation, it is helpful to first look at some of the many challenges and risks currently faced by the motor carrier transportation industry, because in recent years, the most-discussed topics in this field are driver shortage, security, and fuel prices.

The driver shortage, which is coming from both the demand side and the supply side, is slowing the future growth of motor carriers (Solomon, 2010). This is a concern to the entire logistics industry since motor carriers generate about 31% of total ton-miles in freight transportation (Coyle, Novack, Gibson, & Bardi, 2011). In other words, nearly one third of the transportation field could potentially be affected by the expected shortfall of qualified motor carrier operators.

On the demand side, as the economy recovers from the severe downfalls seen from 2007 to 2009, quarterly reports from many freight carriers such as United Parcel Service, JB Hunt, Landstar System, and CSX Corporation consistently show increasing freight demand (Vertical Alliance Group, Inc., 2011). This increase in freight demand then leads to a higher demand for drivers.

On the supply side, there are two major factors: (1) the Comprehensive Safety Analysis 2010 (known as CSA2010); and, (2) the retiring baby boomers (Vertical Alliance Group, Inc., 2011). The CSA2010, administrated by the Federal Motor Carrier Safety Administration (FMCSA), was implemented to improve the overall safety of motor carrier operations. This updated system documents and inspects both driving records for individual drivers and the overall fleet scores of the carriers. Drivers and motor carriers with poor records could be suspended. Such new regulation is expected to crimp the supply of qualified and eligible drivers (Vertical Alliance Group, Inc., 2011). Then there is the aging workforce. The imminent retirement of the baby-boomer generation will soon cause a dramatic shortfall in labor supply and therefore a decrease in the supply of drivers in the labor market.

Facing this expected driver shortage, motor carriers will have to increase their investment in recruitment of qualified drivers while also retaining eligible drivers by maintaining satisfactory benefits. While competing for qualified and eligible motor carrier operators, companies are also challenged to run their businesses more efficiently in order to sustain continued growth with this dwindling labor supply.

Next, security has long been a concern in the freight industry. In the emerging global supply chain network, the emphasis within freight transport security is on the

integrity of the cargo, the route of the cargo, and the information management of the supply chain. Problems in freight security mostly revolve around the increasing risks of smuggling, piracy, and sub-standard vehicles. The issue of freight security is extremely difficult to address because of the large number of origins and destinations, the number of carriers, and the wide range of products carried. Dealing with so many variables makes it hard to detect security risks. While the best solution is yet to be found, implementing multiple points of inspections and raising the security standards of facilities, personnel, and the data are some ways to improve the security of freight transport. (Rodrigue, Slack, & Comtois, 2009).

Finally, escalating fuel prices are increasing the cost of tranportation and reducing industry profitability. Since all modes of tranportation rely on fuel, fuel surcharges have been implemented in every mode of transportation in order to combat increasing costs. Moreover, expensive fuel negatively affects manufacturers and suppliers which are the sources of freight demand. Therefore, climbing energy prices concern the entire network. Due to the low likelihood of a new fusion of energy replacing the dominance of fossil fuels in the next few decades, freight carriers are forced to commit to more energy-efficient strategies and must find new ways to run their businesses more effectively.

Given the aforementioned costs and risks faced by transportation, it is clear that an improvement in transportation efficiency would help companies sustain future growth and remain competitive in the market. However, depending on elements such as price elasticity, cost structure, and product/service demand, companies in different industries may develop very different strategies to increase their transportation efficiency. For example, industries that deal with innovative products (e.g. fashion and high-tech

electronics) face very unpredictable demand and focus mainly on responsiveness. Therefore, being unable to ship products in bulk, it is harder for companies in these industries to maximize vehicle-fill rates or reduce the number of delivery trips. On the contrary, industries that serve functional products see very predictable consumer behavior and are able to design distribution and delivery plans that would enable shipments to have maximized vehicle-fill rates.

1.3 Background of the Consumer Packaged Goods Industry

The consumer packaged goods industry deals primarily with functional goods, including food, beverages, apparel, tobacco, and cleansing products, all of which get consumed and require frequent replacement (TechTarget, 2010). For this particular industry, there are three keys to success: managing retail customers, managing consumers, and managing supply chains (McKinsey & Company, 2010).

In the supply chain of today's world, retailers have an ever-expanding influence on manufacturers, with Wal-Mart being one of the best-known examples. Consumer packaged goods companies are left with very little margin for error and are constantly being pushed to eliminate waste and inefficiency in every way they can. This drive challenges companies to develop a strong and effective supply chain system. Hence, the top-line growth of consumer packaged goods companies has an increased emphasis on logistics optimization and supply chain management. Being a crucial part in logistics management and a key determinant of an operation's efficiency, improvements made in the transportation process directly reflect a company's ability to cut costs, optimize vehicle fill, shorten lead-time, increase efficiency, reduce carbon footprints, and enhance the overall level of service delivered to their customers.

Chapter 2

ACCESSING THE DATA

2.1 The Objective

The selected consumer packaged goods manufacturer has invested heavily in building information systems that collect all sorts of data throughout its transportation process. However, the company has yet to develop a methodology to weed through massive amounts of data, trace out existing patterns, and identify areas of opportunities. This thesis will apply different analysis techniques to explore the data extracted from the data warehouse and try to develop a data analysis methodology that this manufacturer can use for further transportation network analysis.

2.2 Data Description

This analysis focuses on a particular set of 65,533 shipment data entries collected from Sunday, November 29, 2009 through Saturday, February 6, 2010. Recorded shipments are all performed by a variety of different motor carriers. These data, extracted from the company's data warehouse, recorded the following information for each shipment:

- week number,
- actual shipment date,
- shipment type,
- customer name,
- customer location,

- mode of transportation,
- carrier used,
- SAP number,
- shipment origin and destination,
- stop count[1],
- product family.

The set of data consists of two types of shipments delivered from various plants and distribution centers to locations across forty-eight continental states: customer shipments and interplant shipments. A customer shipment is delivered to a customer distribution center while an interplant shipment is loaded with goods that need to be transferred from one plant to another. Consequently, these two types of shipments may have a very different mix of products. Whether a shipment is going to a customer or going through interplant transfers could affect the vehicle-fill rates of the trailer. Therefore, the two types of shipments need to be analyzed separately. This research will only focus on the customer shipments.

The selected customer-shipment data covers nine different product families ranging from fabric care to snacks[2]. Various product families involve different types of commodities. The density of each load will have a noticeable difference based on the

[1] The number of stops that a particular shipment goes through between its origin and final destination.

[2] Data included a tenth group for all shipments with an "unassigned" product family. The "unassigned" family is considered an outage in the master data systems. This research ignores all of the unassigned shipments.

nature of the products. Therefore, it would make sense for the firms to use a separate set of order/shipment policies for each product family.

Also contained within the data were utilization measurements, such as, shipment total weight, shipment total cube[3], shipment total floor positions[4] (FP), and shipment total cube ordering factor[5] (COF). Generally, a trailer is capable of containing goods with a net weight less or equal to 45,500lbs and a volume not to exceed 3,244 ft^3. A full load offers 30 floor positions and has a 3,750 COF[6]. The data system divides each shipment total number (e.g. weight or volume) by the relevant threshold in order to generate vehicle-fill rates (or percentages) of weight, cube, floor positions, and cube ordering factor. The maximum value from these four vehicle-fill metrics is that particular shipment's constrained vehicle utilization (constrained VU). For example, a shipment that fills up 90% of the weight capacity, 47% of the cube capacity, 93% of floor positions, and 91% of its cube-ordering factor will have a constrained VU equal to 93% since that is the maximum value of the four metrics. Because of the diversity of the product families involved in these shipments, measuring the constrained VU is an extremely helpful way of accessing the capacity utilization of the trailer. Depending on the products being

[3] Same as volume.

[4] Floor levels set up for pallets.

[5] The Cube Ordering Factor is a value assigned to the shipping case that measures the item's size relative to the other products being shipped. The COF represents the amount of space taken up by an item in the truck and is therefore useful for allocating freight costs. The COF for each shipping unit is determined by dividing the quantity of that item needed to fill a standard size truck by an arbitrarily defined "full-truck" quantity.

[6] For certain product families, the maximum COF can be different. For example, the maximum COF for pet care products is 3200.

shipped, each shipment may have a very different density. As such, some products may weigh out the trailer while others may cube out the trailer. In other words, it is possible for a shipment to fill up 100% of the volume but at the same time only meet 10% of the maximum weight capacity (e.g. a shipment of cotton). Therefore, focusing on just one of the four metrics (i.e. weight, cube, FP, or COF) will not reveal the true vehicle utilization of a particular shipment. The constrained VU is a metric that allows comparisons across shipments with different density, floor-position usages, and package sizes.

2.3 Explorative Data Visualization

After understanding the elements involved in the data, data were visualized in order to perform an initial assessment. This is an effective method to attain a first-impression of the data, especially when analyzing such a large data set. Generating scatter plots that depict the relationships between variables may be the only analytic method needed to reveal hidden patterns, trends, and correlations.

For the initial assessment in this particular case, it was helpful to analyze the relationship that values of each of the four metrics have with respect to the constrained VUs. This assessment helped recognize whether or not one or multiple metrics consistently dominated and dictated the constrained vehicle utilization. It was also possible to find that one or multiple metrics out of the four were constantly constrained by the other measures, which means there is still “space” in the trailer even though some other metrics had reached their maximum capacities. To create visualized assessment, the entire set of shipment data was graphed on four scatter plots where X-axes always represent constrained vehicle utilizations and the four Y-axes each represents one of the

four metrics measured by weights, cubes, floor positions, and cube ordering factors. On each scatter plot, different colors were used for data points representing different product families. Results are shown in Figure 1 through Figure 4:

Figure 1. Weight vs Constrained VU

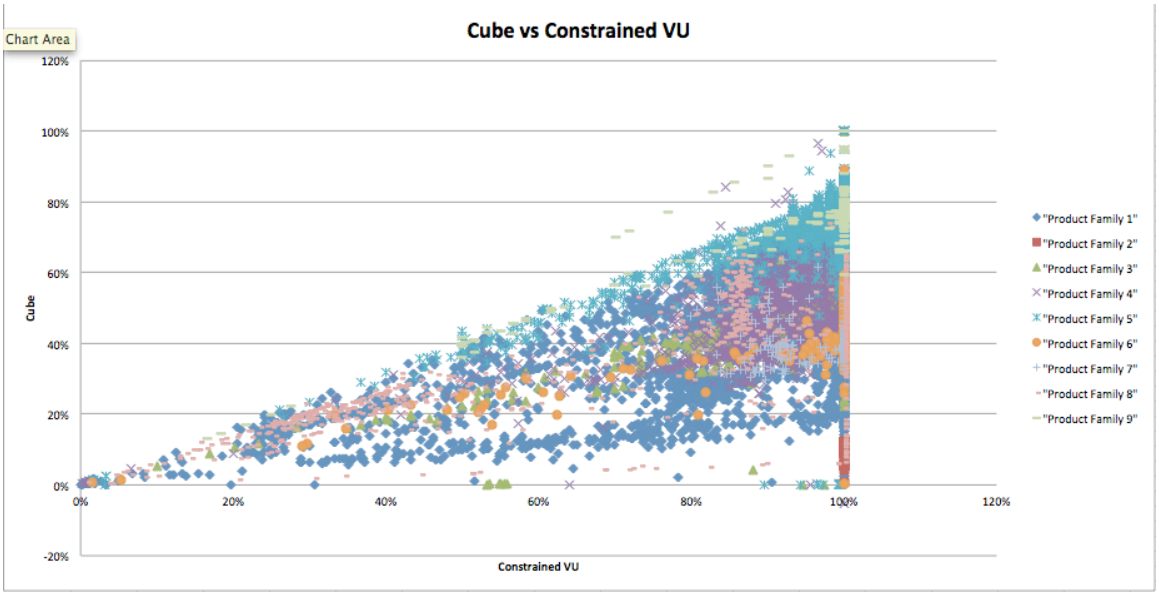


Figure 2. Cube vs Constrained VU

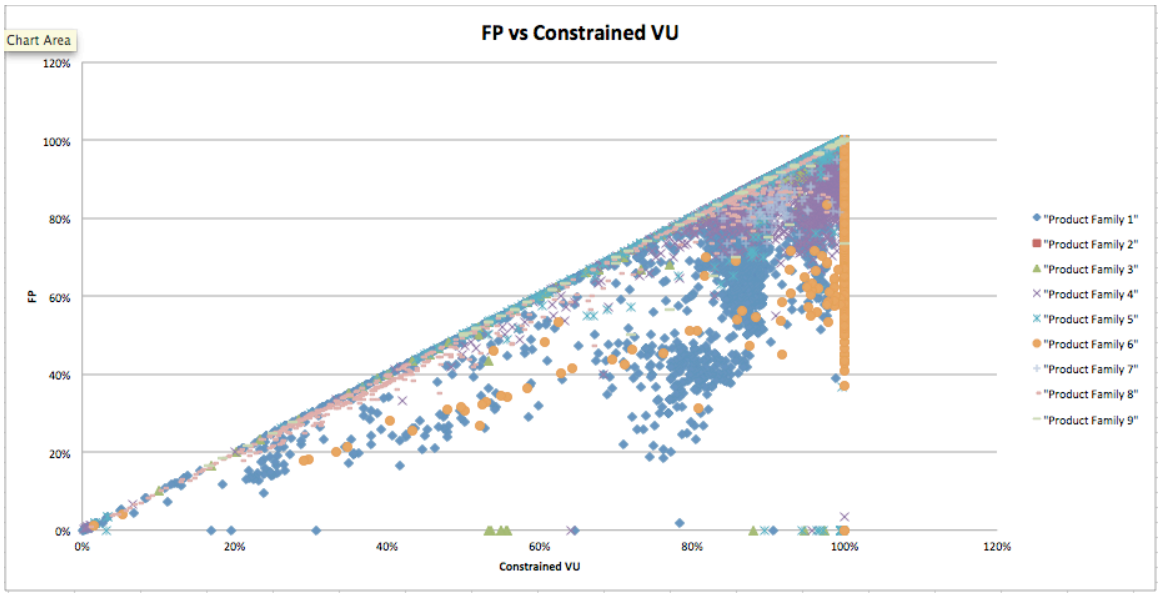


Figure 3. FP vs Constrained VU

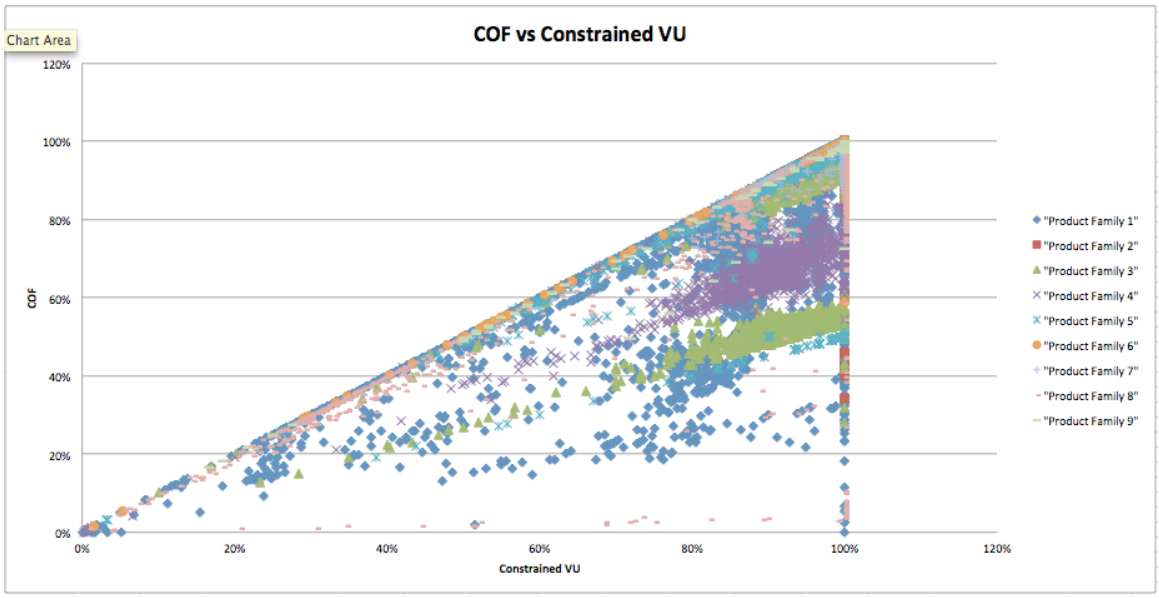


Figure 4. COF vs Constrained VU

As explained in Section 2.2, the constrained vehicle utilization is the maximum value of a single shipment's weight rate, cube rate, floor-position rate, and cube-ordering-factor rate. Therefore, any given shipment cannot have a single value of these four rates that is greater than its constrained vehicle utilization. The series of graphs shown above reflect this characteristic. In each graph, all the data points are within the triangle bounded by the line Y = X, the line X = 100%, and the X-axis.

Based on the description above, it is ideal to have most data points concentrating at the paramount area in every graph, which would indicate that majority of shipments are optimizing their vehicle utilization values in every aspect. However, that is clearly not the case shown in the four graphs. Figure 1, representing the ratios of weight rates over constrained vehicle utilization values, shows that data points widely spread within the triangle. At this point, no specific patterns between weight rates and constrained vehicle utilization values can be detected. From Figure 2, it is clear that the data plots for cube rates versus constrained vehicle utilization values form a big gap right under the Y = X line, which means that the shipments' vehicle utilizations were typically not dominated by its volumes. In other words, cube rates of this set of data are consistently constrained by other measures, and there are many cases in which there is still unfilled volume in the trailer when the other measurements have high rates. Figure 3 and Figure 4, on the other hand, show two sets of data plots that are relatively more concentrated under the Y = X line, which indicates that vehicle utilizations are more often dominated by a shipment's floor-position rate and cube-ordering-factor rate.

Further, each graph is broken down into a series of nine product-family-specific graphs illustrating corresponding relationships between constrained vehicle utilization

and weight, volume, floor position, or cube ordering factor (see Appendix A-D[7]). From the detailed graphs, some more specific and distinctive patterns are detected:

1. **Product Family 1** has an extreme amount of variability in vehicle-fill rates compared with all other product families. Clearly seen in Appendix A1, B1, C1, and D1, data points cover a wide area within the triangles. Floor-position rates seem to have a more significant dominance in constrained vehicle utilization values, as shown in Appendix C1, data points of the *Weight vs Constrained VU* graph have less wide of a spread within the triangle and concentrate relatively more towards the Y = X line. Similarly, data points of **Product Family 8** also contain a large amount of variability as illustrated in Appendix A8, B8, C8, and D8.
2. **Product Family 2** is strictly dominated by the floor-position rates. In Appendix C2, it is clear that all data points fall on the point where floor-position rate = constrained vehicle value = 100%. This leads to the shapes of Appendix A2, B2, and D2. In these three other graphs, data points scatter on the line where constrained vehicle utilization values are always equal to100%.
3. **Product Family 3**, **Product Family 4**, and **Product Family 7** have the most consistent vehicle-fill rates across all metrics. Their scatter-plot charts show that most of shipments in these three product families have *Weight vs*

[7] Appendix A includes detailed *Weight vs Constrained VU* graphs of nine product families; Appendix B contains detailed *Cube vs Constrained VU* graphs of nine product families; Appendix C comprises detailed *FP vs Constrained VU* graphs of nine product families; and Appendix C covers detailed *COF vs Constrained VU* graphs of nine product families.

Constrained VU, *Cube vs Constrained VU*, *FP vs Constrained VU*, and *COF vs Constrained VU* ratios that mostly concentrate in the paramount areas in the triangles.

4. Most of graphs shown in Appendix A – D confirm that cube rates overall have the least dominance in constrained vehicle utilization values. Most of the *Cube vs Constrained VU* ratio scatter plots repeatedly show that cube rates in many product families consistently score below the line Y = X.

These patterns listed above are very helpful for initial assessment of this large data set. However, knowing that some of the product families have high variability in vehicle fill-rates and that cube rates are constantly constrained by other metrics is not enough evidence for further decision making by the manufacturer. One option of the next logical step to further the search of more hidden trends, correlations, and patterns is data mining.

Chapter 3

DATA TRANSFORMATION

3.1 Background of Data Mining

Data mining, defined as the extraction of hidden predictive information from large databases, is a very useful tool to effectively trace out utilizable information from large amounts of collected data. For that reason, it has gained increasing importance in both resource planning and decision-making processes in today's business world (Mathew, 2005). Some of the applications of data mining in supply chain management include forecasting the market and trends, reducing inventory costs, improving efficiency, classifying customer classes, analyzing consumer values, and optimizing transportation fill-rates and paths (Fei, Zhang, & Zhou, 2010).

Despite being defined as a sub-domain of classic statistical analyses that tries to find hidden patterns of data, data mining does not require a predetermined hypothesis like any other classic statistical analyses. However, it is similar to the statistical approach in a sense that it requires defining a business goal or problem in order to effectively develop a focus and efficiently apply the right procedures to the data analysis (Mathew, 2005). Also, data mining should be a continuously ongoing process. It is critical for enterprises to frequently revisit their databases and make adjustments based on new facts and figures from the information systems.

This thesis is particularly interested in the patterns of four vehicle-fill metrics and how they constrain each other, especially when vehicle utilization is low. So, for this

particular data mining study, the objective is to explore and analyze information extracted from the database and try to identify specific correlations and patterns of the 65,533 recorded shipment entries under the four vehicle-fill metrics and constrained vehicle utilization. Hopefully, the correlations and patterns found could help discover potential opportunities to increase vehicle fill and in turn improve overall transportation efficiency.

3.2 Preparing the Data for Mining

Before proceeding to the actual mining part of this project, the data set needed to be cleansed in a sense so that only the information useful for the analysis remained.

The first step of preparation was to eliminate elements that were less important to the analysis and focus on the ones that were critical in the search of significant associations. As described in Section 2.2, the set of data recorded many variables of each shipment including information regarding the dates, locations, carriers used, SAP numbers, etc. All of these variables may or may not affect vehicle fill in some way, therefore elimination of any variable should be done under careful consideration. It is very possible that omitting important variables that were not thought to be significant can cause problems in the data mining results. Oftentimes, it may be an option to run multiple trials to determine whether or not some of the variables are significant to the results. Since this thesis is particularly interested in the patterns of four vehicle-fill metrics and how they constrain each other, the data mining process focused on each shipment's four vehicle-fill rates and its constrained vehicle utilization.

The second step involved dividing the data into different sections that may be independent from each other. By separating the data into smaller groups, data analysis

could be performed on one group of the data at a time. This step is especially important when navigating through a huge set of data. Keeping numbers that are evidently independent from each other together only adds unnecessary complication to the analysis; separating them will increase the chance of findings. For this set of data, shipments were divided based on the nine different product families. Various product families involve different types of commodities. The density of each load will obviously differ based on the nature of the products. It is a common practice for companies in the consumer packaged goods industry to have different sets of ordering requirements for different product families. Therefore, it made sense to divide the data based on product families and to perform data analysis on one product family at a time.

This thesis began the mining process by focusing only on the data regarding Product Family 1. The same methodology can apply to the other product families in future studies.

With the selected set of data in hand, the next step was to take out any values that were corrupt, inaccurate, or inconsistent. This necessary step is known as data cleansing. For this selected set of data, all entries with zero or negative input values for any of the vehicle fill rates and constrained vehicle utilization were eliminated. Any of these zero and negative values are clearly input errors.

Since this thesis was interested in shipments that did not fill up the trailers and intended to offer suggestions to improve vehicle utilization, it was helpful to take out the records with high vehicle-fill rates and only focus on those that have unused capacity. Hence, other than cleaning out the corrupt, inaccurate, or inconsistent values, the data

analysis also excluded all shipment records with constrained vehicle utilization values that equaled 95% or higher.

After a series of preparation steps, the final selected set of data was ready for the data mining process. The final data set included 3996 Product Family 1 observations of vehicle-fill rate and constrained vehicle utilization values, which are greater than zero and lower than 95%.

3.3 Data Mining

For the purpose of this thesis, the data mining process engaged the affinity analysis technique to extract association and co-occurrence rules based on statistical significance. The first step of this mining process was to discretize[8] the weight rates, cube rates, FP rates, and COF rates into categories. Each category was discretized based on its percentage weight of total selected observations and reflected the scale of the numeric values. Computation was done on Microsoft Excel, as attached in Appendix E, and followed these steps:

1. Record frequency counts of weight rates, cube rates, FP rates, and COF rates based on 5%-scale intervals
2. Calculate accumulated counts
3. Compute accumulated weight of total selected observations
4. Discretize results into categories

[8] Discretization concerns the process of transferring continuous models and equations into discrete counterparts. This process is usually carried out as a first step toward making them suitable for numerical evaluation and implementation on digital computers.

The results of discretization were as follows:

Weight Rate Discretization	
0<x<=25%	WGT 1
25%<x<=50%	WGT 2
50%<x<=80%	WGT 3
80%<x<=85%	WGT 4
85%<x<95%	WGT 5

Table 1. Weight Rate Discretization

Cube Rate Discretization	
0<x<=15%	CUBE 1
15%<x<=30%	CUBE 2
30%>x>=50%	CUBE 3
50%>x>=55%	CUBE 4
55%>x>=60%	CUBE 5
60%>x>=75%	CUBE 6
75%>x>95%	CUBE 7

Table 2. Cube Rate Discretization

FP Rate Discretization	
0<x<=40%	FP 1
40%<x<=60%	FP 2
60%<x<=65%	FP 3
65%<x<=70%	FP 4
70%<x<=75%	FP 5
75%<x<=80%	FP 6
80%<x<95%	FP 7

Table 3. FP Rate Discretization

COF Rate Discretization	
0<x<=40%	COF 1
40%<x<=55%	COF 2
55%<x<=65%	COF 3
65%<x<=70%	COF 4
70%<x<=75%	COF 5
75%<x<=85%	COF 6
85%<x<95%	COF 7

Table 4. COF Rate Discretization

The logic behind the categorization was to scale the numeric levels of the four vehicle-fill metrics by equal statistical weight. Important to note, since the discretization particularly focuses on the data with lower numeric values, the metric values with high numeric levels might be categorized in the same group leaving that category with extremely high statistical weight. However, the analysis was specifically interested in the associations involving the lower-value categories, therefore the fact that high-value categories have a larger weight of the observations should not affect the mining results. For example, the *Weight Rate Distribution* chart (Figure 5 shown below) shows that most of the shipments, 2328 out of 3996 observations, utilized 85% to 90% of the trailer's weight capacity.

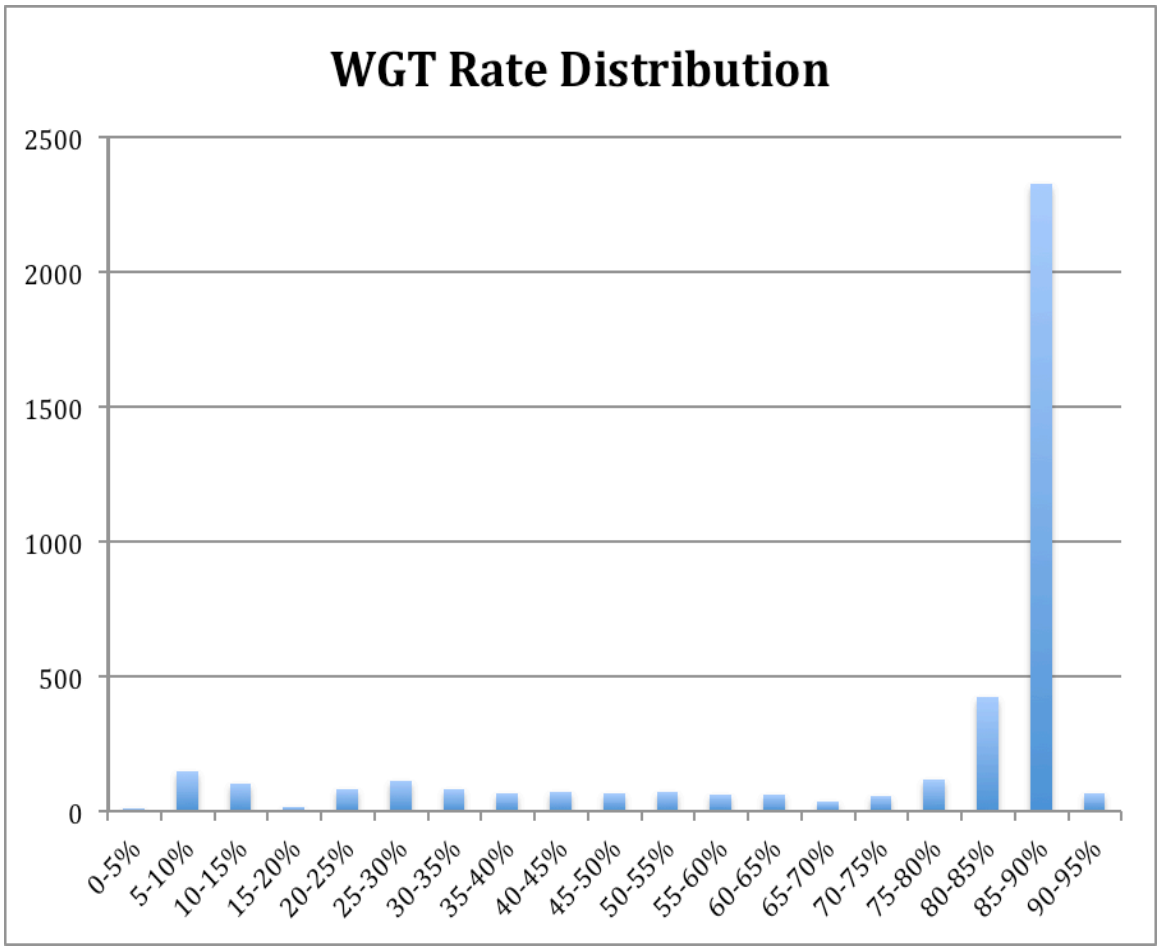


Figure 5. WGT Rate Distribution

After discretization, the numeric levels of weight rate were categorized into a scale that would smooth out the distribution of the low-value levels.

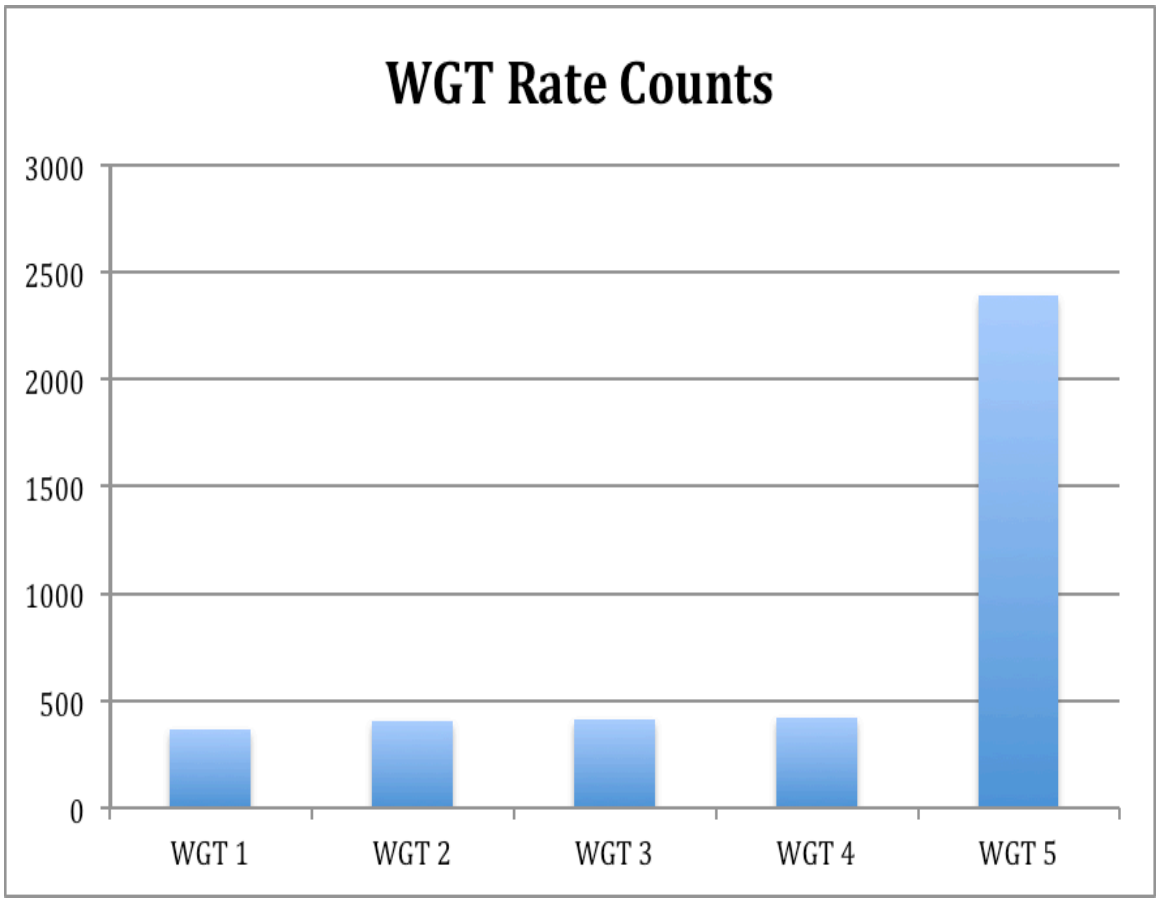


Figure 6. WGT Rate Counts

As shown in the new distribution after discretization (see Figure 6 above), categories WGT 1, WGT 2, WGT 3, WGT 4 had nearly equal statistical significance in the data. Category WGT 5, having a much higher weight, would not affect the results since this analysis was less interested in shipment data with high vehicle-fill rates. Cube rates, FP rates, and COF rates were discretized under the same logic.

Most of the lower categories uniformly represented approximately 10% of the total observations; therefore, the maximum support[9] of the affinity analysis was set to 0.15 to cover the co-occurrences of any of these lower categories. Besides support, another metric that could greatly affect the results was confidence[10]. For this analysis, minimum confidence was set at 0.5; in other words, any rules found with a lower-than-50% confidence were considered coincidences that indicated no significant application. It should be noted that data mining analysis usually involves many explorative trials. Oftentimes, it is helpful to experiment with a few runs by setting different support and confidence bounds to see if any association rules arise.

Affinity analyses for this project were all done on WEKA[11]. The first explorative test was run with the listed discretization results as input; support was constrained within 0.01 to 0.15, and minimum confidence was set to be 0.5. There were 76 outcomes found from this first explorative run, and they are listed in Appendix F. As shown, the highlighted outcomes were the ones with occurrence count over 100 and were thought to be more significant rules. Below is the list of results that carried more weight (copied from the run information in Appendix F):

1. Wgt Tot=2 FP=1 105 ==> COF=1 105 conf:(1)

[9] The *support* supp(X) of an itemset X is defined as the proportion of transactions in the data set which contain the itemset. In the example database, the itemset {milk,bread,butter}has a support of 1 / 5 = 0.2 since it occurs in 20% of all transactions (1 out of 5 transactions).

[10] The *confidence* of a rule is defined conf($X \Rightarrow Y$)=sup($X \cup Y$)/sup(X). For example, the rule {milk, bread}⇒{butter} has a confidence of0.2 / 0.4 = 0.5 in the database, which means that for 50% of the transactions containing milk and bread the rule is correct.

[11] Data Mining Tool used is WEKA (Version 3-6-4) operated on Windows 7.

4. Cube=1 FP=1 110 ==> COF=1 109 conf:(0.99)
9. FP=1 311 ==> COF=1 305 conf:(0.98)
10. COF=2 323 ==> FP=2 316 conf:(0.98)
11. Cube=2 FP=1 201 ==> COF=1 196 conf:(0.98)
12. Cube=2 FP=2 142 ==> COF=2 138 conf:(0.97)
15. Cube=2 COF=2 145 ==> FP=2 138 conf:(0.95)
22. FP=3 165 ==> COF=3 147 conf:(0.89)
23. COF=4 257 ==> FP=4 228 conf:(0.89)
26. Cube=2 COF=1 228 ==> FP=1 196 conf:(0.86)
31. COF=5 529 ==> FP=5 443 conf:(0.84)
34. FP=5 532 ==> COF=5 443 conf:(0.83)
35. COF=1 370 ==> FP=1 305 conf:(0.82)
37. FP=4 278 ==> COF=4 228 conf:(0.82)
47. FP=2 454 ==> COF=2 316 conf:(0.7)
56. FP=1 311 ==> Cube=2 201 conf:(0.65)
57. FP=1 COF=1 305 ==> Cube=2 196 conf:(0.64)
59. FP=1 311 ==> Cube=2 COF=1 196 conf:(0.63)
60. COF=1 370 ==> Cube=2 228 conf:(0.62)
63. Cube=1 244 ==> COF=1 142 conf:(0.58)
65. Cube=1 244 ==> Wgt Tot=2 139 conf:(0.57)
70. COF=1 370 ==> Cube=2 FP=1 196 conf:(0.53)
72. Cube=2 440 ==> COF=1 228 conf:(0.52)
73. COF=3 285 ==> FP=3 147 conf:(0.52)

As underlined in orange, the rules found show that lower-category FP rates and lower-category COF rates were very correlated. When the Product Family 1 shipments were in categories FP 1, FP 2, FP 3, FP 4, and FP 5, meaning they have an FP rate greater than 0 and less than or equal to 75%, the COF rate was somewhat predictable. Also, as highlighted in green, COF rates mostly fell into the same category of the FP rate as well, when the COF rate was less or equal to 75%. WGT categories and CUBE categories, however, did not show any meaningful patterns. Any rules that involved WGT categories or CUBE categories produced inconsistent cross-occurrences.

For the purpose of explorative experiment, it was appropriate to also include a run with a higher-bound support to see if that would help discover more rules. With that

being said, the second run adjusted the support control to between 0.025 and 0.25, and minimum confidence remained at 0.5. The outcomes are listed in Appendix G. No major significant rules were found.

To ensure the thoroughness of the research, the analysis also includes another run with support set to be 0.25 to 1.0 to make sure that no other significant findings will be missed; and the outcomes are listed in Appendix H. Similar to the second run, no major associations were found.

3.4 Results and Discussion

The discretization process revealed that a majority of the shipment data for Product Family 1 showed high vehicle-fill rates in weight. Their cube rates were relatively low and all were below or equal to 75%. In other words, Product Family 1 products have relatively high density and were most likely to weigh out the trailer rather than cube it out. For FP and COF rates, the numbers were heavily distributed between the 70% - 90% intervals. Also, it was clear that for most of the shipments in Product Family 1, this company managed to utilize trailer capacity very well by weight and utilized approximately 80% of floor positions. Although the overall vehicle-fill performance is not bad, the company certainly has room for improvement in vehicle utilization.

Through mining runs focusing on lower-category vehicle-fill rates, low-category FP rates were found highly correlated with the respective-category COF rates. In other words, when the trailer had more than 30% of utilized trailer capacity in terms of FP and COF, the floor-space utilization depended mostly on the relative size of this product's shipping case. Since more than 40% of the Product Family 1 shipments had FP and COF

rates in the lower category, this finding might be helpful to management for any future updates of ordering policies regarding this particular product family.

Moreover, data mining involves many explorative experiments, and significant results might be found after much trial and error. It is clear that more mining experiments are needed for this project in order to find more meaningful patterns and actionable information.

Chapter 4

CONCLUSION

4.1 Summary

With the increasing investment and use of data warehouses, data mining has increasing significance in helping companies to gain competitive advantage in today's market. For consumer packaged goods companies, improving vehicle utilization and the overall transportation performance can help cut costs and increase efficiency and in turn expand a company's competitive advantage.

A key aspect learned from this explorative data analysis was that data mining is a time-intensive and experimental process in a lot of cases. Meaningful patterns are usually hidden and actionable information is hard to find. Hence, it is sometimes difficult to determine what elements have little influence on the outcomes and on what variables the data mining process should devote its focus.

4.2 Limitations and Future Research

One possible approach to enhance the affinity analysis done with the current data would be to include more variables in the mining process. It is very possible that some of the variables eliminated, such as stop counts and shipment dates, could affect the patterns of the truck-fill rates. Also, this thesis focused strictly on all the customer shipments of Product Family 1 products. Thus, applying the same methodology undertaken in this

thesis to the shipment data of products from other families and the interplant shipment data might reveal more findings. Finally, one last possible approach could be discretizing the data in a way other than statistical weight might reveal new findings.

BIBLIOGRAPHY

Wu, J. (2002, February 1). *The Value in Mining Data.* Retrieved January 18, 2011, from Information Management Online: http://www.information-management.com/news/4618-1.html

Vertical Alliance Group, Inc. (2011). *Factors that Affect Drivers Shortage*. Retrieved April 03, 2011, from CSA2010.COM: http://www.csa2010.com/articles/Future_Driver_Shortage.htm

Coyle, J. J., Novack, R., Gibson, B., & Bardi, E. J. (2011). *Transportation: A Supply Chain Perspective* (7th Edition ed.). Mason: South-Western Cengage Learning.

Fei, Z., Zhang, J., & Zhou, X. (2010). Research on the Application of Data Mining in Logistics Enterprise. *ICLEM 2010: Logistics for Sustained Economic Development* (pp. 2292-2296). American Society of Civil Engineers.

Hipp, J., Guntzer, U., & Nakhaeizadeh, G. (2000). Algorithms for Association Rule Mining - A General Survey and Comparisons. (pp. 2(2): 1 - 58). SIGKDD.

Mathew, N. (2005, May). Use of Data Mining to Improve Supply Chain Operational Execution. *Graduate Thesis* .

McKinsey & Company. (2010). *Consumers Packaged Goods - Executive Insight*. Retrieved December 4, 2010, from McKinsey & Company: http://www.mckinsey.com/clientservice/consumerpackagedgoods/insight.asp

Solomon, M. B. (2010, June 04). *Worst-Ever Driver Shortage Looming, Trucking Executives Warn*. Retrieved April 03, 2011, from DC Velocity: http://www.dcvelocity.com/articles/20100623_worst_driver_shortage_looming/

Rodrigue, J.-P., Slack, B., & Comtois, C. (2009). *Transport Safety and Security*. New York: Routledge.

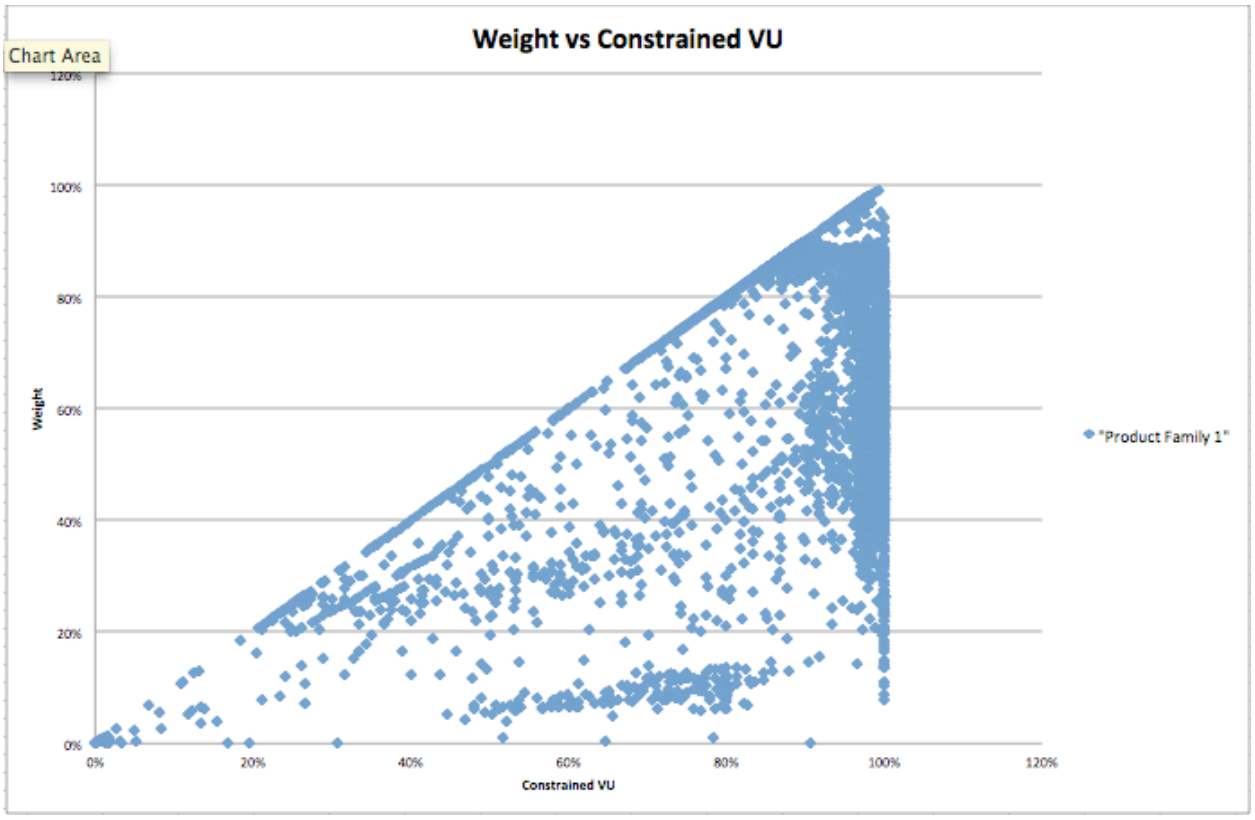
TechTarget. (2010). *Consumer Packaged Goods Industry White Papers*. Retrieved December 4, 2010, from The TechTarget Library of White Papers, Product Literature, Webcasts and Case Studies: http://www.bitpipe.com/tlist/Consumer-Packaged-Goods-Industry.html

Tseng, Y.-y., Yue, W., & Taylor, M. (2005). The Role of Transportation in Logistics Chain. *The Eastern Asia Society for Transportation Studies*, *5*, pp. 1657-1672.

APPENDIX A

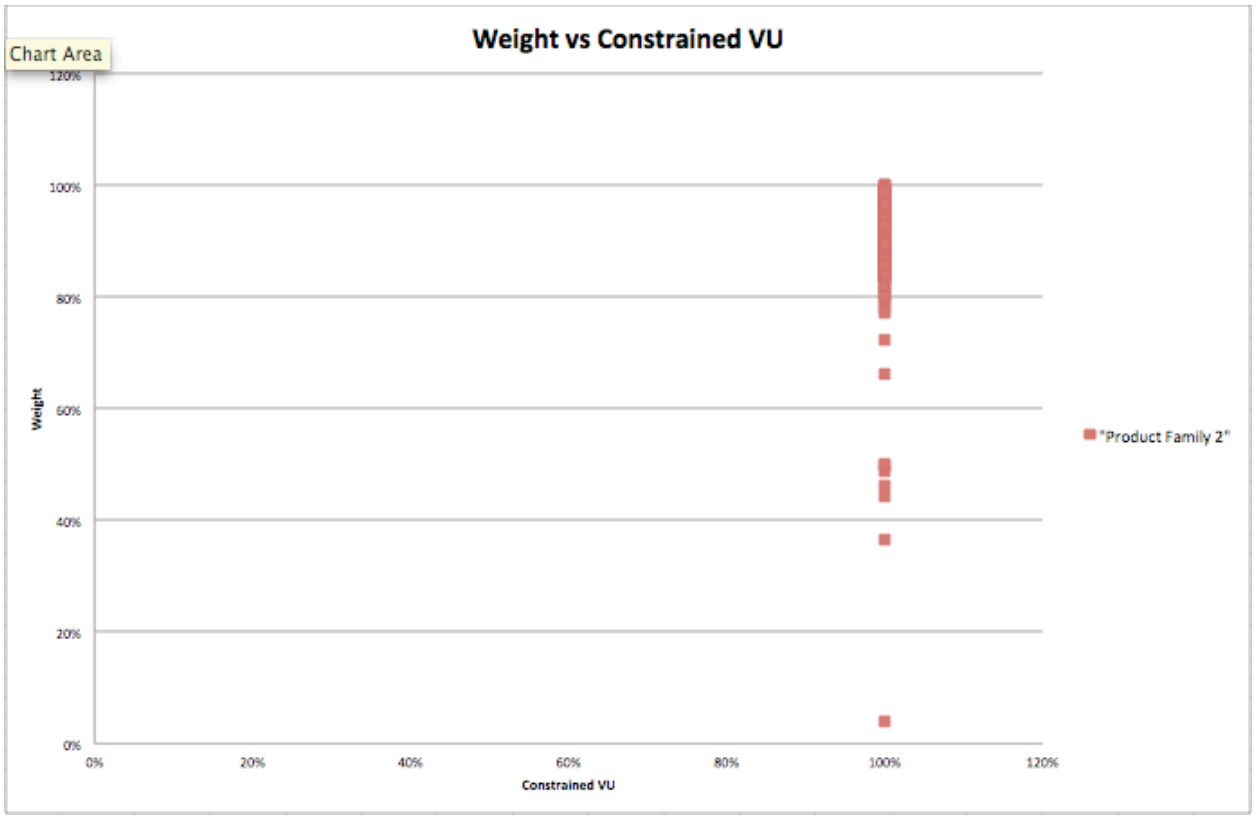
Appendix A1

Weight vs Constrained VU – Product Family 1



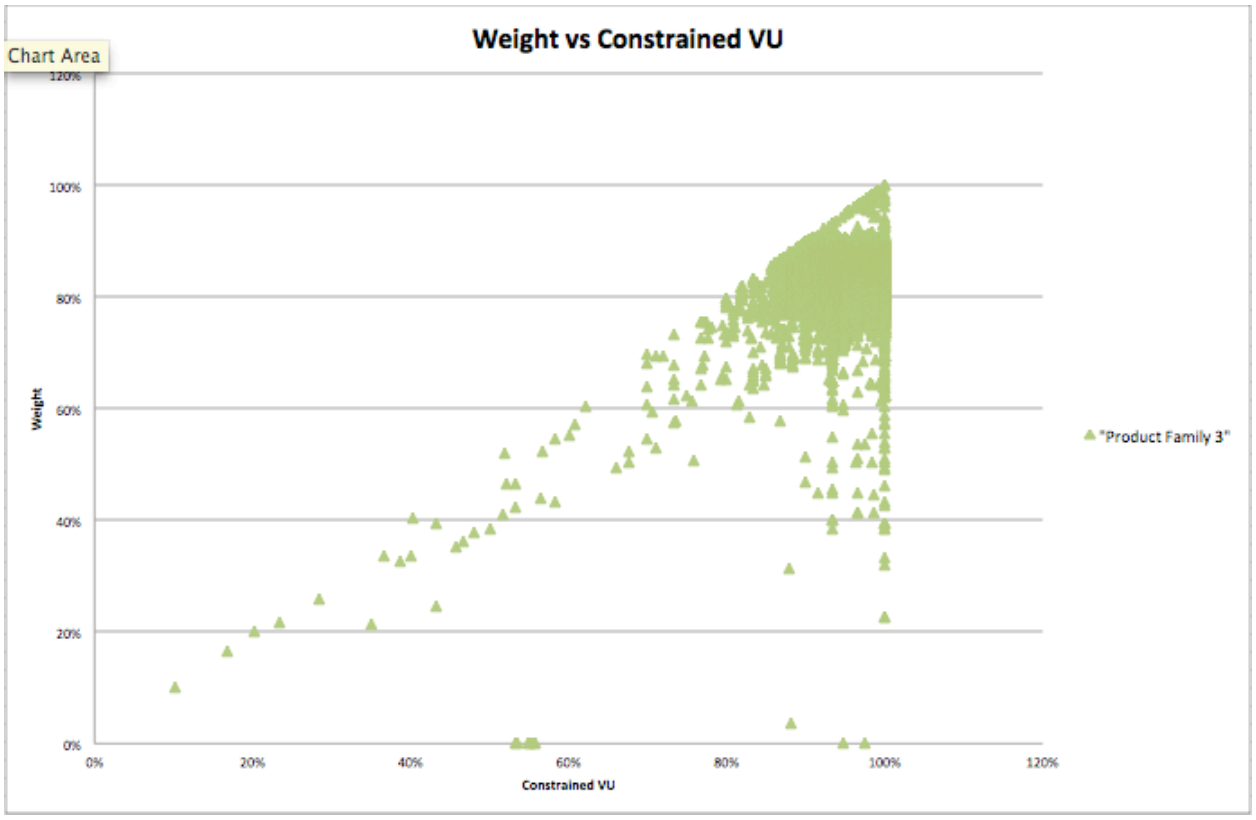
Appendix A2

Weight vs Constrained VU – Product Family 2



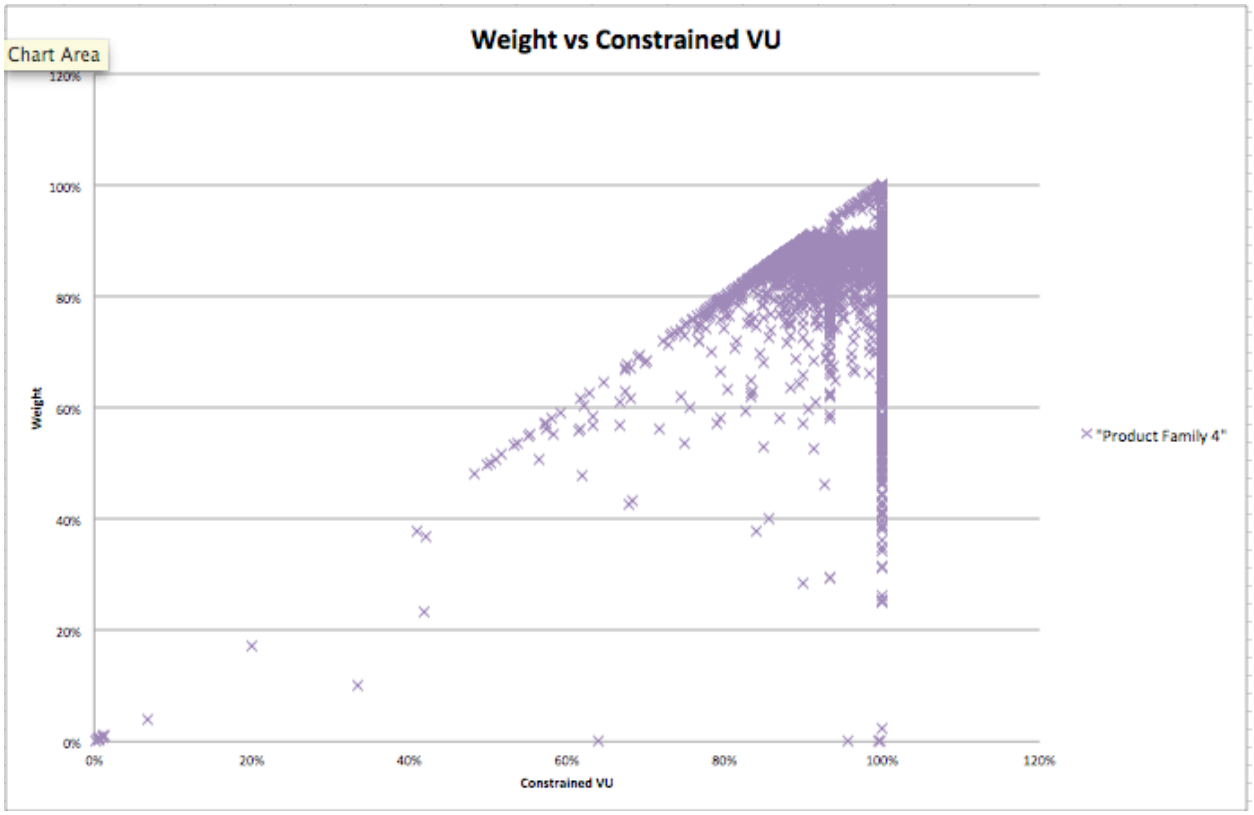
Appendix A3

Weight vs Constrained VU – Product Family 3



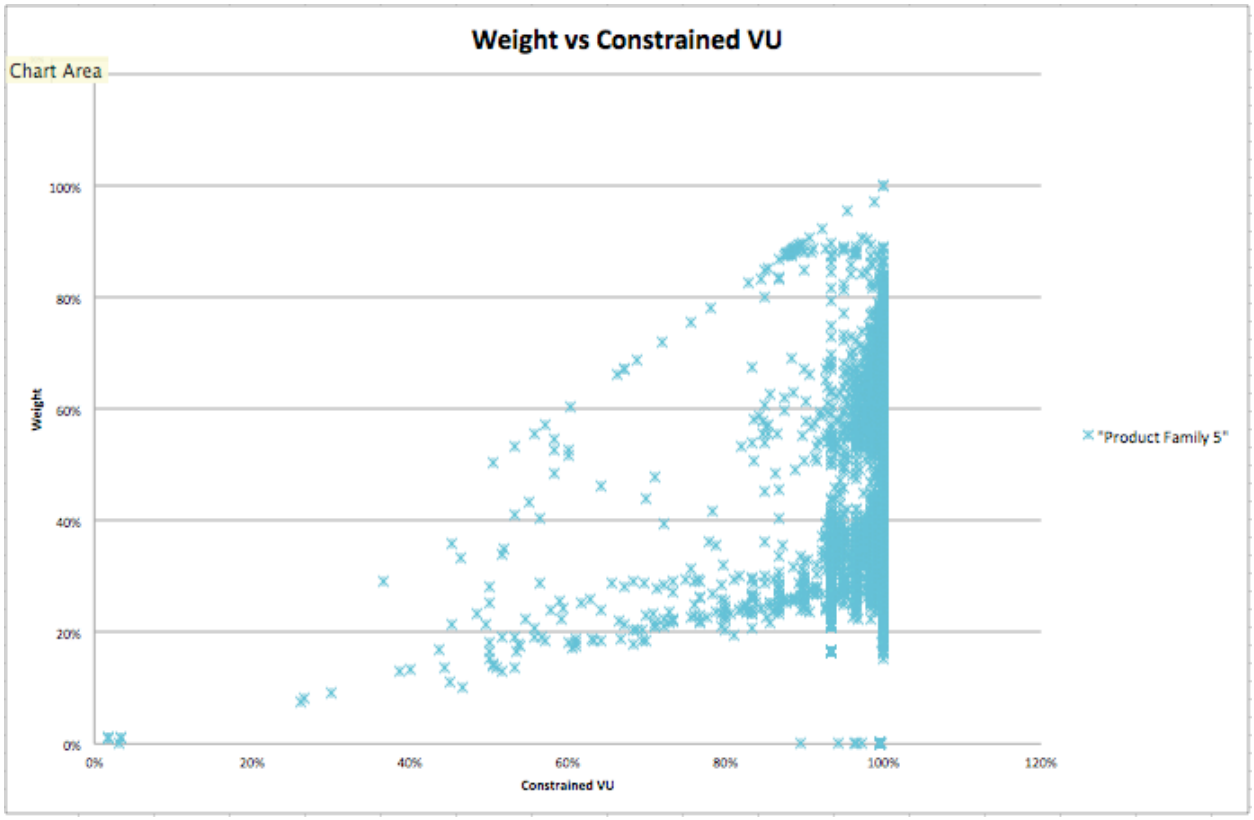
Appendix A4

Weight vs Constrained VU – Product Family 4



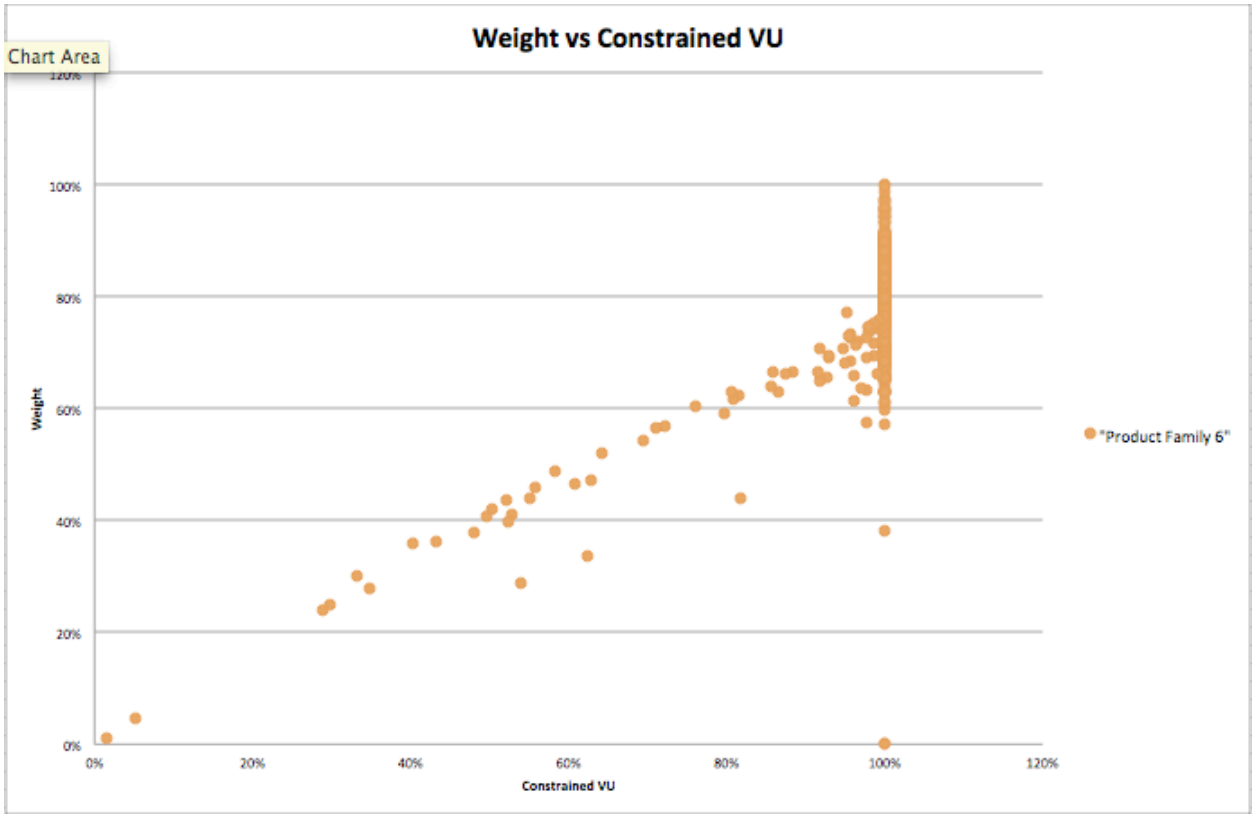
Appendix A5

Weight vs Constrained VU – Product Family 5



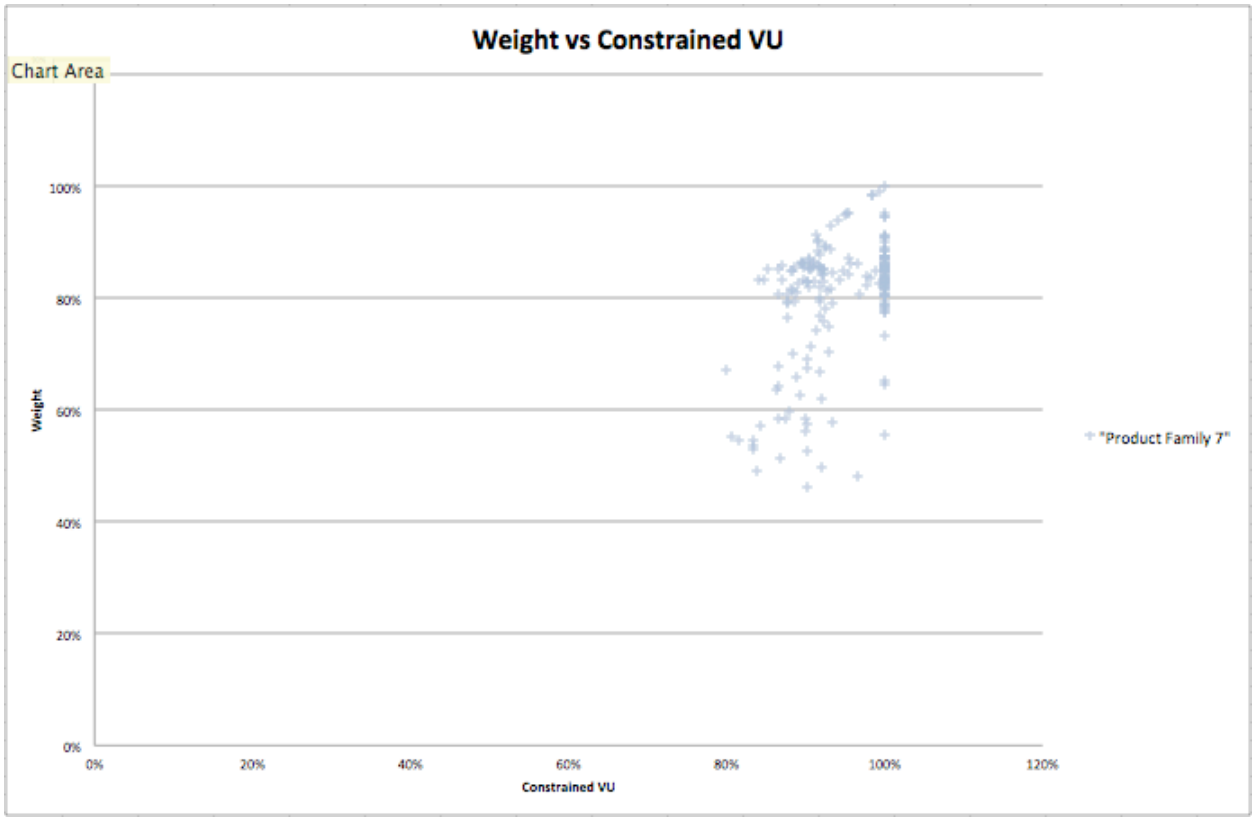
Appendix A6

Weight vs Constrained VU – Product Family 6



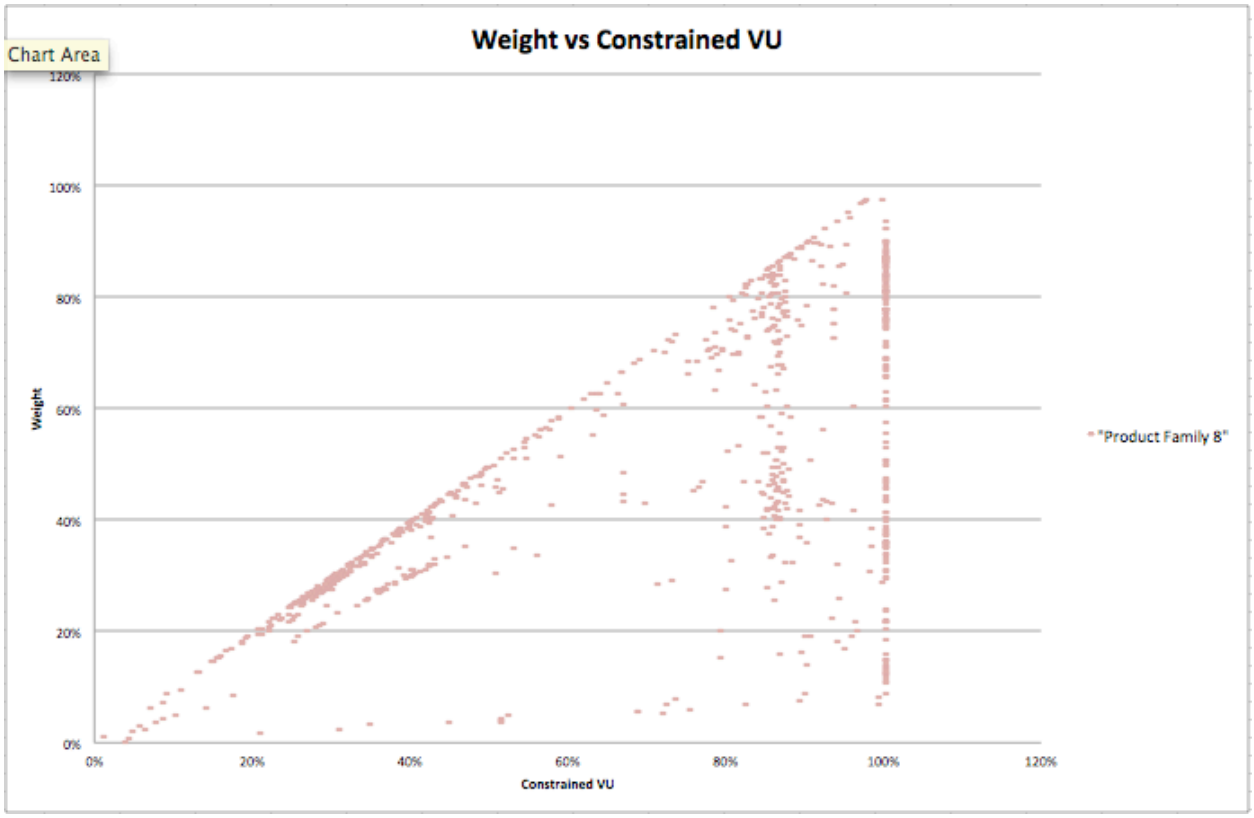
Appendix A7

Weight vs Constrained VU – Product Family 7



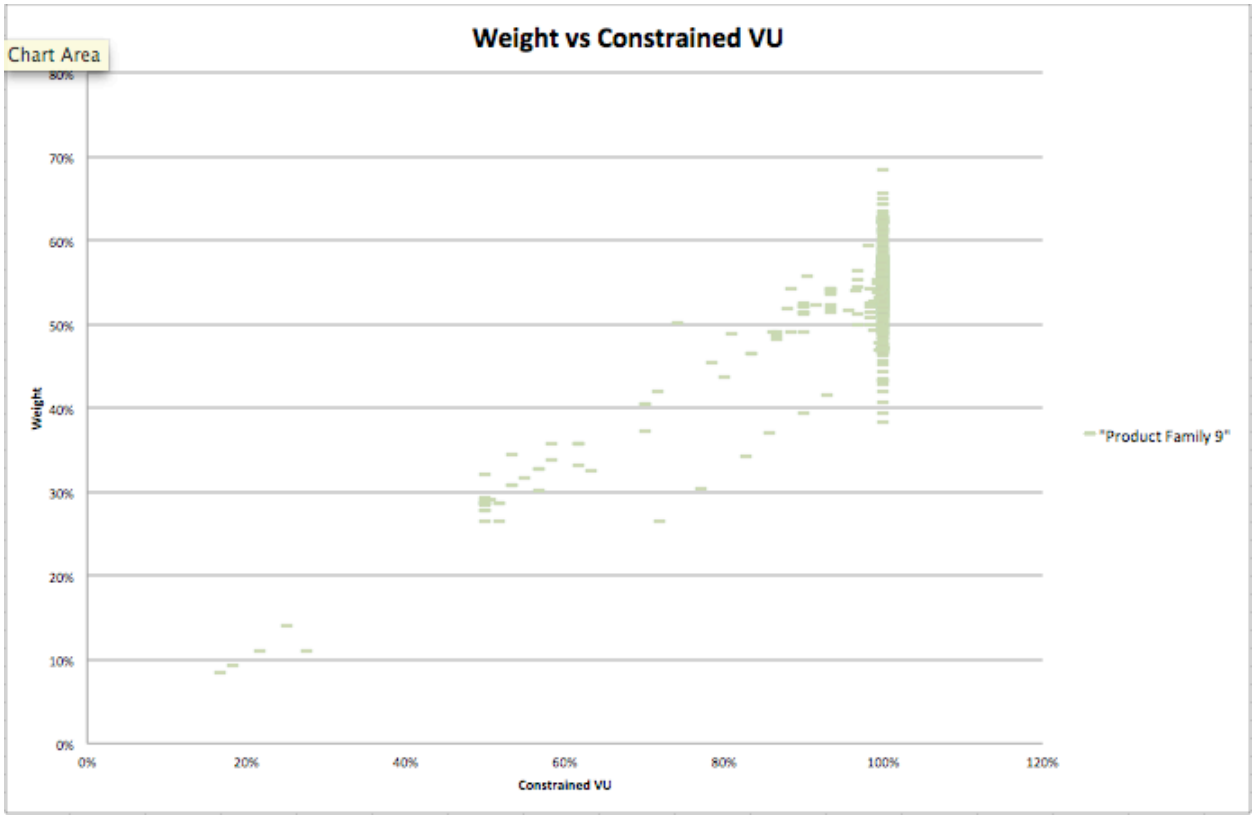
Appendix A8

Weight vs Constrained VU – Product Family 8



Appendix A9

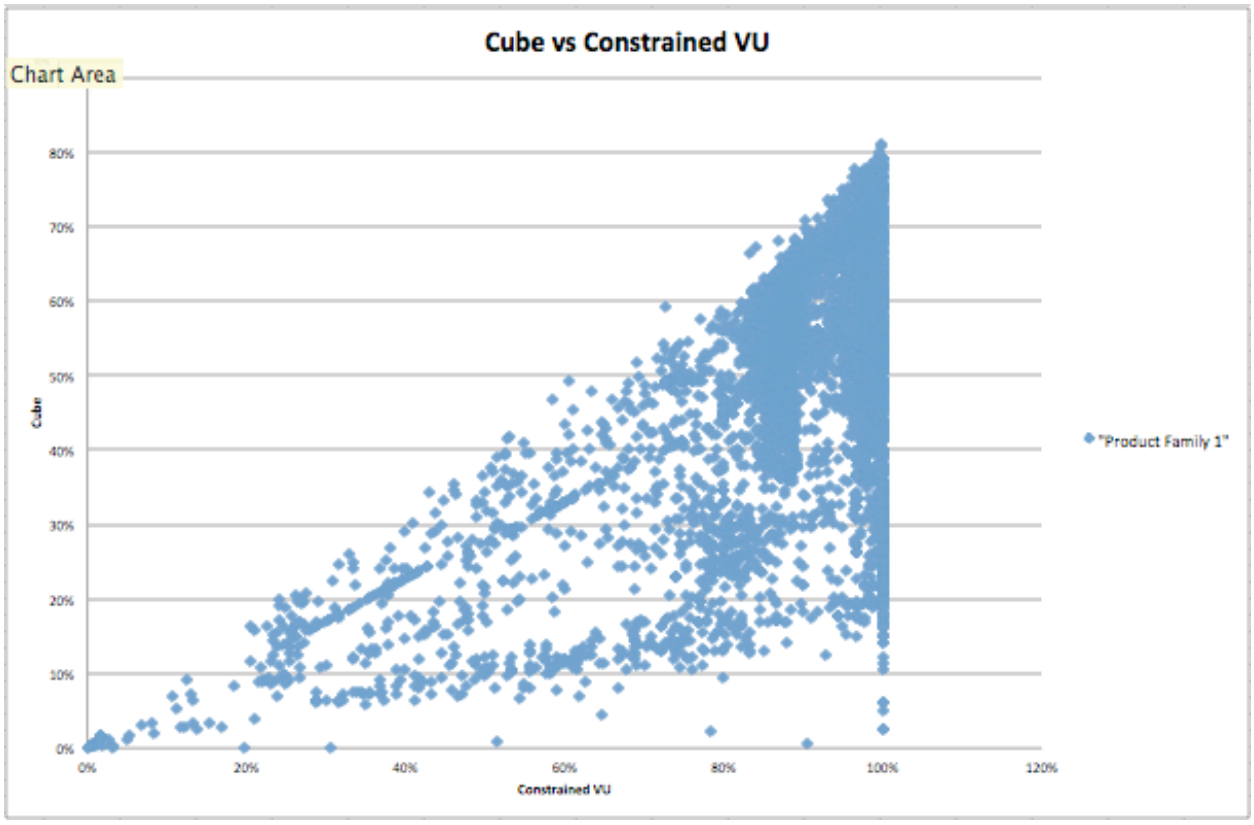
Weight vs Constrained VU – Product Family 9



APPENDIX B

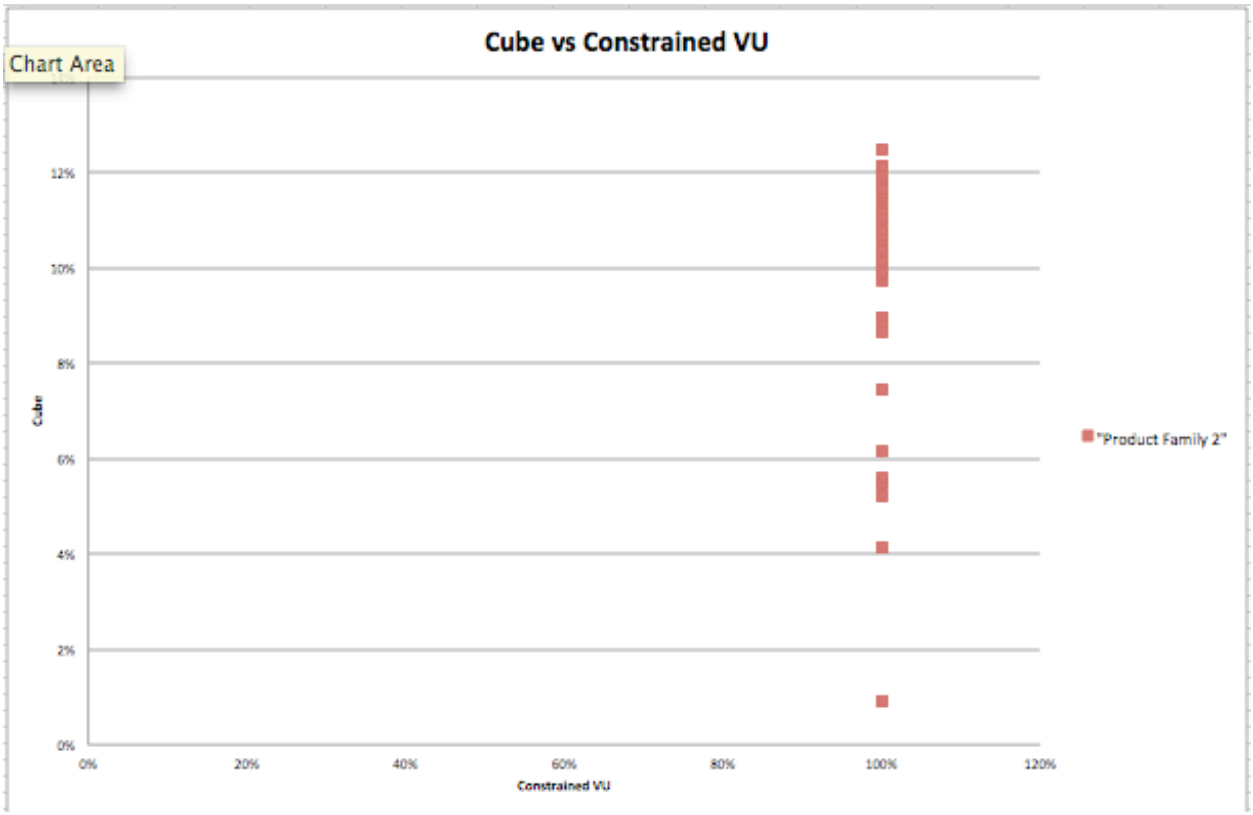
Appendix B1

Cube vs Constrained VU – Product Family 1



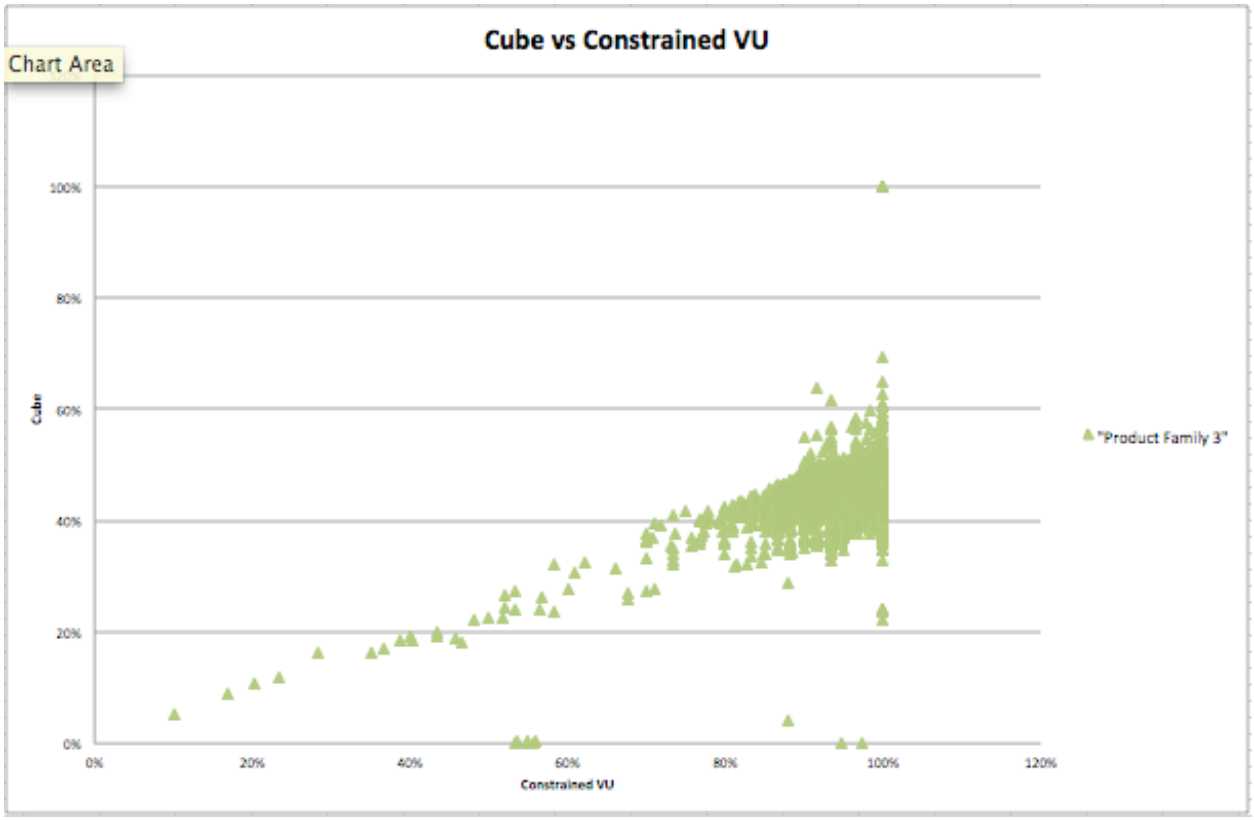
Appendix B2

Cube vs Constrained VU – Product Family 2



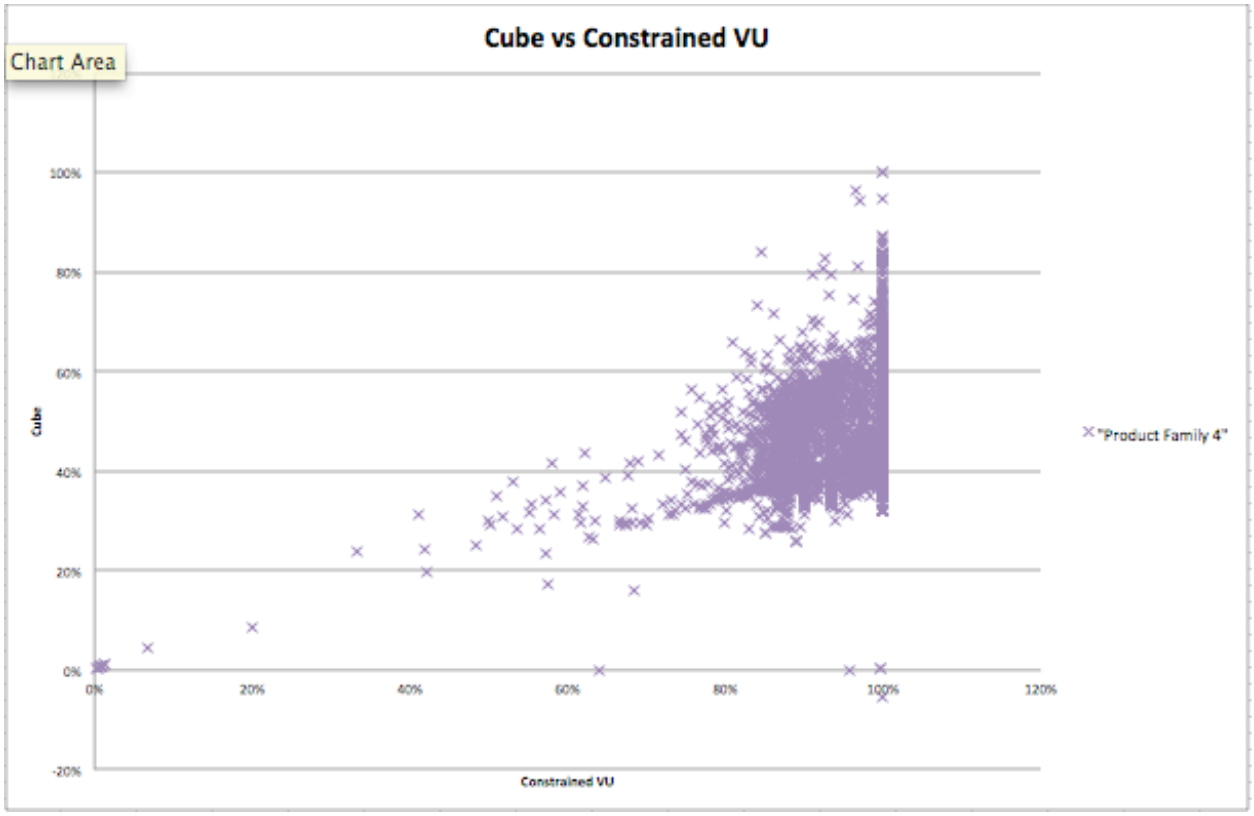
Appendix B3

Cube vs Constrained VU – Product Family 3



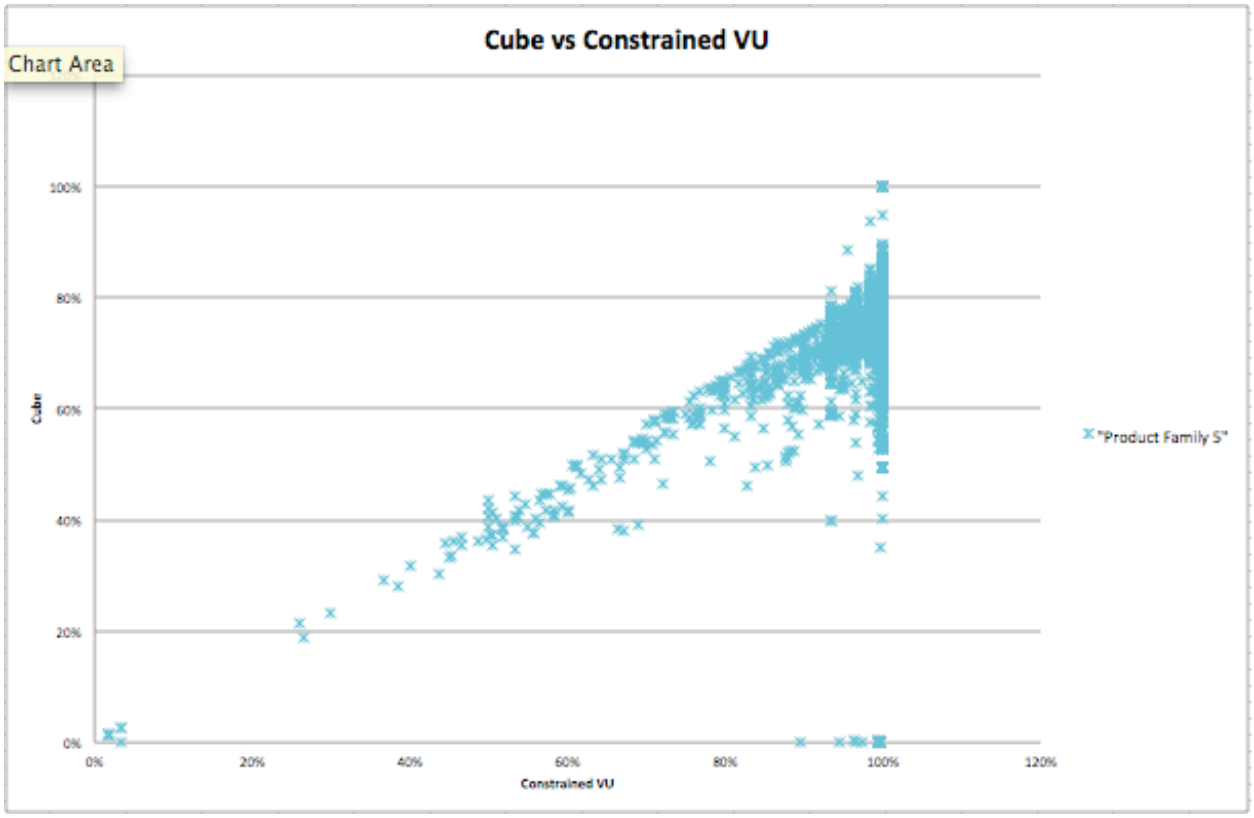
Appendix B4

Cube vs Constrained VU – Product Family 4



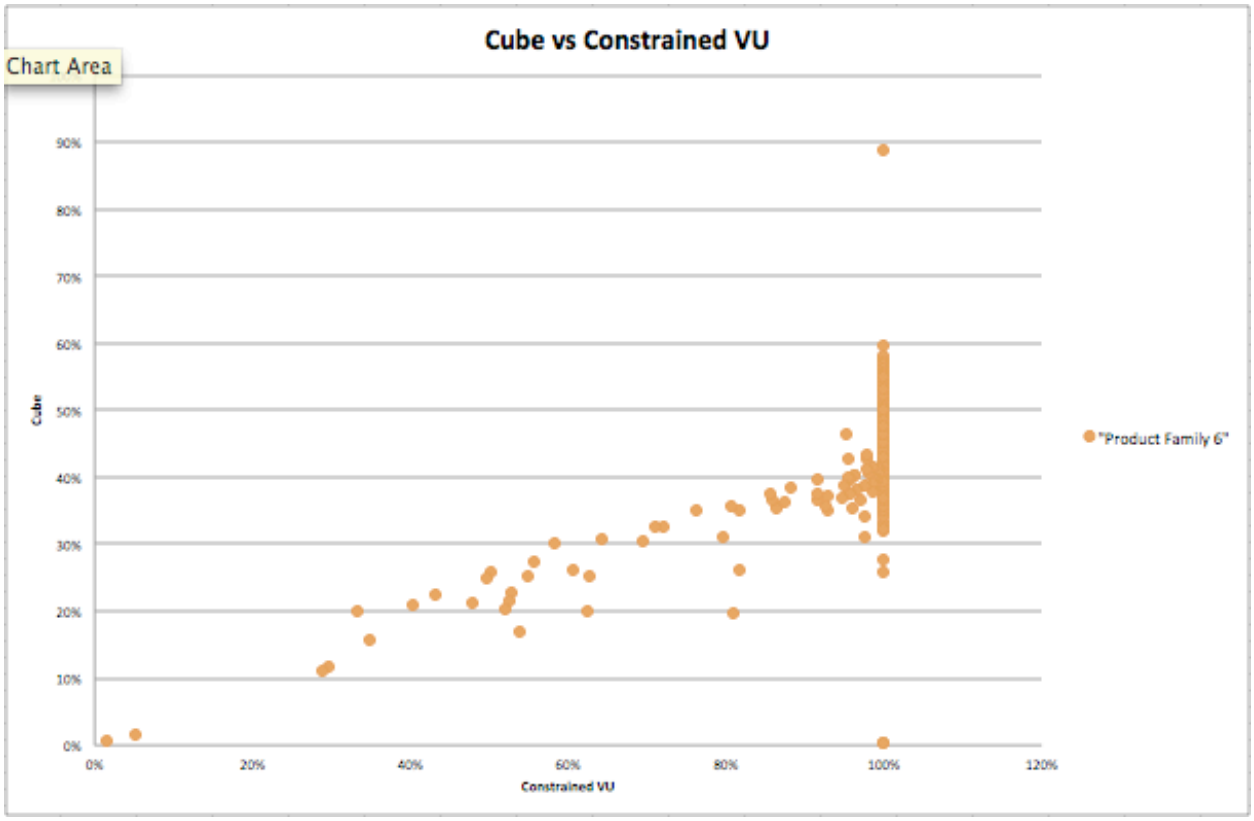
Appendix B5

Cube vs Constrained VU – Product Family 5



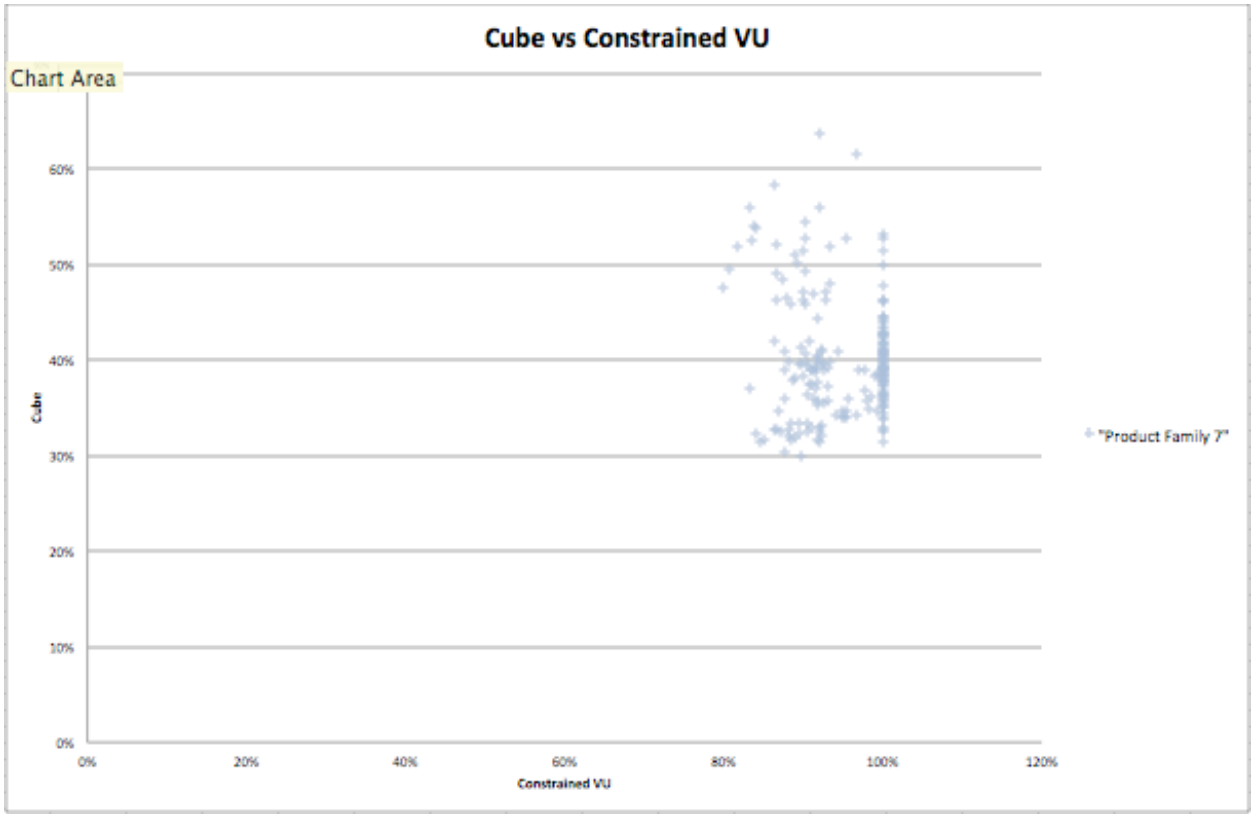
Appendix B6

Cube vs Constrained VU – Product Family 6



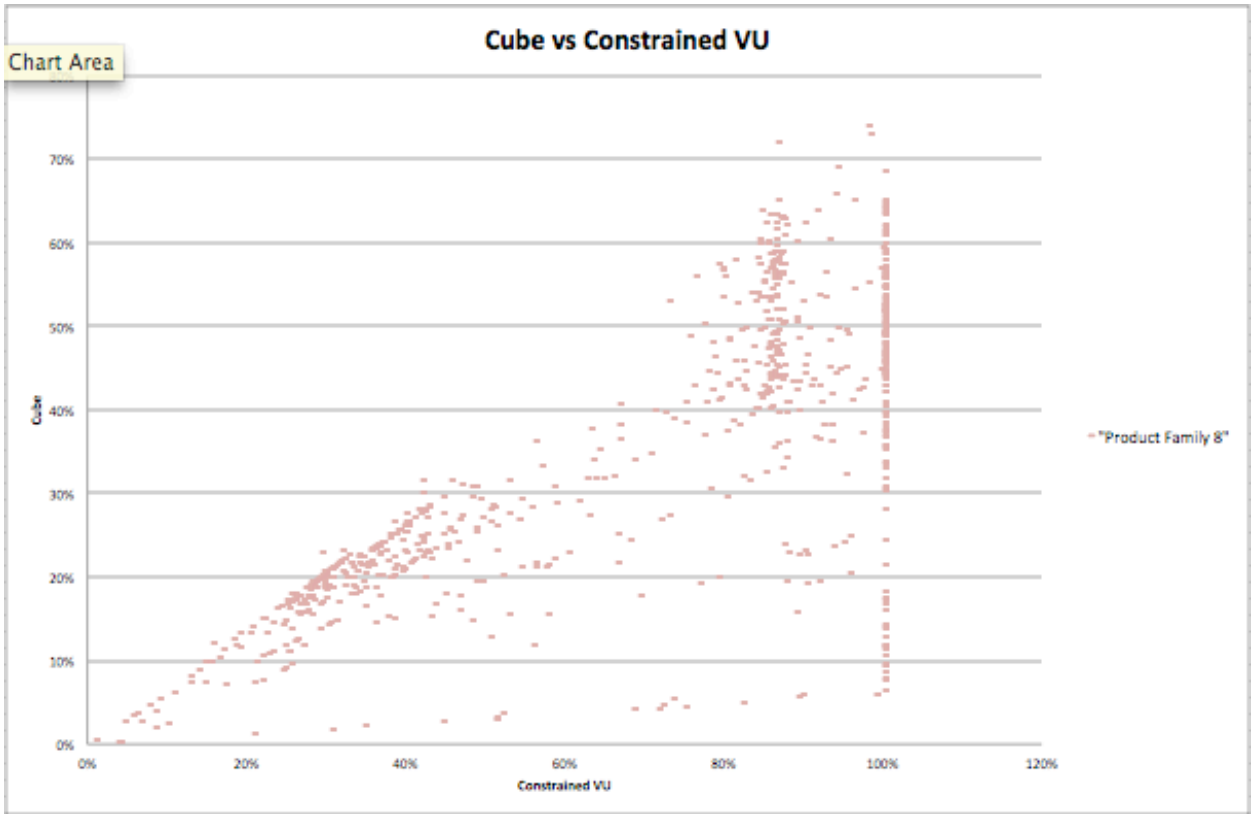
Appendix B7

Cube vs Constrained VU – Product Family 7



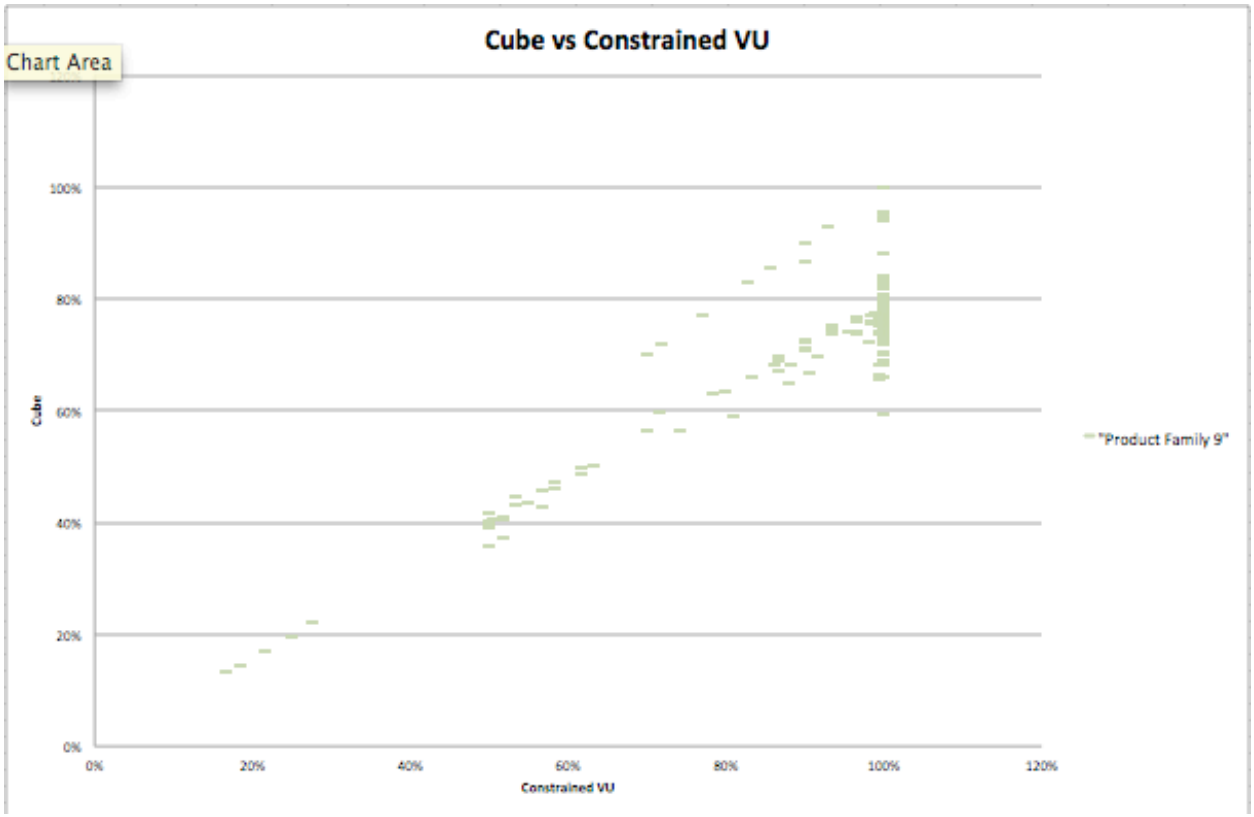
Appendix B8

Cube vs Constrained VU – Product Family 8



Appendix B9

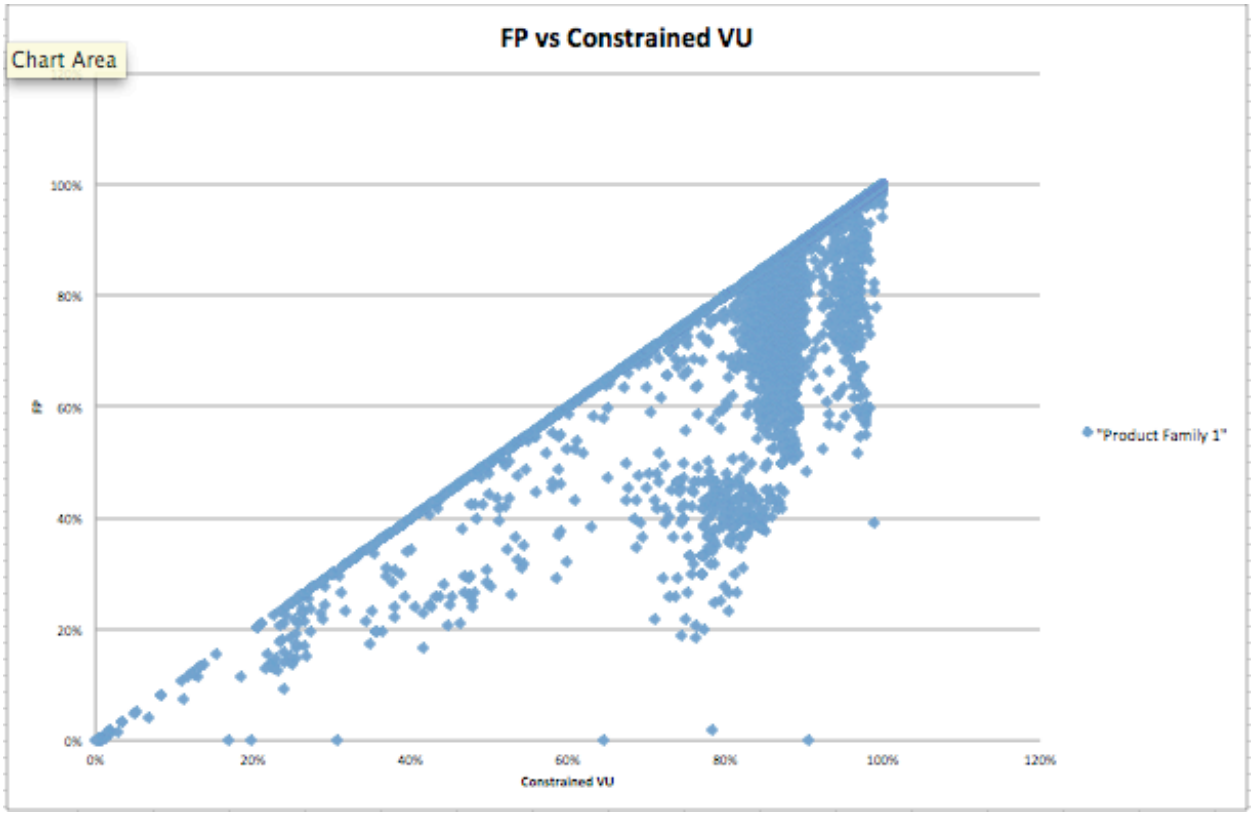
Cube vs Constrained VU – Product Family 9



APPENDIX C

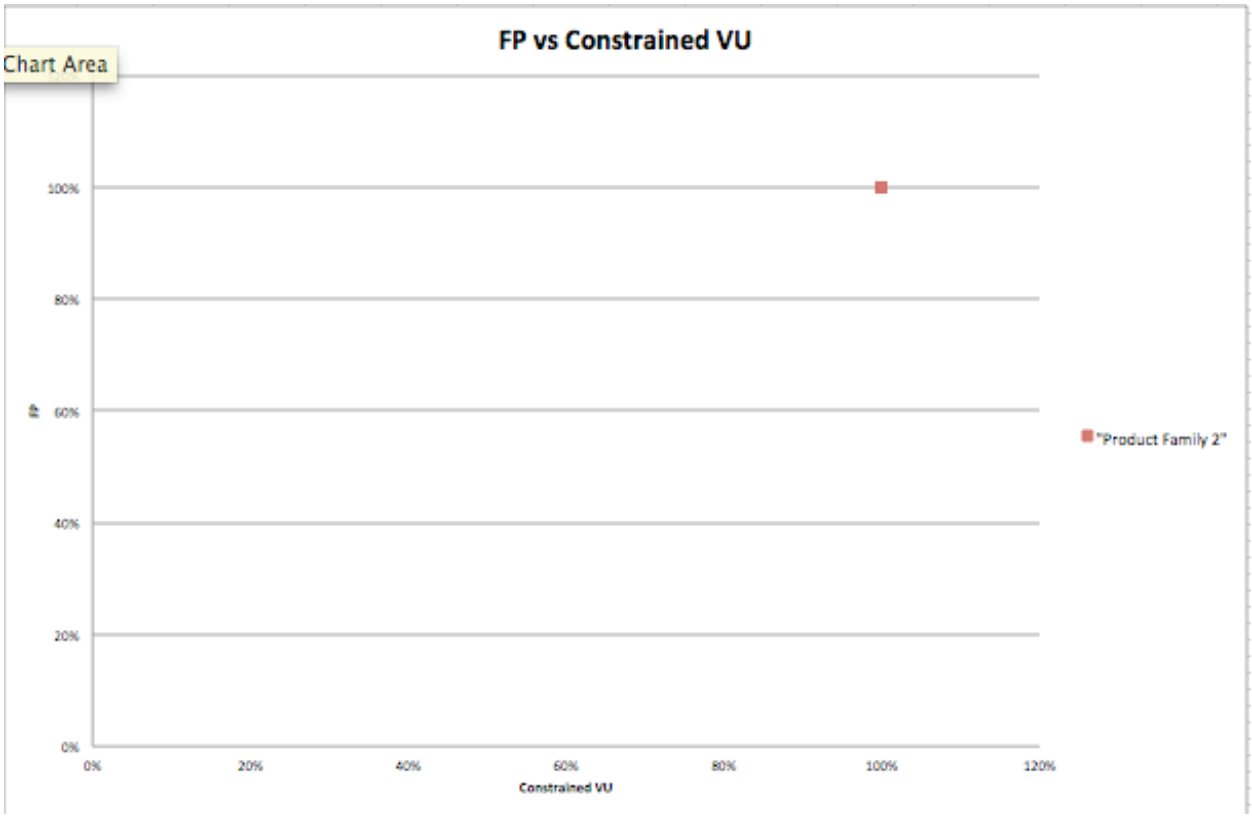
Appendix C1

FP vs Constrained VU – Product Family 1



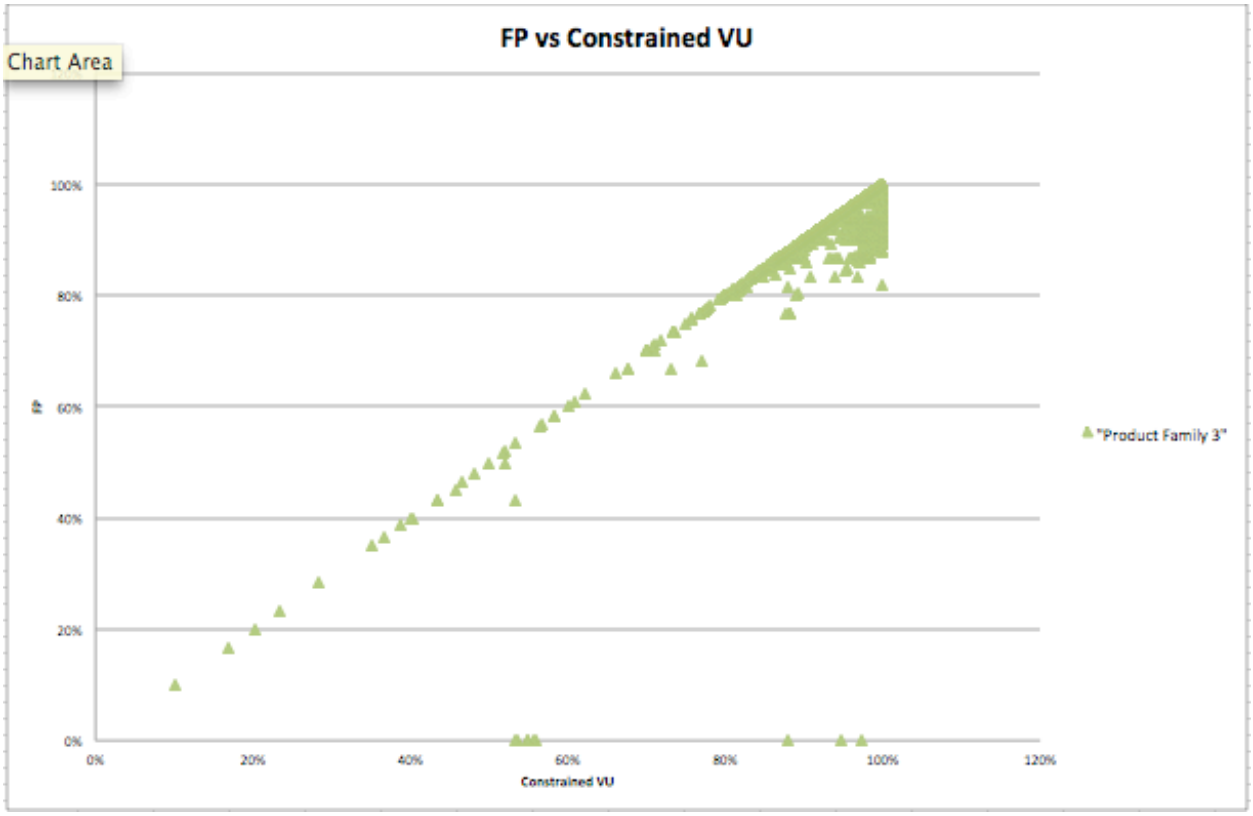
Appendix C2

FP vs Constrained VU – Product Family 2



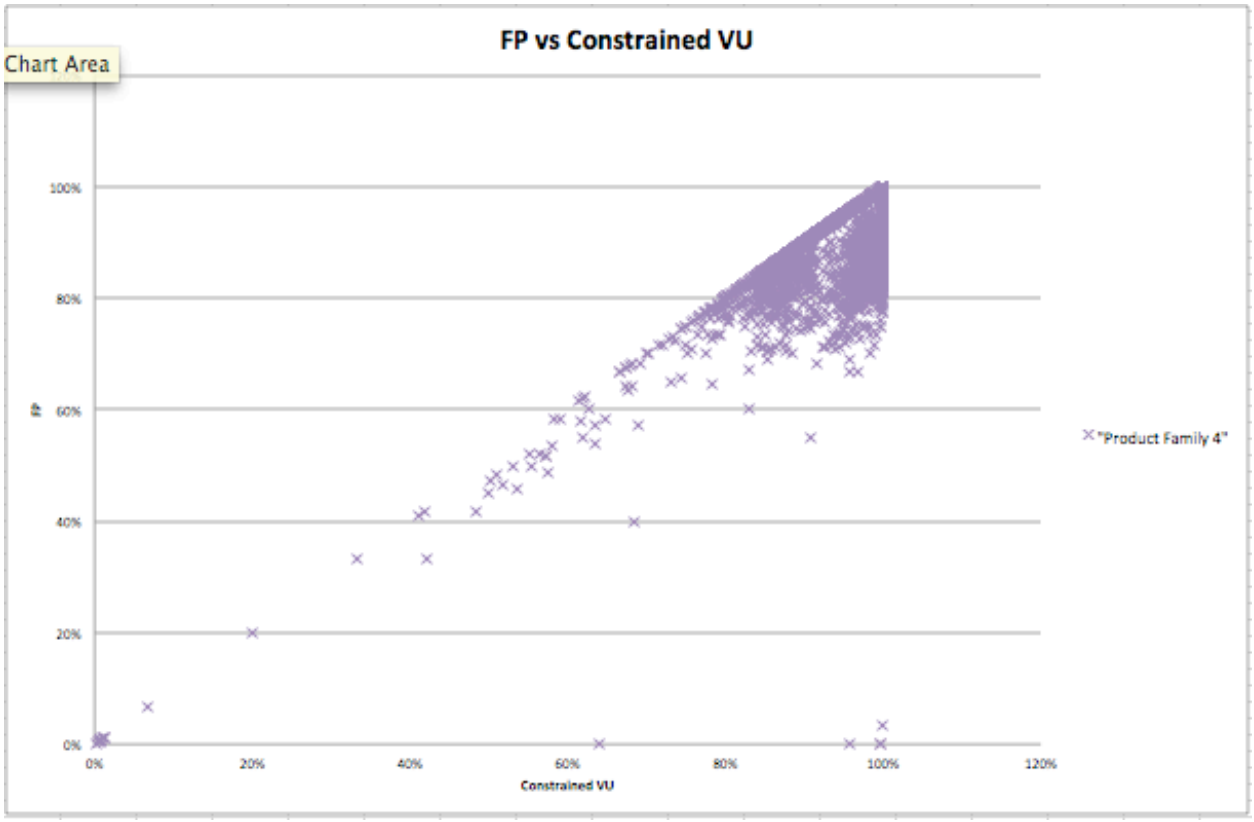
Appendix C3

FP vs Constrained VU – Product Family 3



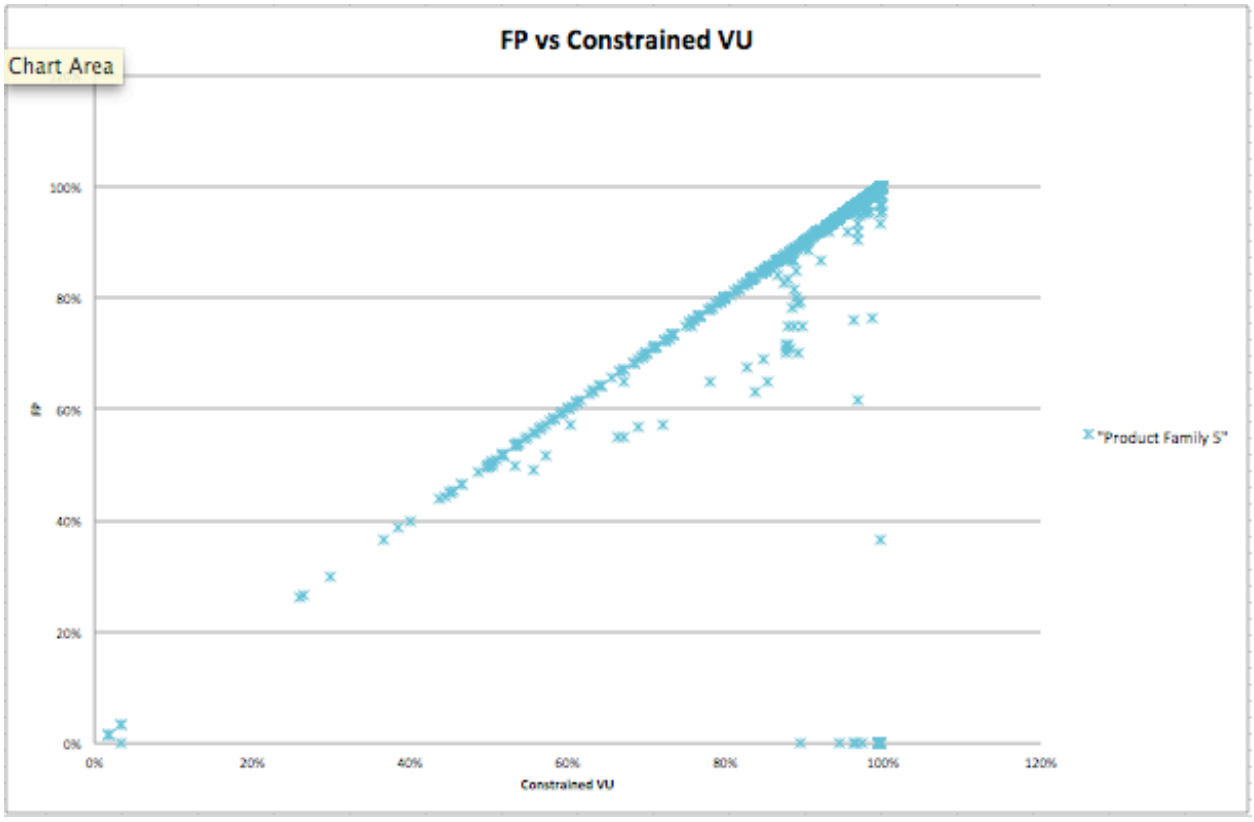
Appendix C4

FP vs Constrained VU – Product Family 4



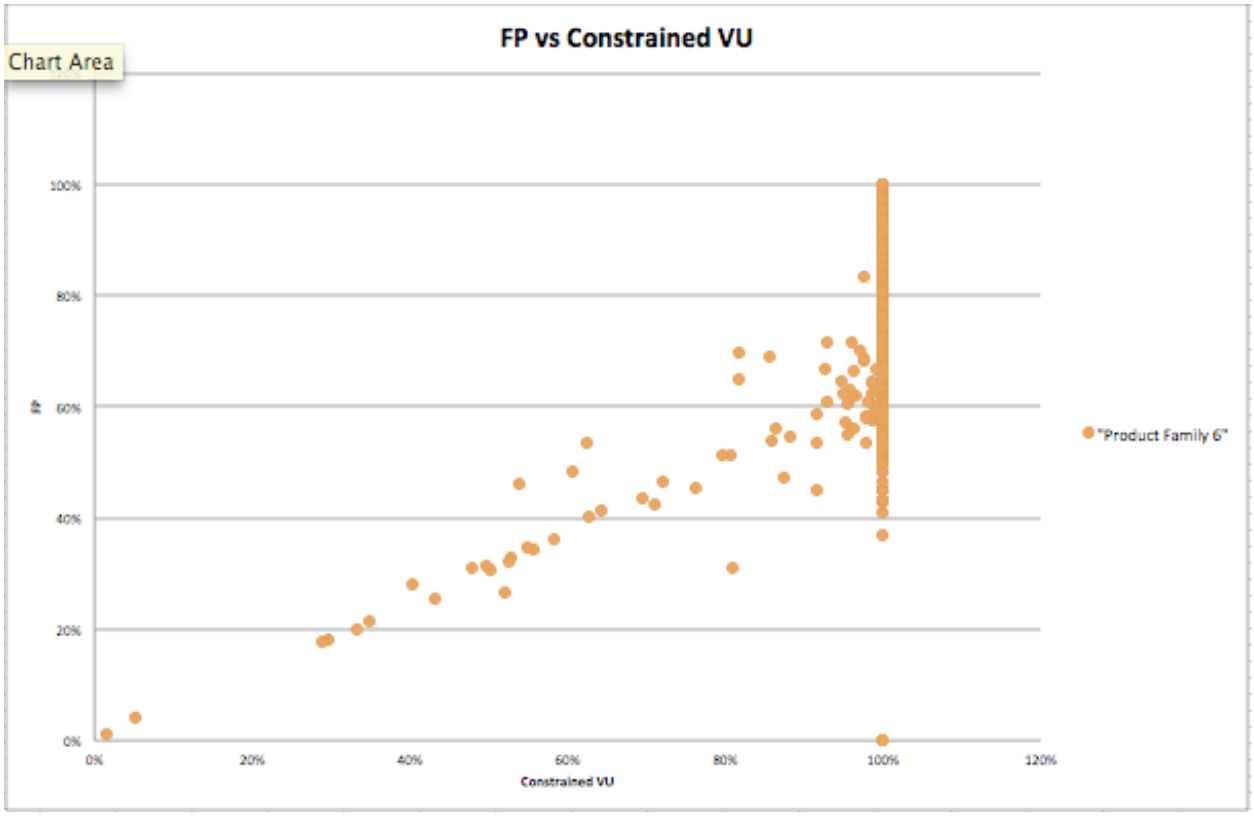
Appendix C5

FP vs Constrained VU – Product Family 5



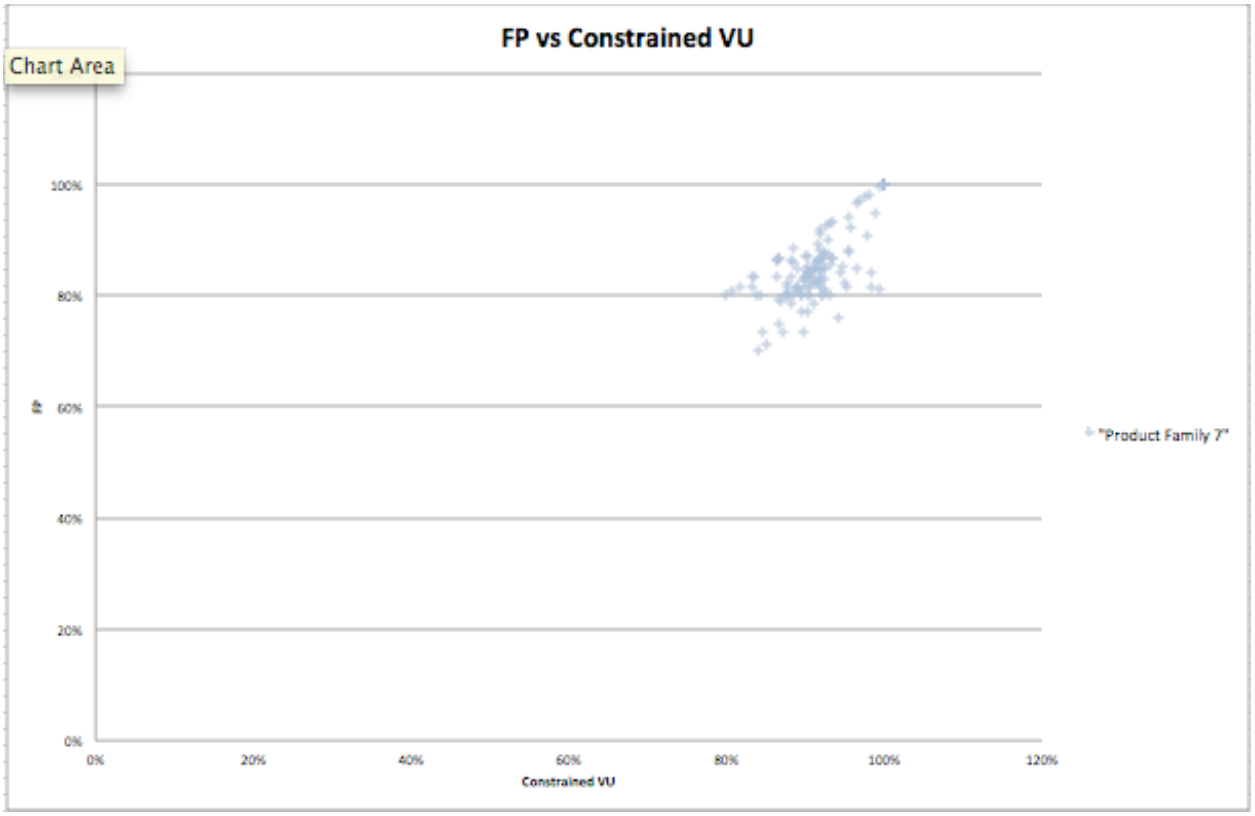
Appendix C6

FP vs Constrained VU – Product Family 6



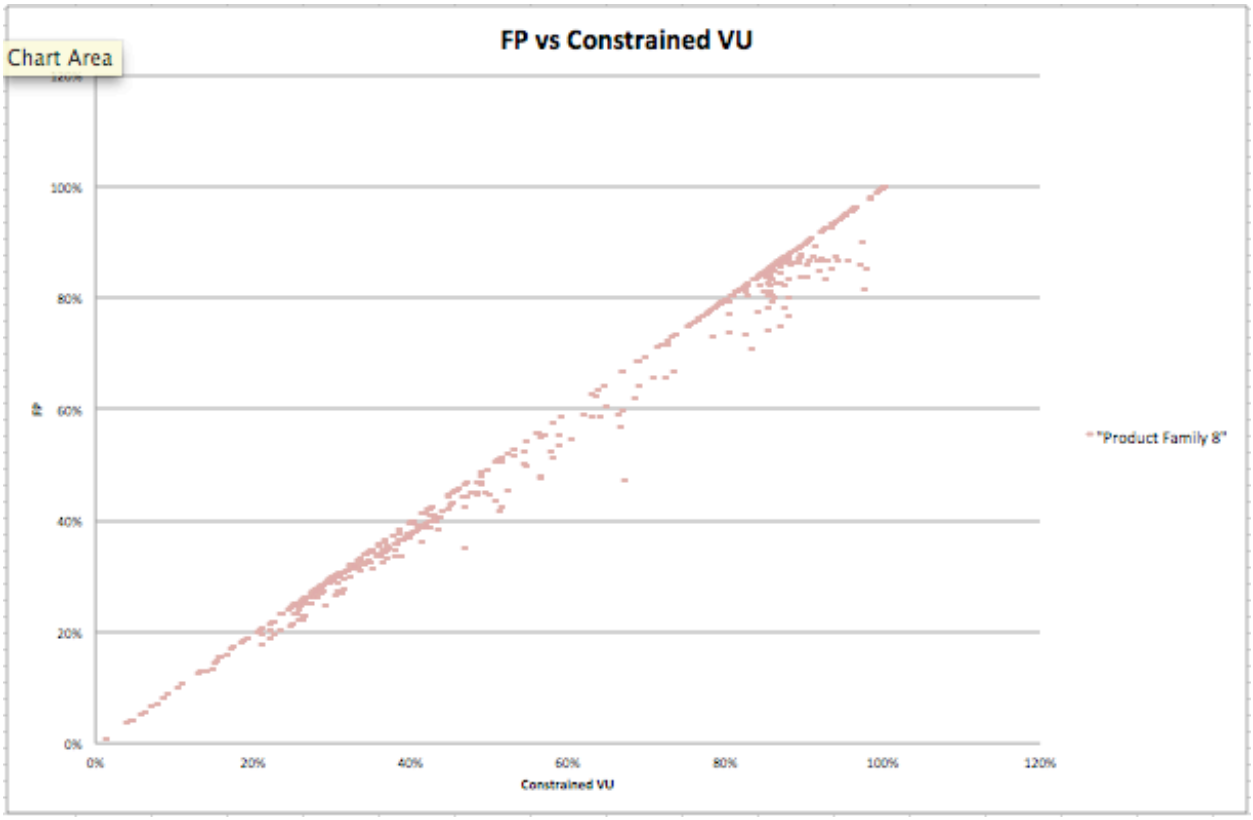
Appendix C7

FP vs Constrained VU – Product Family 7



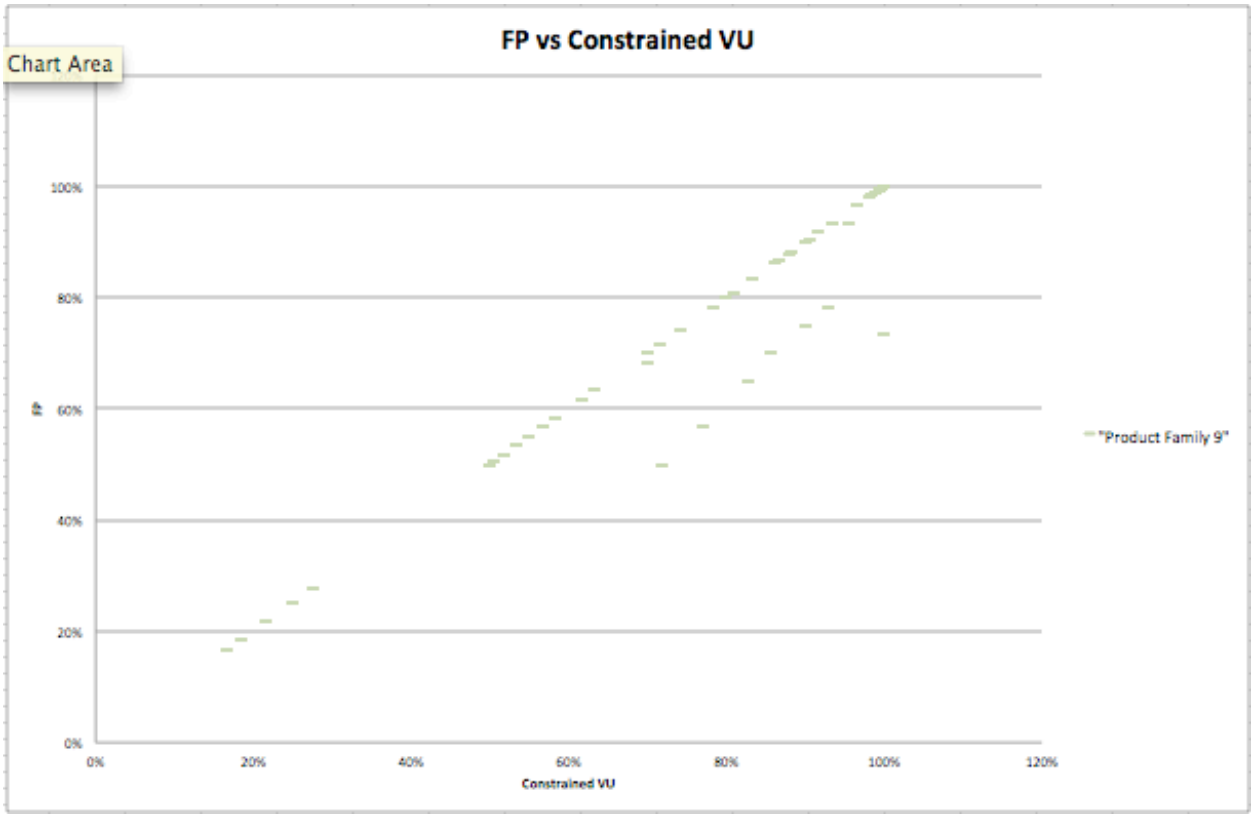
Appendix C8

FP vs Constrained VU – Product Family 8



Appendix C9

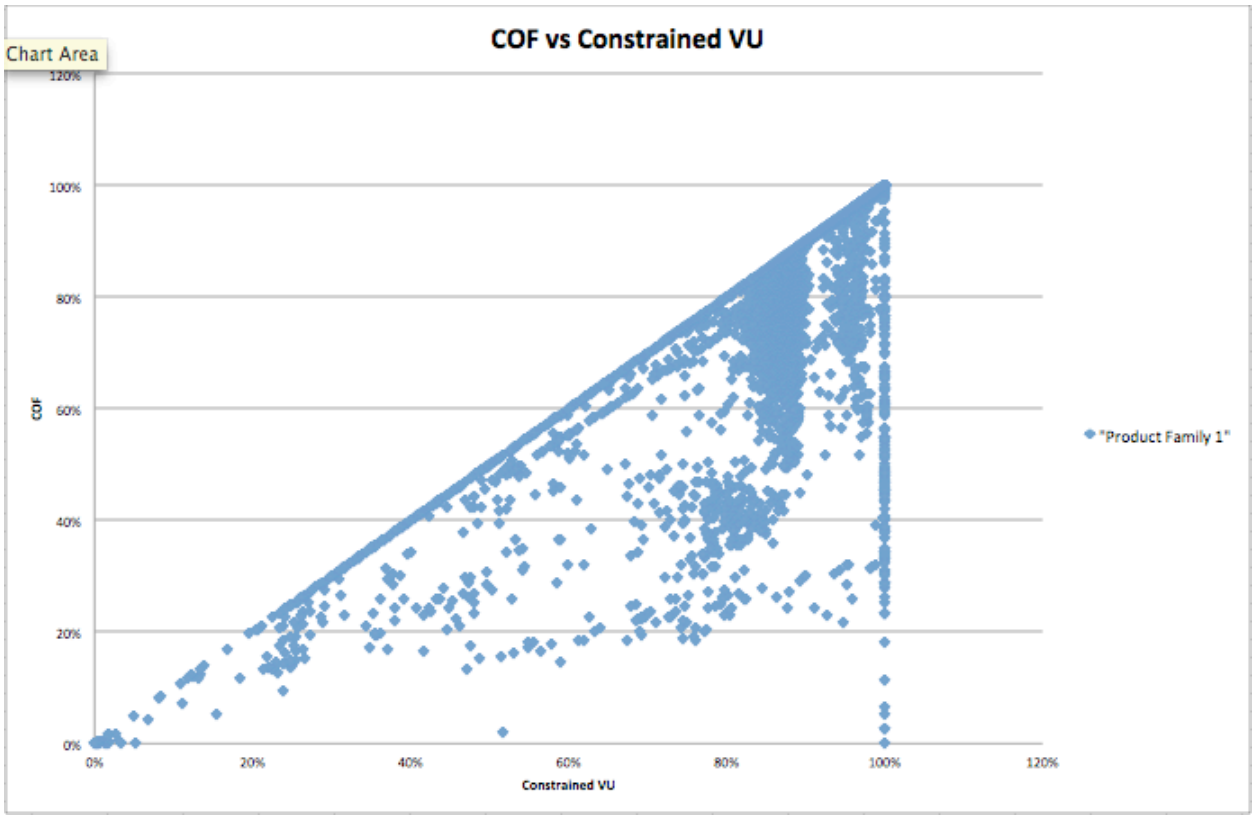
FP vs Constrained VU – Product Family 9



APPENDIX D

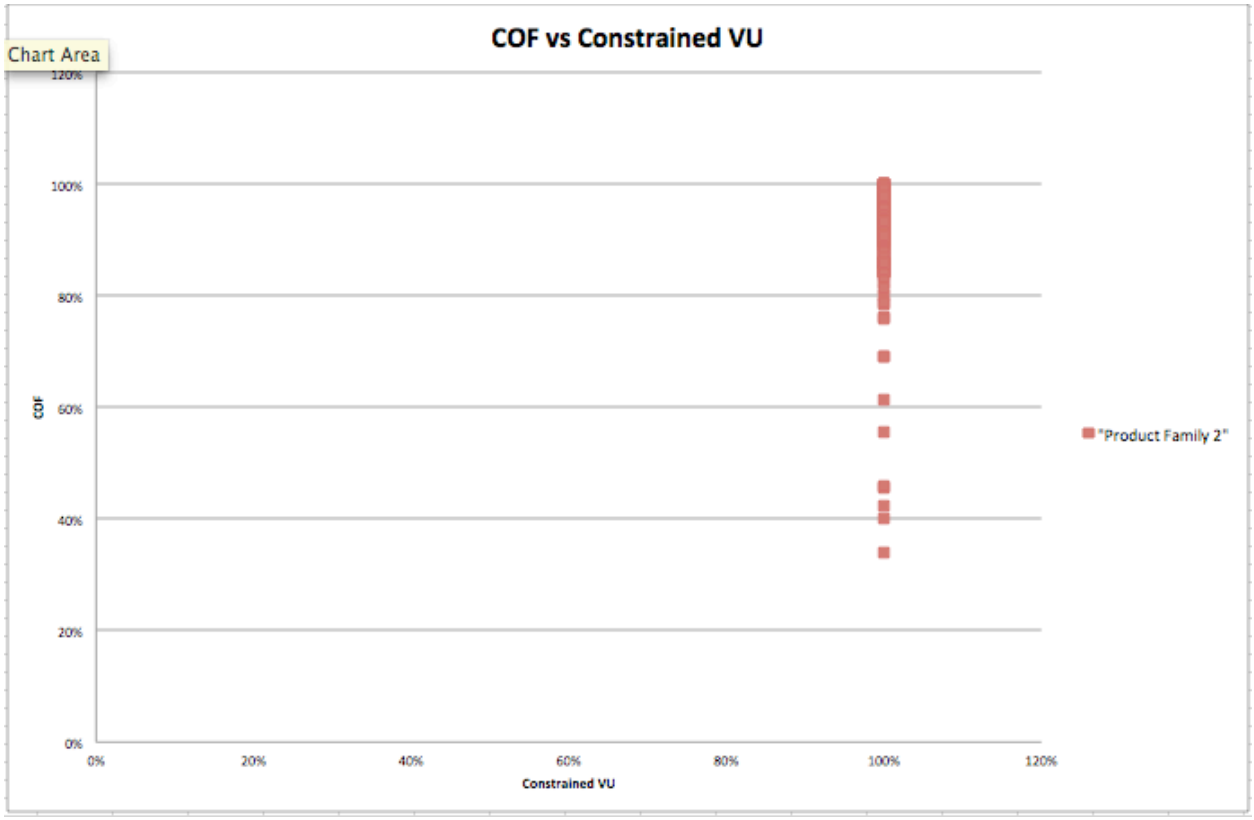
Appendix D1

COF vs Constrained VU – Product Family 1



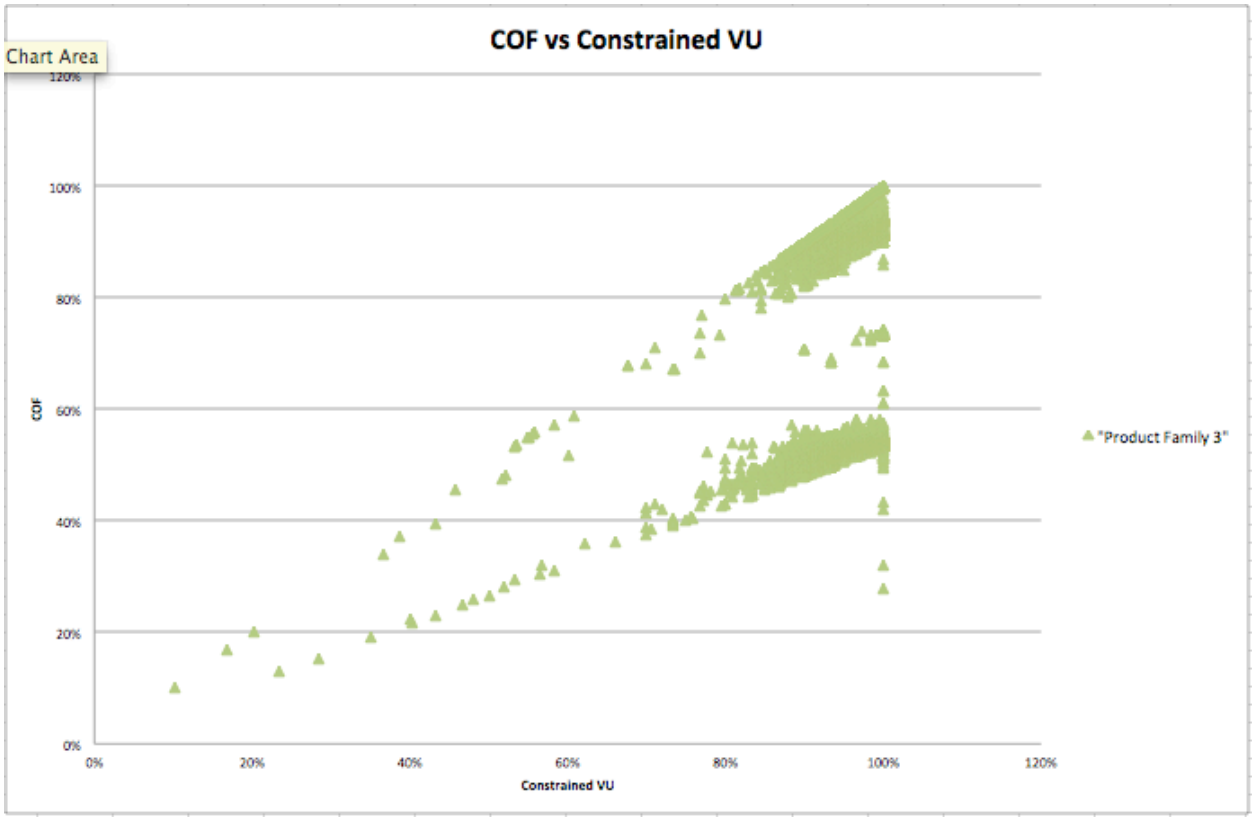
Appendix D2

COF vs Constrained VU – Product Family 2



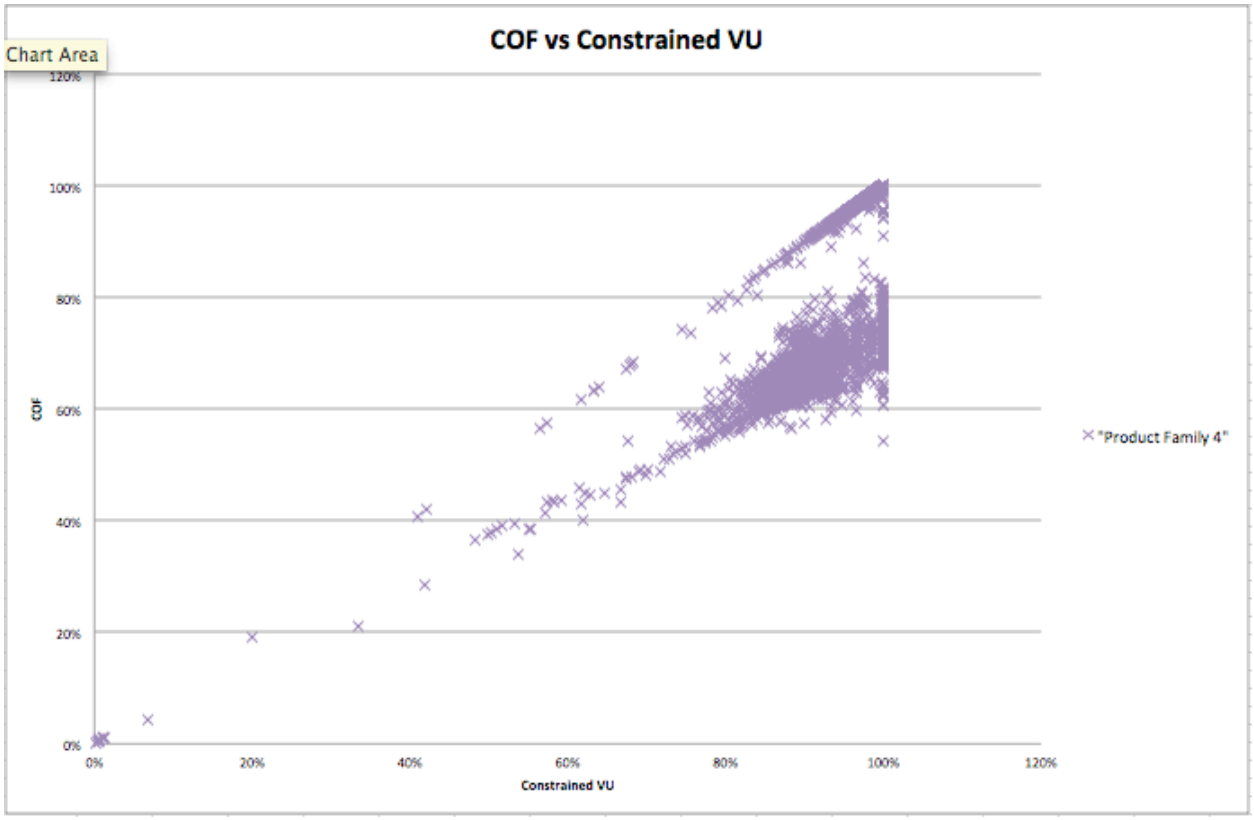
Appendix D3

COF vs Constrained VU – Product Family 3



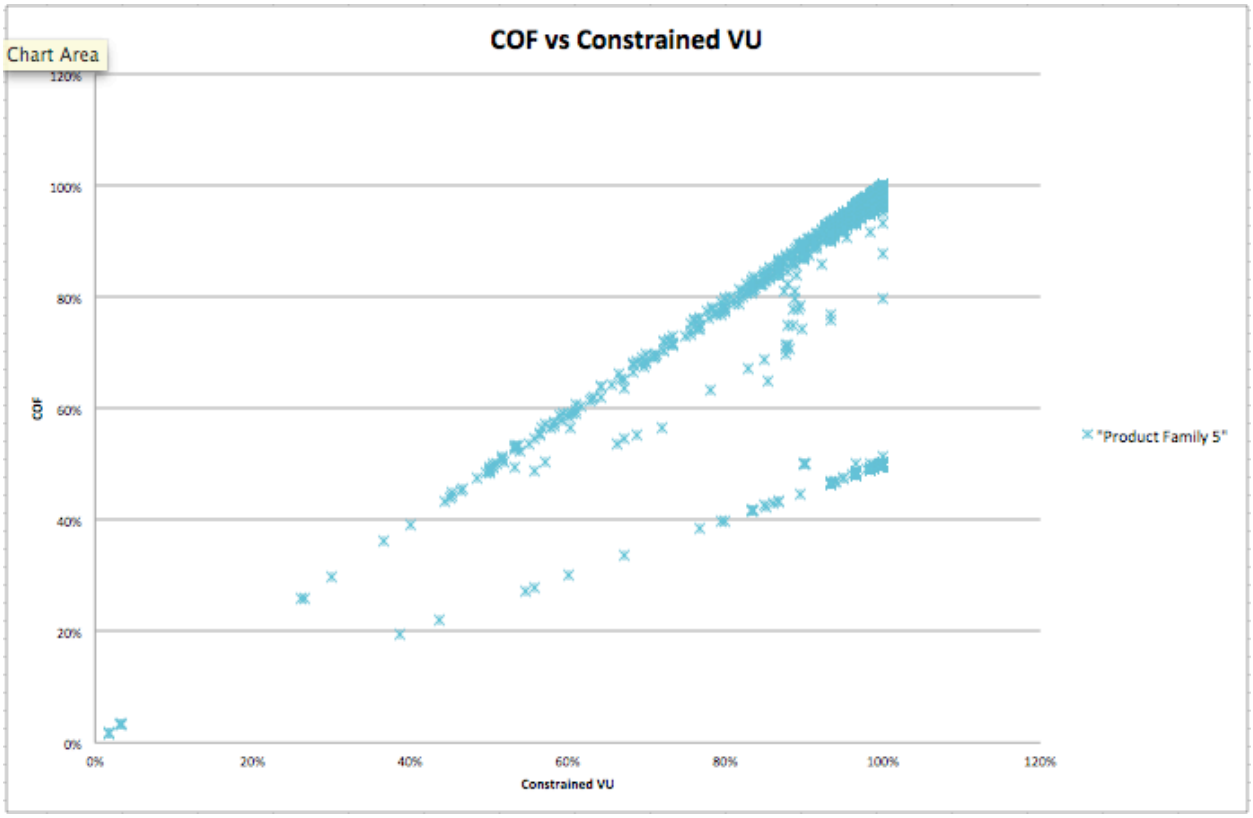
Appendix D4

COF vs Constrained VU – Product Family 4



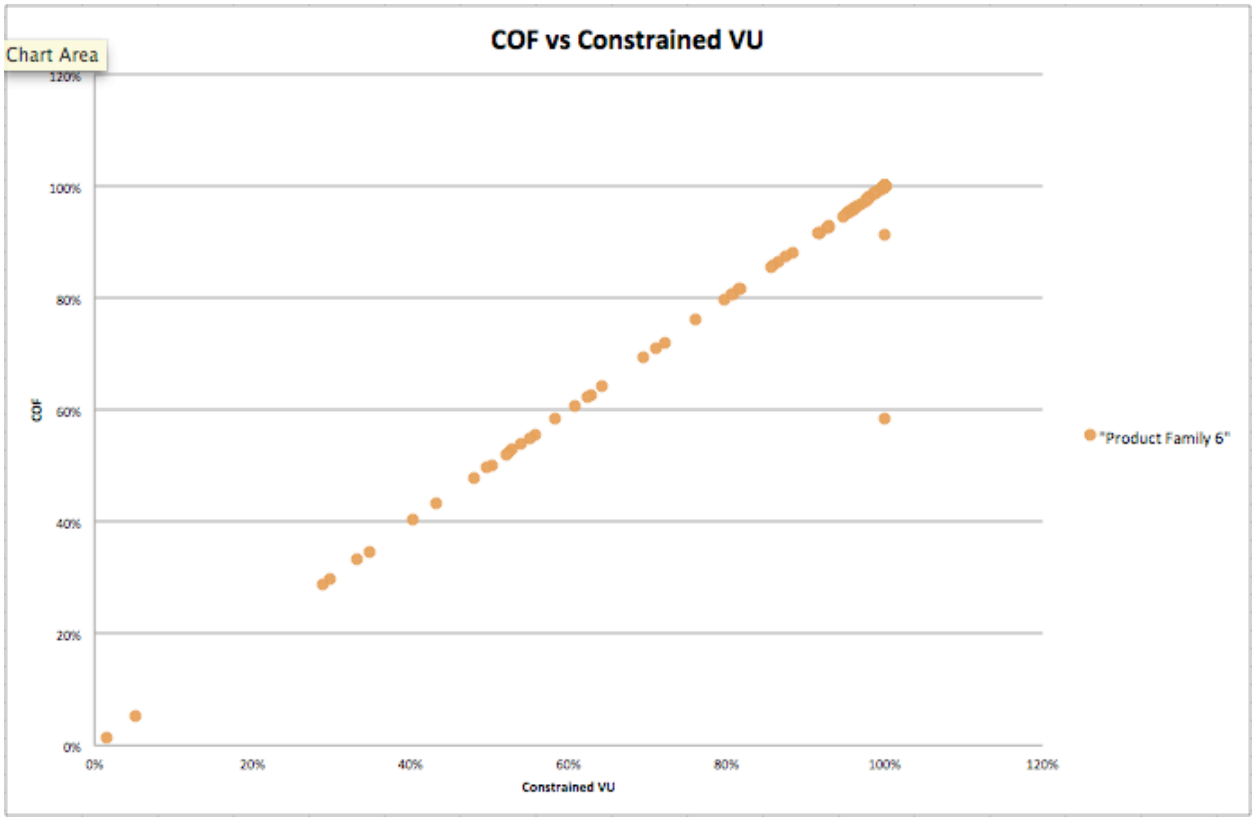
Appendix D5

COF vs Constrained VU – Product Family 5



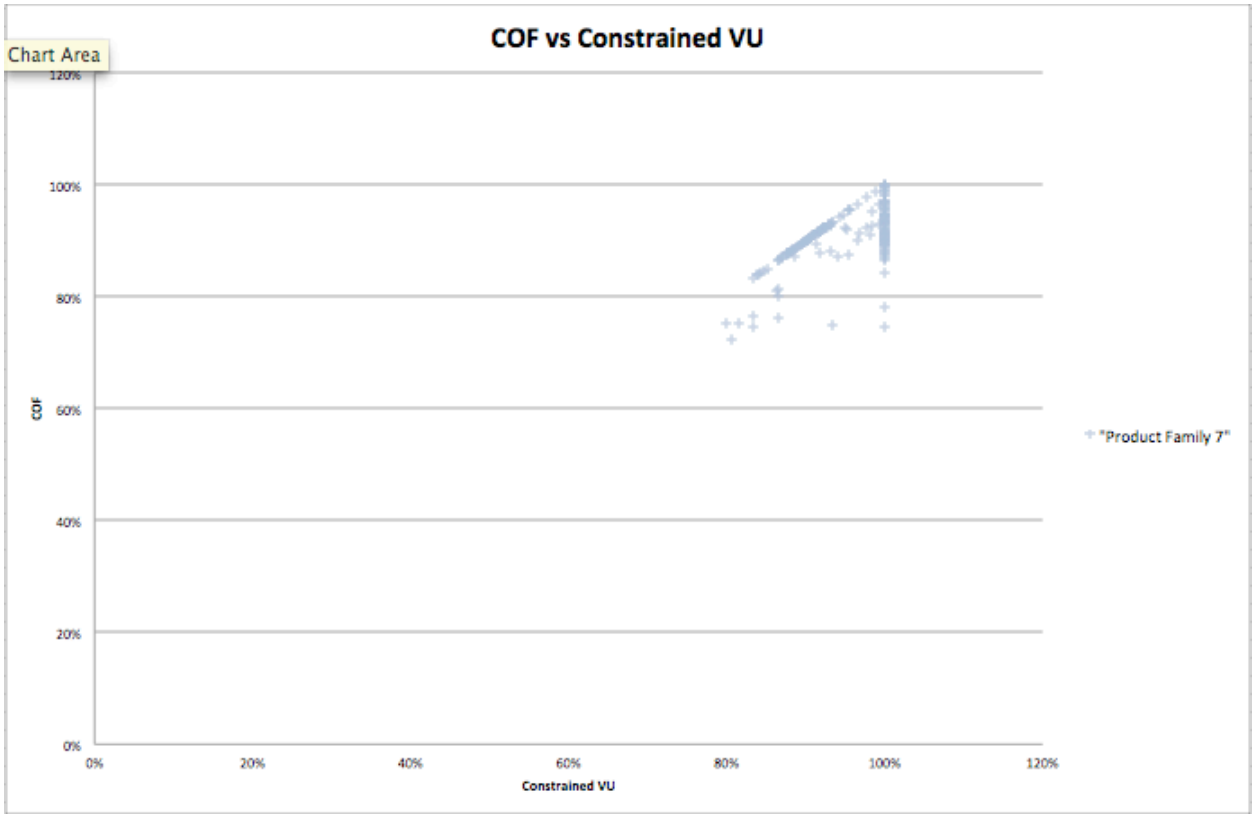
Appendix D6

COF vs Constrained VU – Product Family 6



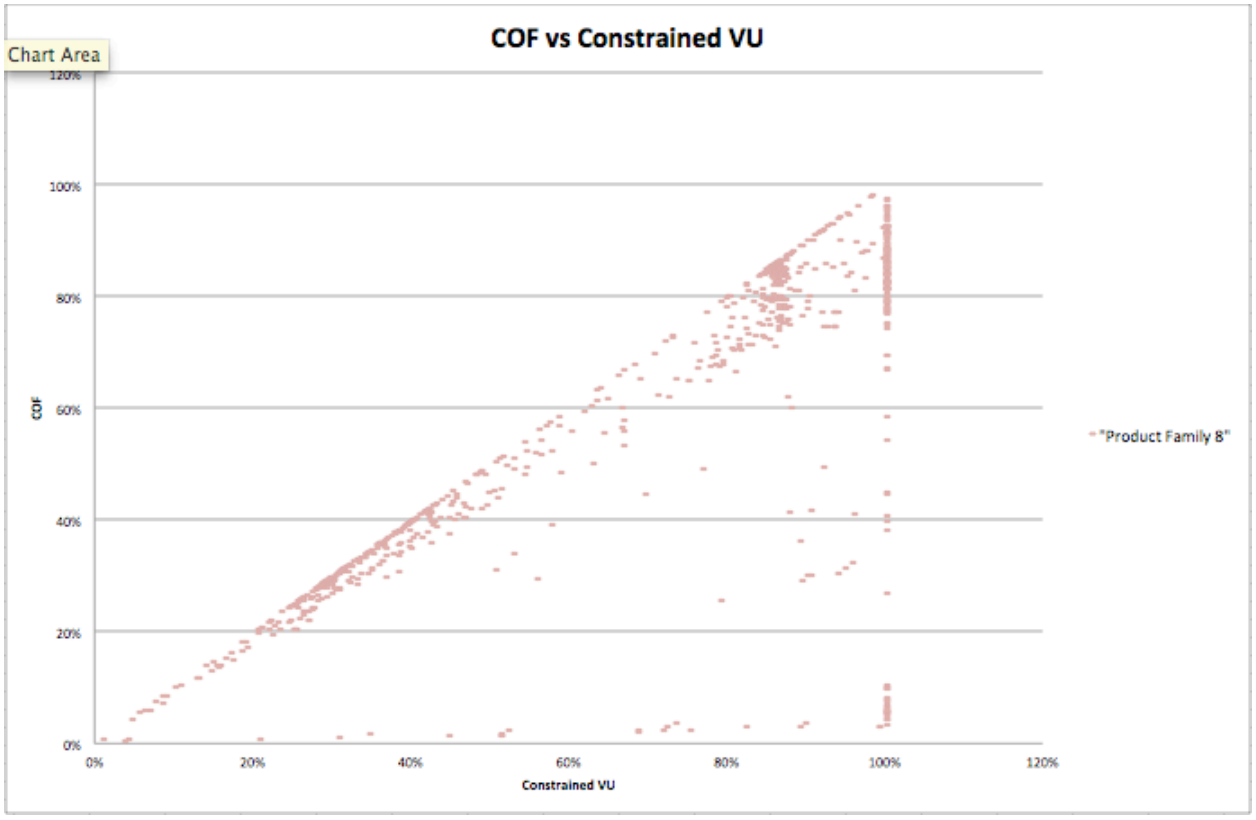
Appendix D7

COF vs Constrained VU – Product Family 7



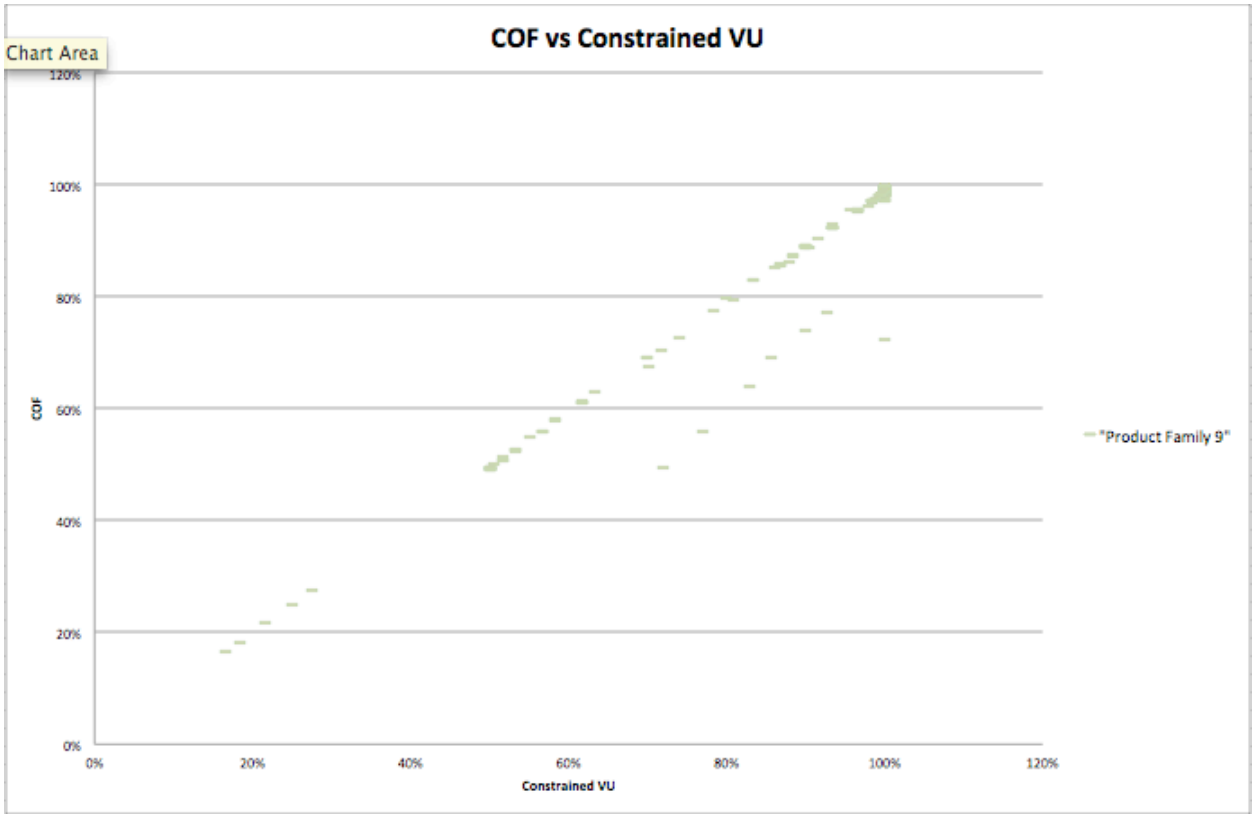
Appendix D8

COF vs Constrained VU – Product Family 8



Appendix D9

COF vs Constrained VU – Product Family 9



APPENDIX E

Appendix E1

Weight Rate Discretization

Limits	Frequency	Accu. Count	Accu. % of Observations	Categorization
0-5%	13	13	0.33%	WGT 1
5-10%	149	162	4.05%	
10-15%	101	263	6.58%	
15-20%	18	281	7.03%	
20-25%	85	366	9.16%	
25-30%	113	479	11.99%	WGT 2
30-35%	82	561	14.04%	
35-40%	66	627	15.69%	
40-45%	72	699	17.49%	
45-50%	69	768	19.22%	
50-55%	74	842	21.07%	WGT 3
55-60%	60	902	22.57%	
60-65%	61	963	24.10%	
65-70%	39	1002	25.08%	
70-75%	58	1060	26.53%	
75-80%	118	1178	29.48%	
80-85%	423	1601	40.07%	WGT 4
85-90%	2328	3929	98.32%	WGT 5
90-95%	67	3996	100.00%	

Appendix E2

Cube Rate Discretization

Limits	Frequency	Accu. Count	Accu. % of Observations	Categorization
0-5%	17	17	0.43%	CUBE 1
5-10%	75	92	2.30%	
10-15%	152	244	6.11%	
15-20%	162	406	10.16%	CUBE 2
20-25%	118	524	13.11%	
25-30%	160	684	17.12%	
30-35%	146	830	20.77%	CUBE 3
35-40%	163	993	24.85%	
40-45%	233	1226	30.68%	
45-50%	358	1584	39.64%	
50-55%	656	2240	56.06%	CUBE 4
55-60%	938	3178	79.53%	CUBE 5
60-65%	564	3742	93.64%	CUBE 6
65-70%	239	3981	99.62%	
70-75%	15	3996	100.00%	
75-80%	0	3996	100.00%	CUBE 7
80-85%	0	3996	100.00%	
85-90%	0	3996	100.00%	
90-95%	0	3996	100.00%	

Appendix E3

FP Rate Discretization

Limits	Frequency	Accu. Count	Accu. % of Observations	Categorization
0-5%	8	8	0.20%	FP 1
5-10%	4	12	0.30%	
10-15%	22	34	0.85%	
15-20%	21	55	1.38%	
20-25%	43	98	2.45%	
25-30%	62	160	4.00%	
30-35%	55	215	5.38%	
35-40%	82	297	7.43%	
40-45%	127	424	10.61%	FP 2
45-50%	92	516	12.91%	
50-55%	109	625	15.64%	
55-60%	140	765	19.14%	
60-65%	165	930	23.27%	FP 3
65-70%	278	1208	30.23%	FP 4
70-75%	532	1740	43.54%	FP 5
75-80%	797	2537	63.49%	FP 6
80-85%	663	3200	80.08%	FP 7
85-90%	423	3623	90.67%	
90-95%	373	3996	100.00%	

Appendix E4

COF Rate Discretization

Limits	Frequency	Accu. Count	Accu. % of Observations	Categorization
0-5%	7	7	0.18%	COF 1
5-10%	6	13	0.33%	
10-15%	24	37	0.93%	
15-20%	35	72	1.80%	
20-25%	70	142	3.55%	
25-30%	73	215	5.38%	
30-35%	63	278	6.96%	
35-40%	91	369	9.23%	
40-45%	128	497	12.44%	COF 2
45-50%	103	600	15.02%	
50-55%	93	693	17.34%	
55-60%	129	822	20.57%	COF 3
60-65%	156	978	24.47%	
65-70%	257	1235	30.91%	COF 4
70-75%	529	1764	44.14%	COF 5
75-80%	728	2492	62.36%	COF 6
80-85%	677	3169	79.30%	
85-90%	428	3597	90.02%	COF 7
90-95%	399	3996	100.00%	

APPENDIX F

Run Information (support: 0.01-0.15; minimum confidence: 0.5)

=== Run information ===

Scheme: weka.associations.Apriori -N 10000 -T 0 -C 0.5 -D 0.05 -U 0.15 -M 0.01 -S -1.0 -c -1

Relation: Data2-2-weka.filters.unsupervised.attribute.Remove-R5-weka.filters.unsupervised.attribute.MathExpression-unset-class-temporarily-Eifelse(A>0.25, ifelse(A>0.5, ifelse(A>0.8, ifelse(A>0.85, ifelse(A>0.95, 6, 5), 4), 3), 2), 1)-R2,3,4-weka.filters.unsupervised.attribute.MathExpression-unset-class-temporarily-Eifelse(A>0.15, ifelse(A>0.3, ifelse(A>0.5, ifelse(A>0.55, ifelse(A>0.6, ifelse(A>0.75, ifelse(A>0.96, 8, 7), 6), 5), 4), 3), 2), 1)-R1,3,4-weka.filters.unsupervised.attribute.MathExpression-unset-class-temporarily-Eifelse(A>0.4, ifelse(A>0.6, ifelse(A>0.65, ifelse(A>0.7, ifelse(A>0.75, ifelse(A>0.8, ifelse(A>0.95, 8, 7), 6), 5), 4), 3), 2), 1)-R1,2,4-weka.filters.unsupervised.attribute.MathExpression-unset-class-temporarily-Eifelse(A>0.4, ifelse(A>0.55, ifelse(A>0.65, ifelse(A>0.7, ifelse(A>0.75, ifelse(A>0.85, ifelse(A>0.95, 8, 7), 6), 5), 4), 3), 2), 1)-R1,2,3-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Instances: 3996
Attributes: 4
Wgt Tot
Cube
FP
COF
=== Associator model (full training set) ===

Apriori
=======
Minimum support: 0.01 (40 instances)
Minimum metric <confidence>: 0.5
Number of cycles performed: 3

Generated sets of large itemsets:

Size of set of large itemsets L(1): 16
Size of set of large itemsets L(2): 39
Size of set of large itemsets L(3): 21
Size of set of large itemsets L(4): 4

Best rules found:

1. Wgt Tot=2 FP=1 105 ==> COF=1 105 conf:(1)
2. Wgt Tot=2 Cube=2 FP=1 70 ==> COF=1 70 conf:(1)
3. Wgt Tot=1 COF=2 53 ==> FP=2 53 conf:(1)
4. Cube=1 FP=1 110 ==> COF=1 109 conf:(0.99)
5. Wgt Tot=1 FP=1 97 ==> COF=1 96 conf:(0.99)
6. Wgt Tot=2 COF=2 75 ==> FP=2 74 conf:(0.99)
7. Wgt Tot=1 Cube=1 FP=1 65 ==> COF=1 64 conf:(0.98)
8. Wgt Tot=4 COF=2 52 ==> FP=2 51 conf:(0.98)
9. FP=1 311 ==> COF=1 305 conf:(0.98)
10. COF=2 323 ==> FP=2 316 conf:(0.98)
11. Cube=2 FP=1 201 ==> COF=1 196 conf:(0.98)
12. Cube=2 FP=2 142 ==> COF=2 138 conf:(0.97)

13. Wgt Tot=3 Cube=2 FP=2 47 ==> COF=2 45 conf:(0.96)
14. Wgt Tot=3 COF=2 90 ==> FP=2 86 conf:(0.96)
15. Cube=2 COF=2 145 ==> FP=2 138 conf:(0.95)
16. Wgt Tot=3 FP=1 71 ==> COF=1 67 conf:(0.94)
17. Wgt Tot=3 Cube=2 FP=1 62 ==> COF=1 58 conf:(0.94)
18. Wgt Tot=4 COF=5 70 ==> FP=5 65 conf:(0.93)
19. Wgt Tot=3 Cube=2 COF=2 49 ==> FP=2 45 conf:(0.92)
20. Wgt Tot=3 Cube=2 COF=1 64 ==> FP=1 58 conf:(0.91)
21. Wgt Tot=1 Cube=1 85 ==> COF=1 76 conf:(0.89)
22. FP=3 165 ==> COF=3 147 conf:(0.89)
23. COF=4 257 ==> FP=4 228 conf:(0.89)
24. Wgt Tot=3 FP=1 71 ==> Cube=2 62 conf:(0.87)
25. Wgt Tot=3 FP=1 COF=1 67 ==> Cube=2 58 conf:(0.87)
26. Cube=2 COF=1 228 ==> FP=1 196 conf:(0.86)
27. Wgt Tot=1 FP=2 62 ==> COF=2 53 conf:(0.85)
28. Cube=1 FP=2 68 ==> Wgt Tot=2 58 conf:(0.85)
29. Wgt Tot=3 COF=1 79 ==> FP=1 67 conf:(0.85)
30. Wgt Tot=1 Cube=1 COF=1 76 ==> FP=1 64 conf:(0.84)
31. COF=5 529 ==> FP=5 443 conf:(0.84)
32. Wgt Tot=3 FP=2 103 ==> COF=2 86 conf:(0.83)
33. Wgt Tot=1 COF=1 115 ==> FP=1 96 conf:(0.83)
34. FP=5 532 ==> COF=5 443 conf:(0.83)
35. COF=1 370 ==> FP=1 305 conf:(0.82)
36. Wgt Tot=4 FP=5 79 ==> COF=5 65 conf:(0.82)
37. FP=4 278 ==> COF=4 228 conf:(0.82)
38. Wgt Tot=3 FP=1 71 ==> Cube=2 COF=1 58 conf:(0.82)
39. Wgt Tot=3 COF=1 79 ==> Cube=2 64 conf:(0.81)
40. Wgt Tot=4 FP=2 63 ==> COF=2 51 conf:(0.81)
41. Wgt Tot=2 Cube=2 COF=1 89 ==> FP=1 70 conf:(0.79)
42. Cube=1 COF=1 142 ==> FP=1 109 conf:(0.77)
43. Wgt Tot=1 Cube=1 85 ==> FP=1 65 conf:(0.76)
44. Wgt Tot=2 COF=1 139 ==> FP=1 105 conf:(0.76)
45. Wgt Tot=1 Cube=1 85 ==> FP=1 COF=1 64 conf:(0.75)
46. Wgt Tot=3 COF=1 79 ==> Cube=2 FP=1 58 conf:(0.73)
47. FP=2 454 ==> COF=2 316 conf:(0.7)
48. Wgt Tot=1 FP=1 97 ==> Cube=1 65 conf:(0.67)
49. Wgt Tot=2 FP=1 105 ==> Cube=2 70 conf:(0.67)
50. Wgt Tot=2 FP=1 COF=1 105 ==> Cube=2 70 conf:(0.67)
51. Wgt Tot=2 FP=1 105 ==> Cube=2 COF=1 70 conf:(0.67)
52. Wgt Tot=1 FP=1 COF=1 96 ==> Cube=1 64 conf:(0.67)
53. Wgt Tot=1 COF=1 115 ==> Cube=1 76 conf:(0.66)
54. Wgt Tot=1 FP=1 97 ==> Cube=1 COF=1 64 conf:(0.66)
55. Wgt Tot=2 FP=2 113 ==> COF=2 74 conf:(0.65)
56. FP=1 311 ==> Cube=2 201 conf:(0.65)
57. FP=1 COF=1 305 ==> Cube=2 196 conf:(0.64)
58. Wgt Tot=2 COF=1 139 ==> Cube=2 89 conf:(0.64)
59. FP=1 311 ==> Cube=2 COF=1 196 conf:(0.63)
60. COF=1 370 ==> Cube=2 228 conf:(0.62)
61. Cube=1 FP=1 110 ==> Wgt Tot=1 65 conf:(0.59)
62. Cube=1 FP=1 COF=1 109 ==> Wgt Tot=1 64 conf:(0.59)
63. Cube=1 244 ==> COF=1 142 conf:(0.58)
64. Cube=1 FP=1 110 ==> Wgt Tot=1 COF=1 64 conf:(0.58)
65. Cube=1 244 ==> Wgt Tot=2 139 conf:(0.57)
66. Wgt Tot=2 Cube=2 157 ==> COF=1 89 conf:(0.57)
67. Wgt Tot=1 COF=1 115 ==> Cube=1 FP=1 64 conf:(0.56)
68. Wgt Tot=3 COF=2 90 ==> Cube=2 49 conf:(0.54)
69. Cube=1 COF=1 142 ==> Wgt Tot=1 76 conf:(0.54)
70. COF=1 370 ==> Cube=2 FP=1 196 conf:(0.53)
71. Wgt Tot=3 FP=2 COF=2 86 ==> Cube=2 45 conf:(0.52)
72. Cube=2 440 ==> COF=1 228 conf:(0.52)

73. COF=3 285 ==> FP=3 147 conf:(0.52)
74. Wgt Tot=2 FP=2 113 ==> Cube=1 58 conf:(0.51)
75. Wgt Tot=2 COF=1 139 ==> Cube=2 FP=1 70 conf:(0.5)
76. Wgt Tot=3 COF=2 90 ==> Cube=2 FP=2 45 conf:(0.5)

APPENDIX G

Run Information (support: 0.025 – 0.25; minimum confidence: 0.5)

=== Run information ===

Scheme: weka.associations.Apriori -N 10000 -T 0 -C 0.5 -D 0.05 -U 0.25 -M 0.025 -S -1.0 -c -1
Relation: Data2-2-weka.filters.unsupervised.attribute.Remove-R5-weka.filters.unsupervised.attribute.MathExpression-unset-class-temporarily-Eifelse(A>0.25, ifelse(A>0.5, ifelse(A>0.8, ifelse(A>0.85, ifelse(A>0.95, 6, 5), 4), 3), 2), 1)-R2,3,4-weka.filters.unsupervised.attribute.MathExpression-unset-class-temporarily-Eifelse(A>0.15, ifelse(A>0.3, ifelse(A>0.5, ifelse(A>0.55, ifelse(A>0.6, ifelse(A>0.75, ifelse(A>0.96, 8, 7), 6), 5), 4), 3), 2), 1)-R1,3,4-weka.filters.unsupervised.attribute.MathExpression-unset-class-temporarily-Eifelse(A>0.4, ifelse(A>0.6, ifelse(A>0.65, ifelse(A>0.7, ifelse(A>0.75, ifelse(A>0.8, ifelse(A>0.95, 8, 7), 6), 5), 4), 3), 2), 1)-R1,2,4-weka.filters.unsupervised.attribute.MathExpression-unset-class-temporarily-Eifelse(A>0.4, ifelse(A>0.55, ifelse(A>0.65, ifelse(A>0.7, ifelse(A>0.75, ifelse(A>0.85, ifelse(A>0.95, 8, 7), 6), 5), 4), 3), 2), 1)-R1,2,3-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Instances: 3996
Attributes: 4
Wgt Tot
Cube
FP
COF
=== Associator model (full training set) ===

Apriori
=======
Minimum support: 0.03 (100 instances)
Minimum metric <confidence>: 0.5
Number of cycles performed: 5

Generated sets of large itemsets:

Size of set of large itemsets L(1): 22
Size of set of large itemsets L(2): 38
Size of set of large itemsets L(3): 8

Best rules found:

1. Cube=3 COF=2 146 ==> FP=2 146 conf:(1)
2. Wgt Tot=2 FP=1 105 ==> COF=1 105 conf:(1)
3. Cube=1 FP=1 110 ==> COF=1 109 conf:(0.99)
4. FP=1 311 ==> COF=1 305 conf:(0.98)
5. COF=2 323 ==> FP=2 316 conf:(0.98)
6. Cube=4 COF=5 334 ==> FP=5 326 conf:(0.98)
7. Cube=2 FP=1 201 ==> COF=1 196 conf:(0.98)
8. Cube=2 FP=2 142 ==> COF=2 138 conf:(0.97)
9. Cube=2 COF=2 145 ==> FP=2 138 conf:(0.95)
10. Cube=3 FP=3 144 ==> COF=3 135 conf:(0.94)
11. FP=3 COF=3 147 ==> Cube=3 135 conf:(0.92)
12. FP=3 165 ==> COF=3 147 conf:(0.89)
13. COF=4 257 ==> FP=4 228 conf:(0.89)
14. Cube=3 COF=4 212 ==> FP=4 187 conf:(0.88)
15. FP=3 165 ==> Cube=3 144 conf:(0.87)
16. Cube=4 FP=5 377 ==> COF=5 326 conf:(0.86)

17. COF=3 285 ==> Cube=3 246 conf:(0.86)
18. Cube=2 COF=1 228 ==> FP=1 196 conf:(0.86)
19. Cube=3 FP=4 218 ==> COF=4 187 conf:(0.86)
20. COF=5 529 ==> FP=5 443 conf:(0.84)
21. FP=5 532 ==> COF=5 443 conf:(0.83)
22. COF=4 257 ==> Cube=3 212 conf:(0.82)
23. COF=1 370 ==> FP=1 305 conf:(0.82)
24. FP=4 COF=4 228 ==> Cube=3 187 conf:(0.82)
25. FP=4 278 ==> COF=4 228 conf:(0.82)
26. FP=3 165 ==> Cube=3 COF=3 135 conf:(0.82)
27. FP=4 278 ==> Cube=3 218 conf:(0.78)
28. Cube=1 COF=1 142 ==> FP=1 109 conf:(0.77)
29. Wgt Tot=2 COF=1 139 ==> FP=1 105 conf:(0.76)
30. Cube=6 818 ==> COF=7 615 conf:(0.75)
31. COF=7 827 ==> Cube=6 615 conf:(0.74)
32. FP=5 COF=5 443 ==> Cube=4 326 conf:(0.74)
33. COF=4 257 ==> Cube=3 FP=4 187 conf:(0.73)
34. FP=5 532 ==> Cube=4 377 conf:(0.71)
35. FP=2 454 ==> COF=2 316 conf:(0.7)
36. FP=4 278 ==> Cube=3 COF=4 187 conf:(0.67)
37. FP=1 311 ==> Cube=2 201 conf:(0.65)
38. FP=1 COF=1 305 ==> Cube=2 196 conf:(0.64)
39. COF=5 529 ==> Cube=4 334 conf:(0.63)
40. FP=1 311 ==> Cube=2 COF=1 196 conf:(0.63)
41. COF=5 529 ==> Cube=4 FP=5 326 conf:(0.62)
42. COF=1 370 ==> Cube=2 228 conf:(0.62)
43. FP=6 797 ==> Cube=5 491 conf:(0.62)
44. FP=5 532 ==> Cube=4 COF=5 326 conf:(0.61)
45. Cube=3 FP=2 244 ==> COF=2 146 conf:(0.6)
46. Cube=1 244 ==> COF=1 142 conf:(0.58)
47. Cube=4 656 ==> FP=5 377 conf:(0.57)
48. Cube=1 244 ==> Wgt Tot=2 139 conf:(0.57)
49. Cube=3 COF=3 246 ==> FP=3 135 conf:(0.55)
50. Wgt Tot=1 366 ==> Cube=3 200 conf:(0.55)
51. FP=2 454 ==> Cube=3 244 conf:(0.54)
52. COF=1 370 ==> Cube=2 FP=1 196 conf:(0.53)
53. Cube=5 938 ==> FP=6 491 conf:(0.52)
54. Cube=2 440 ==> COF=1 228 conf:(0.52)
55. COF=3 285 ==> FP=3 147 conf:(0.52)
56. Cube=4 656 ==> COF=5 334 conf:(0.51)

APPENDIX H

Run Information (support: 0.25 – 0.1; minimum confidence: 0)

```
=== Run information ===

Scheme: weka.associations.Apriori -N 10000 -T 0 -C 0.0 -D 0.05 -U 1.0 -M 0.25 -S -1.0 -c -1
Relation: Data2-2-weka.filters.unsupervised.attribute.Remove-R5-weka.filters.unsupervised.attribute.MathExpression-unset-class-temporarily-Eifelse(A>0.25, ifelse(A>0.5, ifelse(A>0.8, ifelse(A>0.85, ifelse(A>0.95, 6, 5), 4), 3), 2), 1)-R2,3,4-weka.filters.unsupervised.attribute.MathExpression-unset-class-temporarily-Eifelse(A>0.15, ifelse(A>0.3, ifelse(A>0.5, ifelse(A>0.55, ifelse(A>0.6, ifelse(A>0.75, ifelse(A>0.96, 8, 7), 6), 5), 4), 3), 2), 1)-R1,3,4-weka.filters.unsupervised.attribute.MathExpression-unset-class-temporarily-Eifelse(A>0.4, ifelse(A>0.6, ifelse(A>0.65, ifelse(A>0.7, ifelse(A>0.75, ifelse(A>0.8, ifelse(A>0.95, 8, 7), 6), 5), 4), 3), 2), 1)-R1,2,4-weka.filters.unsupervised.attribute.MathExpression-unset-class-temporarily-Eifelse(A>0.4, ifelse(A>0.55, ifelse(A>0.65, ifelse(A>0.7, ifelse(A>0.75, ifelse(A>0.85, ifelse(A>0.95, 8, 7), 6), 5), 4), 3), 2), 1)-R1,2,3-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Instances:    3996
Attributes:   4
              Wgt Tot
              Cube
              FP
              COF
=== Associator model (full training set) ===

Apriori
=======
Minimum support: 0.25 (999 instances)
Minimum metric <confidence>: 0
Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsets L(1): 3
Size of set of large itemsets L(2): 2

Best rules found:

 1. COF=6 1405 ==> Wgt Tot=5 1146    conf:(0.82)
 2. FP=7 1459 ==> Wgt Tot=5 1062    conf:(0.73)
 3. Wgt Tot=5 2395 ==> COF=6 1146    conf:(0.48)
 4. Wgt Tot=5 2395 ==> FP=7 1062    conf:(0.44)
```

VITA

Heng (Miranda) Li

EDUCATION

The Pennsylvania State University — University Park, PA
Schreyer Honors College — *Class of 2011*

- Smeal College of Business, College of the Liberal Arts
  - — B.S. in *Supply Chain and Information Systems*
  - — B.A. in *Economics*
  - — Minor in *International Business*

The Institute for the International Education of Students (IES) — Freiburg, Germany

- The European Union Program — *Fall 2009*
  - — Studied international economic relations of the European Union, transition economies in Central and Eastern Europe, and the European Union's political system and environmental policies
  - — Took field study trips to Berlin, Riga, Luxembourg, Brussels, Paris, Strasbourg, Krakow, Prague, Budapest, and Geneva

PROFESSIONAL EXPERIENCE

China Unicom (Provincial Branch of Guangdong) — Guangzhou, China
Group Customer Service Dept. Intern — *06/2010 – 07/2010*

- Worked alongside the industry manager who supervised projects targeting clients in the transportation and logistics industry
- Assisted with procurement planning, product package and scheme customization for organizations and companies with special needs

Penn State Schreyer Honors College — University Park, PA
Scholar Assistant (Part-Time) — *04/2010 – 05/2011*

- Work directly under the Associate Dean to identify academic advising resources
- Organize special activities and events and promote a wide range of programs offered to Schreyer Scholars

Penn State Executive Programs — University Park, PA
Logistics Assistant (Part-Time) — *09/2008 – 05/2011*

- Assist with administrative operations, program preparation, and program delivery

DMG – Dynamic Marketing Group — Beijing, China
Strategic Planning Intern — *05/2008 – 08/2008*

- Worked within an international planning team representing non-Chinese brands entering into or expanding within China
- Gained competence in strategic planning, brand development, market review, competitive review, trend analysis, consumer surveying, insight analysis, and mood-board development
- Participated in pitches and campaigns for: *Volkswagen*, *Spalding*, *Under Armour*, *Ibis Hotels*, *Johnson & Johnson*, *and Nike*

RESEARCH EXPERIENCE

The Smart Spaces Center for Adaptive Aging in Place — University Park, PA
Undergrad Researcher for the International Comparison of Successful Aging Project—China — *11/2008 – 06/2009*

- Conducted research on changes in Chinese citizens' lives, national policies, and economic structure due to an aging demography
- Presented the research paper, *Analysis of the Graying China: Current and Future Challenges of the Nation's Aging Population*, on July 1st, 2009, at *The First International Symposium on Quality of Life Technology* sponsored by Carnegie Mellon University and the University of Pittsburgh

LEADERSHIP/ACTIVITIES

Hong Kong Student Association — University Park, PA
Vice President — *03/2010 – 05/2011*

- In charge of the club's public relations; work with other officers to coordinate activities and events that promote Cantonese culture and friendship

SAP Interest Group — University Park, PA
Member — *01/2010 – 05/2011*

- Learning to use SAP and other ERP systems; enhancing software skills needed in the supply chain industry

Schreyer Honors College Speaker Series Committee — University Park, PA
Board Member — *03/2008 – 12/2010*

- Conduct research on scholars and nobles, select candidates, and coordinate funding and partnerships for Speaker Series events

Habitat for Humanity — University Park, PA
Member — *09/2007 – 12/2010*

- Execute fundraising events and build houses to support low-income families

HONORS

- Penn State Academic Excellence Scholarship (Fall 07~ Present)
- Schreyer Honors College Summer Internship Grant (Summer 08)
- The National Society of Collegiate Scholars
- Schreyer Ambassador Travel Grant (Fall 09)

SKILLS AND INTERESTS

- Proficient in Microsoft Office products including Word, Excel, PowerPoint, Publisher, Access
- Fluent in English, Mandarin and Cantonese; working towards proficiency in German and Japanese
- Passionate for philately, contemporary art, all phases of theater, and ceramics