ASSORTMENT PLANNING FOR A LARGE RETAILER
WITH SHELF-SPACE CONSTRAINT

A Thesis in
Industrial Engineering and Operations Research
by
KuangYu Wu

© 2012 KuangYu Wu

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2012
The thesis of KuangYu Wu was reviewed and approved* by the following

Saurabh Bansal
Assistant Professor of Supply Chain Management
Thesis Co-advisor

Robert Novack
Associate Professor of Supply Chain and Information Systems

Paul Griffin
Professor of Industrial Engineering
Peter and Angela Dal Pezzo Department Head Chair
Head of the Department of Industrial Engineering
Thesis Co-advisor

M. Jeya Chandra
Professor of Industrial Engineering
Academic Programs & Graduate Program Coordinator

*Signatures are on file in the Graduate School
ABSTRACT

Assortment planning has become an important issue for most retailers due to the growing variety of the products. Every retailer tries to determine her best product combination in order to maximize profit. In this research, we present an in-depth analysis for product assortment planning for a large retail chain that seeks to reduce its assortment to reduce the floor size of the store.

Keywords: assortment planning
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>viii</td>
</tr>
<tr>
<td>Chapter 1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Background of research problem</td>
<td>1</td>
</tr>
<tr>
<td>1.2 ABC Analysis and Association Rules</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Objective and Thesis Organization</td>
<td>3</td>
</tr>
<tr>
<td>Chapter 2 Literature Review</td>
<td>4</td>
</tr>
<tr>
<td>2.1 Category Management</td>
<td>4</td>
</tr>
<tr>
<td>2.2 Assortment Planning</td>
<td>5</td>
</tr>
<tr>
<td>2.3 ABC Analysis</td>
<td>7</td>
</tr>
<tr>
<td>2.3.1 R-model</td>
<td>8</td>
</tr>
<tr>
<td>2.3.2 ZF-model</td>
<td>10</td>
</tr>
<tr>
<td>2.4 Market Basket Analysis</td>
<td>12</td>
</tr>
<tr>
<td>2.4.1 Definition</td>
<td>13</td>
</tr>
<tr>
<td>2.4.2 Concepts</td>
<td>13</td>
</tr>
<tr>
<td>Chapter 3 Methodology and Analysis</td>
<td>15</td>
</tr>
<tr>
<td>3.1 Problem Statement</td>
<td>15</td>
</tr>
<tr>
<td>3.2 Data Description</td>
<td>16</td>
</tr>
<tr>
<td>3.3 Analysis Process</td>
<td>18</td>
</tr>
<tr>
<td>3.4 Empirical Study</td>
<td>20</td>
</tr>
<tr>
<td>3.4.1 Data Observation</td>
<td>20</td>
</tr>
<tr>
<td>3.4.2 Analysis Result</td>
<td>21</td>
</tr>
<tr>
<td>3.5 Discussion</td>
<td>23</td>
</tr>
<tr>
<td>3.5.1 Basket Component</td>
<td>23</td>
</tr>
<tr>
<td>3.5.2 Percentage Suggestion</td>
<td>25</td>
</tr>
<tr>
<td>3.5.3 Revenue vs. Volume</td>
<td>27</td>
</tr>
<tr>
<td>Chapter 4 Conclusion and Future Research</td>
<td>31</td>
</tr>
<tr>
<td>4.1 Summary</td>
<td>31</td>
</tr>
<tr>
<td>4.2 Conclusion</td>
<td>32</td>
</tr>
<tr>
<td>4.3 Future Research</td>
<td>32</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 2-1. The Category Management Process (Basuoy et al., 2001) ......................5
Figure 3-1. Basket SKUs Component .....................................................................24
Figure 3-2. Ratio of Guaranteed Revenue .................................................................25
Figure 3-3. Marginal Benefit of SKUs .....................................................................26
Figure 3-4. Pickiness for Baskets ............................................................................27
Figure 3-5. Percentage of Incomplete Basket ..........................................................28
LIST OF TABLES

Table 2-1. Transactions Data ................................................................. 13
Table 3-1. Criterion................................................................................. 20
Table 3-2. Point of Sale Data................................................................. 20
Table 3-3. Results base on Volume ......................................................... 21
Table 3-4. Results base on Revenue ....................................................... 22
Table 3-6. Impact on Divisions............................................................... 29
Table A-1. Simple Data Set Example ...................................................... 37
Table A-2. Definition of Dataset ............................................................ 38
ACKNOWLEDGEMENTS

This thesis couldn’t have been completed without the generous assistance of my academic advisor: Dr. Saurabh Bansal. Because of his intellectual guidance, his insightful comments, and his remarkable patience, this thesis was able to progress towards its full-fledged form from an early immature stage. Thanks to his help and encouragement throughout the whole process of composing this thesis, all the bits and pieces of the original ideas were able to be developed and structured into the coherent format of its final version. Therefore, I feel deeply grateful for his support, both academically and personally. Also, I would like to express my special gratitude to my committee member Dr. Paul Griffin, Dr. Robert Novack and Dr. Jeya Chandra, who review my work in the final phase of my M.S. thesis, and whose valuable comments have provided me with new insights into the way we perceive language in use. Due to the help of Dr. Bansal, Dr. Griffin, Dr. Novack and Dr. Chandra, the achievement of this thesis is one of the most rewarding experiences in my graduate years. Also, special thanks to Rashmi Sharma, whose kind help and constructive comments gave me a lot of inspirations in the final phase of conducting this thesis. Finally, thanks to my family and friends who gave me so many support on my journey of graduate life.
Chapter 1

Introduction

The objective of assortment planning is to determine a set of products that maximizes profit subject to different constraints, for example, budget for purchase of products, shelf space for displaying products, and other varieties of different constraints (Kok et al. 2006). This problem has become more severe in recent years since the number of items in the marketplace increased by 16% per year between 1985 and 1992, while shelf space expanded by only 1.5% per year during the same period (Quelch and Kenny 1994). Therefore, assortment planning has gained more and more attention in recent years.

Most retailers organize their stock keeping units (SKU) into different groups called categories, and each category may further be divided into many subcategories. Retailers try to figure out how to balance product quantity and variety between different categories according to their budget and shelf-space; this is called category management (CM) (Basuroy et al., 2001). The major focus of this research is assortment planning which provides the first selection of products that can be then adjusted for considerations regarding categories.

1.1 Background of research problem

This research is motivated by a problem faced by a large retail chain. The retail chain wants to replace its existing stores with 20,000 square feet with smaller stores that
have a 5,000 square-feet area. The main motivation for opening smaller stores is the rising fare of real estate. The firm determined that the revenue from the larger store could no longer cover the rent. Therefore, the firm decided to open smaller stores with a smaller assortment. Many questions arose as a result of this decision:

a. Which products should the stores carry?

b. Which categories should the firm eliminate from the assortment?

c. Within each category, which products should the firm keep?

Each question above can be seen as a decision variable, where the main business objective is to maximize the profit and sales. The retailer hopes to develop a way to maximize her profit with limited space resources. The objective of this thesis is to provide solutions to these questions for the firm.

1.2 ABC Analysis and Association Rules

ABC analysis is a very simple analysis technique that has been used extensively in material planning. It simply classifies materials or products into category A, B and C. ‘A’ segment represents important objects, ‘B’ represents less important objects and ‘C’ represents relatively unimportant objects. There are many ways to determine the classifications, usually by the total score of each product, for example, by the total sales quantity or the sum of revenue.

Association rule is a technique in market basket analysis. It is used to find out if there is any relation between product sales. To be more specific, the association rules
ascertain the probability that a customer who buys product A may also buy product B. If the probability is large enough, then the association rule stands out.

ABC analysis and association rule are two techniques being widely used in retail business, since they are not only simple to apply but very understandable as well. More details about ABC analysis and association rule will be provided in chapter 2.

1.3 Objective and Thesis Organization

The objective of this thesis is to solve the problem mentioned in the section above; maximize the profits with limited resources by determining the types and quantities of SKUs, or simpler, obtain the right SKUs. We look at the literature to understand current possible solutions, and implement a few suggested methods. We analyze the pros and cons of different approaches and identify future solutions. Chapter 2 provides a literature review of the different techniques that have already been implemented in assortment planning.

Chapter 3 shows the processes and results of the analysis, with the data sets displayed in detail and analysis exhibited step by step. Results are presented in the form of graphs and tables, which can be understood clearly.

Chapter 4 presents the conclusion of results, which indicate the best methods that can be applied. Finally, potential ideas for future research are suggested.
Chapter 2

Literature Review

This chapter presents a summary of the existing literature on the chief focus of this thesis. We give a brief introduction of category management in section 2.1, and an in-depth summary of studies related to assortment planning in section 2.2. In Sections 2.3 and 2.4, we introduce studies of ABC analysis and market basket analysis respectively, which are the main implementation focus of this research.

2.1 Category Management

The Institute of Grocery Distribution (IGD) defines category management as “The strategic management of product groups through trade partnerships which aims to maximize sales and profit by satisfying consumer and shopper needs (IGD, 1999).” Basuroy et al. (2001) presented an eight-stage process for category management based on the Joint Industry Report on Efficient Consumer Response (1995). Through these planning stages (see Figure 2-1), retailers can easily implement category management from analysis to practice.

As shown in Figure 2-1, assortment planning is listed in step “category tactics”. Lindblom and Olkkonen (2006) divided category management tactics into four main topics: assortment planning, pricing, space allocation and in-store promotional activity.
2.2 Assortment Planning

Assortment planning is still a relatively new topic in academic fields, but the related research addressing different aspects of assortment planning is growing rapidly. The brief definition for assortment planning is “to decide how many and which products to include in the product line and to determine the inventory levels of different products for a retailer” (Rajaram, 2001). Kök et al. (2006) had published a very detailed literature review with case studies of assortment planning.
Different issues are involved through the whole assortment planning process, such as demand estimation, objective function construction, algorithms applications and constraints implementation. Assortment planning process can be considered as an optimization process, where we are seeking the best solution to meet a certain objective, i.e. maximizing profit or minimizing cost, with different constraints such as budget or space resources. No matter what kind of analytical assortment planning process retailers choose to implement, they have to start with a customer choice model and certain demand parameters from estimation (Cachon et al., 2005). Therefore, constructing a proper demand model is a very important task in assortment planning.

Different analytical approaches to assortment planning have been proposed. The multinomial logit (MNL) model is a utility-based model which is widely used in many areas (e.g., Ben-Akiva and Lerman 1985, Anderson et al. 1992), and has recently been used in assortment planning. In the MNL model, each customer chooses the product that maximizes her own utility. The utility $U_i$ of product $i$ is given by $U_i = \mu_i + \zeta_i$, where $\mu_i \in \mathbb{R}$ represents the mean utility that the customer assigns to product $i$. The model assumes that $\zeta_0, ..., \zeta_N$ are independent and identically distributed random variables having a Gumbel distribution with location parameter 0 and scale parameter 1.

Van Ryzain and Mahajan (1999) formulate customer demand behavior in assortment planning using the MNL model, where they prove an intuitively appealing result; that the optimal assortment will always include the most popular products, which is proved by the newsvendor model (or newsboy model). Cachon et al. (2005) also use the MNL model for customer behavior of searching products. Rusmevichientong et al. (2010) use the MNL choice model to construct customer demand preferences and
optimize them in static and dynamic choice situations. Saure and Zeevi (2009) study the assortment problem by balancing the tradeoff between exploration and exploitation. To be more specific; when there is no historical information, retailers should dynamically adjust their preferences based on demand observations in order to maximize profits. When retailers spend more time exploring customers’ behavior, they have less time to implement exploitation, i.e. optimization of assortment, and vice versa.

Substitution behavior is also an important topic in assortment planning, since it has the potential to change the assortment. Van Ryzin and Mahajan (1999) develop a model where a customer will substitute if her favorite product is not in store permanently, but will leave if the product is temporary out of stock. Kök and Fisher (2007) also propose a method for assortment planning under the concept of substitution, while Goyal et al. (2009) proposed a near-optimal algorithm under dynamic substitution and stochastic demand.

2.3 ABC Analysis

ABC analysis is a very commonly use method in inventory management. Basically, the analysis is based on the Pareto principle and distinguishes products into three types; A is the item with the highest level and C is the lowest, where B is in between. There are numerous of ways to perform ABC analysis; the most straightforward way is to classify items into three category based on one criterion, such as the cost or the dollar-usage. Teunter et al. (2009) proposed a new “cost criterion” for classification; the criterion is computed as $\frac{b_i D_i}{h_i Q_i}$, where $b_i$ is the shortage cost, $D_i$ is the
demand rate, \( h_i \) is the inventory holding cost, and \( Q_i \) is the order quantity for SKU \( i \). Since the shortage cost \( b_i \) models the criticality of a product, they obtained a nearly optimal assortment for minimizing inventory costs.

Flores and Whybark (1988) are the pioneers of integrating the “multi-criteria” ideas into ABC analysis; they suggested that there are many other criteria that can represent significant factors. They stated two methods to implement multi-criteria; the first one is based on the importance while the second one is based on policy driven factors. Ramanathan (2006) used a weighted linear optimization model for the classification of inventory items that is very similar to data envelopment analysis (DEA), which is a methodology used to determine relative efficiencies between decision units (DMU). We give a detailed introduction below:

2.3.1 R-model

Ramanathan (2006) presented a multi-criteria ABC analysis model, which is built on the concept of weighted linear optimization. The model is shown below:

\[
\begin{align*}
\max & \quad \sum_{j=1}^{J} \upsilon_{mj}y_{mj} \\
\text{s. t.} & \quad \sum_{j=1}^{J} \upsilon_{mj}y_{nj} \leq 1, \quad n = 1, 2, \ldots, N, \quad \text{(R-model)} \\
\upsilon_{mj} & \geq 0, \quad j = 1, 2, \ldots, J.
\end{align*}
\]

Where,

\( \upsilon_{mj} \): Weight of the \( m^{th} \) product in \( j^{th} \) criteria
$y_{mj}$: Performance of the $m^{th}$ product in $j^{th}$ criteria

N represents the number of products

J represents the number of criteria

In the R-model, it is assumed that all the criteria have a positive relation with the importance level assigned to every product, i.e. the larger the score that a product gets, the more important it is (have greater chance of being classified as level A). The objective function is a weighted additive function, which is used to aggregate the performance score of a product. Weights ($u_{mj}$) are the decision variables of the R-model, they are computed using the optimization routine subject to the constraint that the weighted sum for all items should be less than or equal to one.

The R-model can create a set of criterion weights for each product and distribute a normalized score to every product for further ABC analysis. Moreover, it is a very flexible model, which can easily support new criteria. However, despite its many advantages, R-model could lead to a situation where a product with a high performance score in one insignificant criterion can be classified as A. Therefore, Zhou and Fan (2007) presented an extension of the R-model, called the ZF-model, with which they tried to improve the situation by using two sets of weights. The ZF-model provides a more comprehensive ABC analysis than the R-model as it can eliminate the extreme case that occurs in the R-model.
2.3.2 ZF-model

Zhou and Fan (2007) proved that in the R-model, if a product has a criterion performance that dominates other criteria, this product will always get an aggregate performance score of one. This situation may cause inappropriate classification to some products. Therefore, they proposed a more comprehensive model, the ZF-model, which incorporated some balancing features for multi-criteria ABC analysis. The extension model is shown below:

\[
\min \sum_{j=1}^{J} \omega_{mj}y_{mj}
\]

s.t. \( \sum_{j=1}^{J} \omega_{mj}y_{nj} \geq 1, n = 1,2,\ldots,N \) \quad \text{(ZF-model)}

\( \omega_{mj} \geq 0, j = 1,2,\ldots,J \).

Where,

- \( \omega_{mj} \): Weight of the \( m^{th} \) product in \( j^{th} \) criteria
- \( y_{mj} \): Performance of the \( m^{th} \) product in \( j^{th} \) criteria
- \( N \) represents the number of products
- \( J \) represents the number of criteria

As we can see, the ZF-model is very similar to the R-model; the weights are also determined by using a weighted linear optimization subjected to the aggregate weighted constraints. As an end result, each product will get a set of weights that are least favorable for it.
Since we can obtained the most favorable weights from the R-model and the least favorable weights from the ZF-model, we can term them as the “good index” and the “bad index”, and build a composite index by combining the two. Steps are shown as follow:

1. From R-model, set \( \max \sum_{j=1}^{l} v_{mj} y_{mj} = gI_m \) (good index)
2. From ZF-model, set \( \min \sum_{j=1}^{l} \omega_{mj} y_{mj} = bI_m \) (bad index)
3. Calculate composite index \( nI_m(\lambda) = \lambda \cdot \frac{gI_m - gI^-}{gI^-} + (1 - \lambda) \cdot \frac{bI_m - bI^-}{bI^-} \)

Where \( gI^+ = \max\{gI_m, m = 1,2, ..., N\}, gI^- = \min\{gI_m, m = 1,2, ..., N\} \),
\( bI^+ = \max\{bI_m, m = 1,2, ..., N\}, bI^- = \min\{bI_m, m = 1,2, ..., N\} \)

\( \lambda \) is a control parameter that can acquire a value between 0 and 1, and is used to reflect the preference of decision maker on the good and the bad index. If \( \lambda = 1 \), \( nI_m \) will become the good index version. If \( \lambda = 0 \), \( nI_m \) will become the bad index version. If there is no preference between the good index and the bad index, \( \lambda = 0.5 \) could be a fair and reasonable selection.

The R-model and the ZF-model lead the research fields through the multi-criteria ABC analysis era; Ng (2007) later proposed a model with multiple criteria classification, which can be solved without using linear programming by proper transformation. Hadi-Vencheh (2010) studied Ng (2007) model and presented an improved model, which maintained the multi criteria classification features and added weights effect through the solutions. Chen (2011) improved the R and ZF models by implementing a peer-estimation approach into the classification procedure, such that the performance index became more comprehensive and meaningful.
While ABC analysis based methods provide several good ways to classify items, it misses the market basket analysis, which is discussed in the next section.

### 2.4 Market Basket Analysis

Market basket analysis is a data analysis technique to find out relationships between customer purchasing behaviors when selecting merchandise through transactional data, i.e. when a customer purchases a stapler, there is a good chance that she will also purchase some staple-pins.

Market basket analysis can assist retail management in making different decisions, such as which items to put in a store, which items to put on sale, which items should be placed together, etc (Agrawal et al., 1993). For example, if a retail store finds out that there is a good chance that the customer will buy butter when she buys milk and bread, they can put these items closer to each other or even run a promotion for this specific product combination.

Agrawal et al. (1993) proposed a data mining technique for market basket analysis, which is called association rules. They defined a confidence parameter and a support parameter calculated using the transaction data, where confidence is a measure of the reliability of association rules and support implies how frequently the rules occur.
2.4.1 Definition

The original definition presented by Agrawal et al. (1993) is as follows: Let $I = \{i_1, i_2, ..., i_n\}$ be a set of $n$ products with binary attributes. Let $T = \{t_1, t_2, ..., t_m\}$ be a set of transaction data, where each transaction $T$ has a unique ID and contains a subset of products in $I$. $X \Rightarrow Y$ is defined as a rule, where $X$ and $Y$ are both products from $I$ while $X \cap Y = \emptyset$. We illustrate a very simple example below:

<table>
<thead>
<tr>
<th>Transaction($t_m$)</th>
<th>Pencil($i_1$)</th>
<th>Ruler($i_2$)</th>
<th>Eraser($i_3$)</th>
<th>Scissors($i_4$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$t_2$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$t_3$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$t_4$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$t_5$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

From Table 3, we can see the set of item $I = \{\text{Pencil, Ruler, Eraser, Scissors}\}$, and an example rule could be $\{\text{Pencil, Ruler}\} \Rightarrow \{\text{Eraser}\}$ which means that if pencil and ruler are sold, eraser will also be sold on the same transaction.

2.4.2 Concepts

There are two main concepts in association rules analysis, support and confidence. The support of a product set $X$ is defined as the proportion of transactions that contain the product set. In the previous example, $\{\text{Pencil, Ruler, Eraser}\}$ has a support of $3/5=0.6$, meaning it occurs in 60% of the transactions. Another concept is the confidence of the rule, which is the calculation of the reliability. The confidence is calculated as a ratio of
the union of support X and support Y, and support X. For the previous example, the confidence of \{Pencil, Ruler\} \Rightarrow \{Eraser\} is \(0.6/0.8=0.75\) (3 transactions out of 4), meaning that 75\% of the transactions which contain pencil and ruler also contain eraser.

It should be noted that support and confidence are different. Support corresponds to statistical significance, while confidence is the measure of rule strength. Normally support and confidence are defined by decision makers. The association rule analysis is usually divided into two phases based on these concepts; 1) minimum support is defined in order to find out frequent products in data. 2) minimum confidence is defined through the frequency of products to form rules; association rules will stand only when the confidence of products is greater than minimum confidence. 

Brijs et al. (1999) presented an assortment selection model called PROFSET which is base on association rules. PROFSET used the PROFitability per frequent SET to determine the optimal selection of products in terms of maximal total profit; the model combined cross-selling effects by using frequent item sets.

Wong et al. (2005) presented a method which gives recommendation for item-sets under the concept of association rules. They proposed a new approach called loss rule, which is defined in a similar way as the association rules; the loss rule is used in the formulation of the total profit of the item selection. They used different mathematical approaches, such as quadratic, heuristic, and evolutionary, to solve the MPIS problem, and showed the approach to be efficient and effective using empirical studies.
Chapter 3
Methodology and Analysis

3.1 Problem Statement

Determining the ideal set of assortment is always a major issue for every retailer. Van der Ster and van Wissen (1993) stated two important goals for optimal product assortment to meet. The first goal is from the aspect of quality; products assortment should satisfy the image of the store, for example, an office supply retail store should contain pen, papers, staplers, etc, which are related to office business. To achieve this goal, products are usually been distinguished into two types, basic products and added products. Basic products are the products that should be kept in a store, even if they didn't provide a very big profit, while they provide a fundamental support of store image. Added products are the products selected by the retailer in order to enhance the store image and make more profit by potential cross sales. For example, for an office supply retailer, black pencils, black pens, and white papers should be basic products, while different colors of pencils, pens and papers can be added products.

A second goal for an optimal product is quantitative; it is very obvious that every retailer’s goal is to make the largest profit. Therefore, the assortment should be quantitatively attractive in the view of profit it can provide to the retailer.

Despite the two desirable features mentioned above, many retailers determine their products assortment base on rules of thumb such as “order more products which
generate high sales, while order less of the ones that don’t sell much”. Many studies discussed earlier have tried to construct models which lead to optimal profit. However, the high cost of the estimation of parameters and the construction of experiments has made regular use of these models difficult. Thus, we try to provide some methods which are straightforward to apply in practice and then measure their likely impact on the assortment’s performance.

Section 3.2 discusses the dataset used in the analysis. The data set is obtained from a large retail store chain, and contains a large amount of information. In section 3.3, analysis process is introduced step by step. In section 3.4, the results of the analysis for a subset of the data set are presented. Finally, in section 3.5 the conclusions are presented.

### 3.2 Data Description

The data set was obtained from a large retail chain. It contains transactions data from over a dozen stores for over a year. The fields available for every transaction are: reporting fiscal date, order id, sales type, sales breakout id, SKU id, sales location, net sales unit and net sales of that transaction.

Reporting fiscal date data is a set of numbers with 7 digits, where the first four represent the year, next two numbers indicate fiscal week and last number represent day of the week. There are 20 kinds of sales types and 8 kinds of sales breakout types; sales type represents different selling types such as regular gross sales, mail-ins, returns, etc., while breakout type indicate the method to decide the net sales, i.e. regular sales,
different promotion sales, and clearance. The transactions were for over 14,000 SKUs, that were distinguished into 6 divisions and 46 departments.

The whole data set is very large and unorganized (A brief pictorial representation of the dataset is presented at Appendix A); therefore we carried some steps to simplify the data. We focused on one store at a time; within a range of timeline and certain specific sales type/break down id. Below are the steps used to narrow down the data:

1. Store location
2. Fiscal day (Timeline of sales data)
3. Sales type/ Break down
4. Division/Department

We considered data from individual stores. The objective of selecting data of a particular store was the ease of implementation. Moreover it is not clear whether the recommendations based on the data from all stores will be useful for individual stores. The objective of analyzing data from specific time window was to eliminate product churn; product churn refers to the continuous process of replacing SKUs. This may lead to spurious sale from multiple products whereas the sale should be attributed to the same “item” that was represented by different SKUs at different point of time. The focus on the study was only on the items sold at the regular price and at normal discount prices in the physical store. Orders made online were not included. For a part of analysis, we removed a category of furniture for the analysis, since the firm was planning to stop carrying furniture in the smaller stores. The data analysis was performed using MS Access.
3.3 Analysis Process

After selecting the data for analysis, a series of steps were implement for an in-depth analysis; basically we wanted to find out the changes between selecting assortment by two criteria, revenue and volume, in different percentages. We monitored the following performance matrices for various assortments considered: percentage for incomplete basket, revenue for both complete and incomplete basket, and revenue base on customer preference. The following step-by-step process outlines the process followed. (For detail SQL queries please see Appendix B):

1. Determine the basket count in the dataset
   *This provides the basic information for further calculations, such as the percentage of the complete/incomplete basket.

2. Determine the count of SKUs in the dataset
   *This provides the basic information for further calculating, such as percentage of SKUs lost in each basket.

3. Choose top X percent SKUs base on volume/revenue (or others)

4. Select (1 – X)%1 SKUs

5. Calculate the number of SKUs missing in each basket, focusing on baskets that have missing SKU by joining total SKU data and bottom SKU data

6. Calculate the total SKU number through the basket that have missing SKU by joining order id of original total data and bottom SKU data
   *Steps 5 and 6 are useful for calculating the percentage of average loosing SKU number in each basket that is affected, i.e. which has lost SKUs.
7. Calculate complete basket total revenue by joining order id of total data and reverse missing SKU basket data

8. Calculate incomplete basket total revenue by joining order id of total data and missing SKU basket data

9. Calculate incomplete basket lost sale by joining SKU of total data and bottom SKU

*Using step 8 and 9, we can calculate the revenue for basket that has lost SKU.

10. Find out incomplete basket with less than Y percent SKU missing by setting constraint through percentage of missing SKU and total SKU

11. Calculate the original revenue of basket with less than Y percent SKU missing by joining order id of total revenue of incomplete X percent SKU basket data (from step 8) and incomplete X percent SKU data with constraint of Y percent

12. Calculate the lost sale of basket with less than Y percent SKU missing by joining order id of lost sale of incomplete X percent SKU basket data (from step 9) and incomplete X percent SKU data with constraint Y percent

*Using steps 11 and 12, we can calculate the revenue of X percent incomplete revenue with less than Y percent SKU missing.

The aforementioned steps provide a rich set of information for analyzing various assortments. In next section a full analysis on real dataset using the Steps 1-12 is discussed.
3.4 Empirical Study

The large retailer chain store we cooperated with wanted to shrink the size of the original stores, from 20000 square feet to 5000 square feet, which is a 75% decrease on space. Our objective was to suggest a smaller set of SKUs that the firm could carry in the smaller stores. We focused on the following characteristics shown in Table 3-1.

Table 3-1. Criterion

<table>
<thead>
<tr>
<th>Criteria for choosing assortment</th>
<th>Revenue/Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data time line</td>
<td>2 weeks/1 month/2 months</td>
</tr>
<tr>
<td>Breakpoints of top percentage</td>
<td>20%/25%/30%</td>
</tr>
</tbody>
</table>

Revenue and volume were chosen as the criteria for choosing assortment, where the revenue stands for net sales and volume represents the total number of units sold. Three different points of sale data sales were considered, 2 weeks, 1 month and 2 months to determine which one is more appropriate for analysis. If the results are not different, that implies that most of the SKUs are fast moving and we can use data from a short time-frame. The percentage of breakpoints is base on the potential of store size.

3.4.1 Data Observation

For a specific large store, we found the total basket counts and SKUs for different point of sales data, shown as below:

Table 3-2. Point of Sale Data

<table>
<thead>
<tr>
<th></th>
<th>2 weeks</th>
<th>1 month</th>
<th>2 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Baskets</td>
<td>4,237</td>
<td>7,609</td>
<td>15,319</td>
</tr>
<tr>
<td>Total SKUs</td>
<td>3,251</td>
<td>4,330</td>
<td>6,055</td>
</tr>
</tbody>
</table>
From Table 3-2, we can see that the number of SKUs increase with the duration of the time-frame; which means 2 weeks data includes only the fast-moving SKUs but does not register a large number of slow moving SKUs that show up on the longer period time data; the shorter time point may consist more noise in sale estimation. Therefore, 2 month horizon data will provide the best balance of avoiding product churn and providing reliable results. Naturally, the longer the time frame we choose, the more robust results will be obtained. However, a larger data set will need more time for analysis. With this tradeoff in mind, we decided to focus on the data from 2 months.

### 3.4.2 Analysis Result

After obtaining the data for 2 month, we followed the steps mentioned in section 3.3 and acquired two sets of results shown in following tables:

<table>
<thead>
<tr>
<th></th>
<th>20%</th>
<th>25%</th>
<th>30%</th>
</tr>
</thead>
<tbody>
<tr>
<td># of baskets missing SKUs</td>
<td>6,261.00</td>
<td>5,509.00</td>
<td>4,533.00</td>
</tr>
<tr>
<td>% of baskets</td>
<td>40.90%</td>
<td>36.00%</td>
<td>30.00%</td>
</tr>
<tr>
<td>Average for each basket SKUs lost</td>
<td>74.80%</td>
<td>71.70%</td>
<td>68.60%</td>
</tr>
<tr>
<td>$ for complete basket</td>
<td>144,609.36</td>
<td>165,902.50</td>
<td>196,585.90</td>
</tr>
<tr>
<td>Original total $ for incomplete basket</td>
<td>276,751.24</td>
<td>255,458.10</td>
<td>224,774.70</td>
</tr>
<tr>
<td>$ lost for incomplete basket</td>
<td>211,332.06</td>
<td>185,509.17</td>
<td>155,600.62</td>
</tr>
<tr>
<td>$ for incomplete basket</td>
<td>65,419.18</td>
<td>69,948.93</td>
<td>69,174.08</td>
</tr>
<tr>
<td>$ for 50% incomplete basket</td>
<td>33,810.49</td>
<td>39,491.10</td>
<td>45,323.27</td>
</tr>
<tr>
<td>$ for complete + 50% incomplete basket</td>
<td>178,419.85</td>
<td>205,393.60</td>
<td>241,909.17</td>
</tr>
<tr>
<td>$ for 25% incomplete basket</td>
<td>5,779.93</td>
<td>8,031.10</td>
<td>12,913.08</td>
</tr>
<tr>
<td>$ for complete + 25% incomplete basket</td>
<td>150,389.29</td>
<td>173,933.60</td>
<td>209,498.98</td>
</tr>
</tbody>
</table>
Table 3-4. Results base on Revenue

<table>
<thead>
<tr>
<th></th>
<th>Top</th>
<th>20%</th>
<th>25%</th>
<th>30%</th>
</tr>
</thead>
<tbody>
<tr>
<td># of baskets missing SKUs</td>
<td>8,222.00</td>
<td>7,418.00</td>
<td>6,708.00</td>
<td></td>
</tr>
<tr>
<td>% of baskets</td>
<td>53.70%</td>
<td>48.40%</td>
<td>43.80%</td>
<td></td>
</tr>
<tr>
<td>Average for each basket SKUs lost</td>
<td>83.20%</td>
<td>80.10%</td>
<td>76.90%</td>
<td></td>
</tr>
<tr>
<td>$ for complete basket</td>
<td>190,324.45</td>
<td>210,709.87</td>
<td>228,837.76</td>
<td></td>
</tr>
<tr>
<td>Original total $ for incomplete basket</td>
<td>231,036.15</td>
<td>210,650.73</td>
<td>192,522.84</td>
<td></td>
</tr>
<tr>
<td>$ lost for incomplete basket</td>
<td>132,018.50</td>
<td>109,423.06</td>
<td>90,870.21</td>
<td></td>
</tr>
<tr>
<td>$ for incomplete basket</td>
<td>99,017.65</td>
<td>101,227.67</td>
<td>101,652.63</td>
<td></td>
</tr>
<tr>
<td>$ for 50% incomplete basket</td>
<td>33,889.74</td>
<td>40,450.41</td>
<td>47,671.15</td>
<td></td>
</tr>
<tr>
<td>$ for complete + 50% incomplete basket</td>
<td>224,214.19</td>
<td>251,160.28</td>
<td>276,508.91</td>
<td></td>
</tr>
<tr>
<td>$ for 25% incomplete basket</td>
<td>4,232.94</td>
<td>5,410.51</td>
<td>6,758.61</td>
<td></td>
</tr>
<tr>
<td>$ for complete + 25% incomplete basket</td>
<td>194,557.39</td>
<td>216,120.38</td>
<td>235,596.37</td>
<td></td>
</tr>
</tbody>
</table>

Below we gave a brief explanation of every row in the two tables:

1. # of baskets missing SKUs: number of baskets that had missing SKUs
   (incomplete basket).
2. % of baskets: percentage of incomplete baskets.
3. Average for each basket SKUs lost: average percentage of SKUs missing in the incomplete baskets.
4. $ for complete basket: total net sales from the baskets that don’t have lost SKUs.
5. Original total $ for incomplete basket: total net sales for the baskets that have lost SKUs.
6. $ lost for incomplete basket: lost revenue through the incomplete baskets.
7. $ for incomplete basket: actual revenue for the incomplete baskets.
8. $ for 50% incomplete basket: revenue for incomplete baskets that have less than 50% missing SKUs, i.e. incomplete baskets that contain over half SKUs.
9. $ for 25% incomplete basket: revenue for incomplete baskets that have less than 25% missing SKUs.
*The purpose of calculating 25% and 50% incomplete basket revenue is to evaluate the pickiness of the customers; if a customer is very picky, she will not buy anything unless everything in her basket is available in the store. We will discuss this in the next section.

3.5 Discussion

3.5.1 Basket Component

Tables 3-3 and 3-4 show that the average percentages of missing SKUs in incomplete basket are high; based on volume is around 70%, and based on revenue is around 80% (83.2%, 80.1%, 76.9%). One conjecture is that this is happening because the average basket size is small. The graph below supports this:
The histograms in Figure 3-1 show the distributions of number of SKUs in incomplete baskets. A large number of baskets contain only one or very few SKUs. This observation leads us to the conjecture that when we remove SKUs from the assortment, the fraction of incompleteness in incomplete baskets might increase substantially. This may be concern for the firm since a very large fraction of incompleteness might lead to basket abandonment and may result in a poor word-of-mouth publicity.
3.5.2 Percentage Suggestion

We chose the top 20%, 25% and 30% SKUs by revenue as the potential assortments and tried to find out which assortment might be the best one to use. To make this decision, we first considered the ratio between complete basket and incomplete basket.

![Figure 3-2. Ratio of Guaranteed Revenue](image)

The graph shows that both assortment by volume and revenue will have the ratio over 1 by choosing 30% (or higher) SKUs; on the other words, only if we choose more than 30% SKUs, the guaranteed revenue will greater than non guaranteed revenue, which implies that keeping at least 30% SKUs in the small stores will be the best choice. It is very obvious that the more percentage we obtain, the larger ratio we will get; but the purpose of this research is to find out the best percentage in order to implement into smaller store, therefore we won’t consider the higher percentage number.
Next we look at the relation between revenue per SKU and the percentage range we set; it is intuitive that the revenue per SKU will decrease by the increasing percentage, but we want to find out if there exists a diminishing-returns point. Such a point will be a natural candidate for a good assortment.

As we can see in the Figure 3-3 above, assortments with top 30% either by revenue or volume reach the point of diminishing return; which basically tell us that the revenue per SKU will stay at the same ranges with percentage over 30%. This also implies that the total revenue won’t explosively increase with the total volume of SKUs.

Through the two basic analysis presented above, we gain a quick conclusion; keep at least 30% SKUs in the smaller stores. This percentage will basically guarantee that the store has at least half of the revenue that it has currently. Given that the real estate cost will decrease by 70% when the firm switches from 20,000 square feet store to 5,000 square feet store, this appears to be an attractive option.
3.5.3 Revenue vs. Volume

_Pickiness_

Some customers may be picky about the completeness of their shopping experience, and may not be willing to buy incomplete baskets; pickiness can be defined as the minimum completeness of a basket at which customer will not abandon the basket. For example, 100% of pickiness may refer to a customer will purchase only if all items in the basket is available in the store; 0% on the other hand represent customer buys even if only 1 present in basket is available in the store.

![Figure 3-4. Pickiness for Baskets](image)

We now investigate the impact of customer pickiness on the revenue for the firm from the smaller store. We restrict our attention on 30% assortments by volume and revenue to analyze the impact of pickiness of customers (graph shown as above). The
graph shows that revenue based assortment is better at all pickiness level, which indicates that choosing assortment based on revenue may be a better idea.

**Incomplete Baskets**

The relation between incomplete basket and the percentage is a critical factor that needs to be considered, since too many incomplete baskets might result in an eventual loss of customer base in the long run. The figure below presents the variation in the number of incomplete baskets with the assortment percentage; it is obvious that the smaller percentage the higher percentage incomplete basket will be, but the graph will still provide some useful information.

![Figure 3-5. Percentage of Incomplete Basket](image-url)
The Figure 3-5 shows that revenue based baskets lead to more incomplete baskets but they earn more revenue (according to the result table in section 3.4.2). The best explanation for this result is based on an observation made earlier in section 3.5.1; most baskets contain very low number of SKUs. Therefore, when we choose the assortment based on revenue, more baskets will appear to be incomplete as compared to the incomplete baskets in a volume-based assortment. However, the revenue-based assortment will provide more revenue which matches the retailer’s main objective of maximizing revenue.

**Division Impact**

Finally we investigate the impact of carrying the suggested 30% assortment by volume or by revenue on the size of various product-divisions for the firm. Table 3-6 shows the total number of SKUs in each division at the larger store if top 30% SKUs by volume or revenue were stored.

<table>
<thead>
<tr>
<th>Division Name</th>
<th>Total SKUs</th>
<th>Assortment by Volume</th>
<th>% Retained</th>
<th>Assortment by Revenue</th>
<th>% Retained</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSD &amp; MISCELLANEOUS</td>
<td>601</td>
<td>50</td>
<td>8.32%</td>
<td>6</td>
<td>1.00%</td>
</tr>
<tr>
<td>COPY AND PRINT DEPOT</td>
<td>42</td>
<td>10</td>
<td>23.81%</td>
<td>17</td>
<td>40.48%</td>
</tr>
<tr>
<td>FURNITURE</td>
<td>2,356</td>
<td>275</td>
<td>11.67%</td>
<td>457</td>
<td>19.40%</td>
</tr>
<tr>
<td>PERIPHERALS</td>
<td>2,102</td>
<td>348</td>
<td>16.56%</td>
<td>574</td>
<td>27.31%</td>
</tr>
<tr>
<td>SUPPLIES</td>
<td>7,820</td>
<td>2727</td>
<td>34.87%</td>
<td>1,911</td>
<td>24.44%</td>
</tr>
<tr>
<td>TECHNOLOGY &amp; SERVICE</td>
<td>1,192</td>
<td>385</td>
<td>32.30%</td>
<td>654</td>
<td>54.87%</td>
</tr>
</tbody>
</table>

It can be clearly observed that the assortment based on revenue generated a more balanced portfolio of divisions (except for an extreme value). This suggests that the
assortment based on revenue will provide customer a shopping environment with a higher variety level. Another key difference in the assortments stands out. The assortment based on volume is more supplies oriented but the assortment based on revenue is more technology oriented. This is not completely surprising since office supplies such as pen and paper tend to be sold in bulk but bring in less revenue than technology products that are sold in fewer numbers but still bring in large revenue due to a high per-unit selling price.

To sum up, choosing top 30% assortment base on revenue is the best alternative we suggest; not only because it provides higher guaranteed revenue but also a higher level of variety in divisions. The assortment may be fine-tuned by comparing the SKUs in the by-volume-assortment and in the by-revenue-assortment, and adding SKUs that are highly desirable in one assortment but not in the other one. By choosing the top assortment by revenue, we can expect the smaller store to have a large number of incomplete baskets which may cause the decreasing number of customer with definite guaranteed revenue. Due to the limitation of resource and time, this is the most comparative result we have been able to obtain so far. We identify future directions for more comprehensive analysis in the next chapter.
Chapter 4

Conclusion and Future Research

4.1 Summary

We can divide this research into two main parts; literature review and empirical case analysis. In the literature review part, we introduce assortment planning and two main theories that can be implemented into assortment planning, ABC analysis and association rules. ABC analysis is a very common analysis technique which classifies each item on its importance on a certain criterion, and we can choose the top x percentage of items base on this criterion. There are many models of ABC analysis around this core idea, with both single criterion and multiple criteria models. Association rules is another technique that can be implement for assortment planning, where it can provide the relation between the potential for simultaneous sale of items using point of sale data. However, association rules required more detail sale information and more technical resources to provide optimal assortments.

In the second part of the research, we presented an in-depth analysis of a real case data set from a large retail chain. For confidentiality reasons, we have scaled the results presented in the thesis by a constant factor. The firm proposed to build smaller stores to save real estate costs; however this needs downsizing to a smaller assortment in order to fit in the space limitation.
4.2 Conclusion

Our analysis of the data shows that the firm should keep in stock at least the top 30% SKUs by revenue in the smaller store. This is a starting point for the assortment; managers at individual stores should add more SKUs to make the portfolio balanced based on their experiences. This assortment has many desirable properties. First, it captures at least 50% of the revenue in complete baskets. Second, when compared to volume based assortment, this assortment provided a higher guaranteed revenue. Third, it provides a more balance portfolio of various product-divisions; in other words, the store will provide a larger variety of products. The chief limitation of revenue based assortment is that it leads to higher fractions of incomplete baskets as compared to volume based assortments. The manager’s discretionary additions of SKUs will help reduce this fraction.

4.3 Future Research

This research has focused on the revenue based and volume based assortments separately. The future research can combine these two criteria together, and investigate the advantages of the multi-criteria analysis model.

Association rules can be used to determine the relation between items in the same basket, and this information can be exploited for assortment planning. The data set we had was extremely large and prohibited us from developing association rules within the short time available to us. We believe there is scope for combining ABC analysis and
association rules to improve assortment planning; this also should be explored in the future research.
References


## Appendix A

### Data Set Introduction

#### Table A-1. Simple Data Set Example

<table>
<thead>
<tr>
<th>Product_Code</th>
<th>Order_Id</th>
<th>Order_Quantity</th>
<th>Order_Shipment_ML</th>
<th>Date_Id</th>
<th>Net_Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC5208</td>
<td>2012</td>
<td>22</td>
<td>0</td>
<td>1</td>
<td>210142</td>
</tr>
<tr>
<td></td>
<td>2013</td>
<td>13</td>
<td>8</td>
<td>1</td>
<td>736500</td>
</tr>
<tr>
<td></td>
<td>2014</td>
<td>13</td>
<td>8</td>
<td>1</td>
<td>737384</td>
</tr>
<tr>
<td></td>
<td>2015</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td>828001</td>
</tr>
<tr>
<td></td>
<td>2016</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td>909785</td>
</tr>
<tr>
<td></td>
<td>2017</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td>632861</td>
</tr>
<tr>
<td></td>
<td>2018</td>
<td>12</td>
<td>8</td>
<td>1</td>
<td>101177</td>
</tr>
<tr>
<td></td>
<td>2019</td>
<td>12</td>
<td>8</td>
<td>1</td>
<td>547715</td>
</tr>
<tr>
<td></td>
<td>2020</td>
<td>14</td>
<td>8</td>
<td>1</td>
<td>154366</td>
</tr>
<tr>
<td></td>
<td>2021</td>
<td>14</td>
<td>8</td>
<td>1</td>
<td>776390</td>
</tr>
<tr>
<td></td>
<td>2022</td>
<td>14</td>
<td>8</td>
<td>1</td>
<td>136192</td>
</tr>
<tr>
<td></td>
<td>2023</td>
<td>15</td>
<td>8</td>
<td>1</td>
<td>576060</td>
</tr>
<tr>
<td></td>
<td>2024</td>
<td>15</td>
<td>8</td>
<td>1</td>
<td>407071</td>
</tr>
<tr>
<td></td>
<td>2025</td>
<td>15</td>
<td>8</td>
<td>1</td>
<td>576060</td>
</tr>
<tr>
<td></td>
<td>2026</td>
<td>15</td>
<td>8</td>
<td>1</td>
<td>576060</td>
</tr>
<tr>
<td></td>
<td>2027</td>
<td>12</td>
<td>8</td>
<td>1</td>
<td>800870</td>
</tr>
<tr>
<td></td>
<td>2028</td>
<td>12</td>
<td>8</td>
<td>1</td>
<td>570085</td>
</tr>
<tr>
<td></td>
<td>2029</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td>916097</td>
</tr>
<tr>
<td></td>
<td>2030</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td>547275</td>
</tr>
<tr>
<td></td>
<td>2031</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td>547275</td>
</tr>
<tr>
<td></td>
<td>2032</td>
<td>12</td>
<td>8</td>
<td>2</td>
<td>629920</td>
</tr>
<tr>
<td></td>
<td>2033</td>
<td>8</td>
<td>8</td>
<td>1</td>
<td>334147</td>
</tr>
<tr>
<td></td>
<td>2034</td>
<td>8</td>
<td>8</td>
<td>1</td>
<td>328001</td>
</tr>
<tr>
<td></td>
<td>2035</td>
<td>14</td>
<td>8</td>
<td>1</td>
<td>503870</td>
</tr>
<tr>
<td></td>
<td>2036</td>
<td>15</td>
<td>8</td>
<td>1</td>
<td>219881</td>
</tr>
<tr>
<td></td>
<td>2037</td>
<td>15</td>
<td>8</td>
<td>1</td>
<td>257771</td>
</tr>
</tbody>
</table>
### Table A-2. Definition of Dataset

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reporting_Fiscal_Data_Id</td>
<td>Fiscal date transaction made</td>
</tr>
<tr>
<td>Order_Hour</td>
<td>At which hour the order was made</td>
</tr>
<tr>
<td>Order_Id</td>
<td>Transaction number</td>
</tr>
<tr>
<td>Rpt_Sales_Breakout_Id</td>
<td>Sales breakout id</td>
</tr>
<tr>
<td>Sales_Type_Code</td>
<td>Sales type code</td>
</tr>
<tr>
<td>SKU</td>
<td>SKU number</td>
</tr>
<tr>
<td>Sales_Location_Id</td>
<td>Sales location id</td>
</tr>
<tr>
<td>Net_Sales_Units</td>
<td>Net sales unit</td>
</tr>
<tr>
<td>Net_Sales</td>
<td>Net sales (revenue)</td>
</tr>
<tr>
<td>IMU</td>
<td>Marginal profit.</td>
</tr>
</tbody>
</table>
Appendix B

SQL queries used in Access

*POSDATA is the original data set
*Each step will generate the table that is highlight in yellow

1.

```sql
SELECT Order_Id, COUNT(POSDATA.SKU) AS Total_SKU_Count INTO BASKET_ITEM_COUNTS
FROM POSDATA
GROUP BY Order_Id;
```

2.

```sql
SELECT POSDATA.SKU INTO SKU_COUNTS
FROM POSDATA
GROUP BY SKU;
```

3.

```sql
SELECT TOP X PERCENT POSDATA.SKU INTO TOP_X_SKUs
FROM POSDATA
GROUP BY POSDATA.SKU
ORDER BY Sum(POSDATA.Net_Sales) DESC;
```

4.

```sql
SELECT DISTINCT POSDATA.SKU INTO BOTTOM_1_MINUS_X_SKUs
FROM POSDATA LEFT JOIN TOP_X_SKUs ON POSDATA.SKU=[TOP_X_SKUs].SKU
WHERE [TOP_X_SKUs].SKU Is Null;
```
5.

```sql
SELECT Order_Id, COUNT(POSDATA.SKU) AS Missing_SKU_Count INTO BASKET_MISSING_X_ITEM_COUNTS FROM POSDATA INNER JOIN BOTTOM_1_MINUS_X_SKUs ON POSDATA.SKU=[BOTTOM_1_MINUS_X_SKUs].SKU GROUP BY Order_Id;
```

6.

```sql
SELECT POSDATA.Order_Id, COUNT(SKU) AS Total_SKU INTO BASKET_MISSING_X_ITEM_TOTAL_COUNTS FROM POSDATA INNER JOIN BASKET_MISSING_X_ITEM_COUNTS ON POSDATA.Order_Id=[BASKET_MISSING_X_ITEM_COUNTS].Order_Id GROUP BY POSDATA.Order_Id;
```

7.

```sql
SELECT POSDATA.Order_Id, POSDATA.Net_Sales INTO COMPLETE_X_BASKET_REVENUE FROM POSDATA LEFT JOIN BASKET_MISSING_X_ITEM_COUNTS ON POSDATA.Order_Id=[BASKET_MISSING_X_ITEM_COUNTS].Order_Id WHERE [BASKET_MISSING_X_ITEM_COUNTS].Order_Id Is Null;
```

8.

```sql
SELECT POSDATA.Order_Id, POSDATA.Net_Sales INTO INCOMPLETE_X_BASKET_ORIGINAL_REVENUE FROM POSDATA INNER JOIN BASKET_MISSING_X_ITEM_COUNTS ON POSDATA.Order_Id=[BASKET_MISSING_X_ITEM_COUNTS].Order_Id;
```

9.

```sql
SELECT POSDATA.Order_Id, POSDATA.Net_Sales INTO INCOMPLETE_X_BASKET_REVENUE FROM POSDATA INNER JOIN BOTTOM_1_MINUS_X_SKUs ON POSDATA.SKU=[BOTTOM_1_MINUS_X_SKUs].SKU;
```
10.

```
SELECT BASKET_MISSING_X_ITEM_COUNTS.Order_Id INTO X_YINCOMPLETE_BASKET
FROM BASKET_MISSING_X_ITEM_COUNTS INNER JOIN
BASKET_MISSING_X_ITEM_TOTAL_COUNTS ON
BASKET_MISSING_X_ITEM_COUNTS.Order_Id=BASKET_MISSING_X_ITEM_TOTAL_COUNTS.Order_Id
WHERE
(BASKET_MISSING_X_ITEM_COUNTS.Missing_SKU_Count/BASKET_MISSING_X_ITEM_TOTAL_COUNTS.Total_SKU)<0.5;
```