Segmentation Analytics with SAS® Viya®

An Approach to Clustering and Visualization

Randall S. Collica

The correct bibliographic citation for this manual is as follows: Collica, Randall S. 2021. *Segmentation Analytics with SAS® Viya®: An Approach to Clustering and Visualization*. Cary, NC: SAS Institute Inc.

Segmentation Analytics with SAS® Viya®: An Approach to Clustering and Visualization

ISBN 978-1-952363-06-1 (Hardcover)
ISBN 978-1-951684-05-1 (Paperback)
ISBN 978-1-951684-06-8 (Web PDF)
ISBN 978-1-951684-07-5 (EPUB)
ISBN 978-1-951684-08-2 (Kindle)

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

July 2021

Contents

About This Book

What Does This Book Cover?

The main purpose of this book is to demonstrate how to accomplish clustering and segmentation in SAS® Viya® through the use of several graphical user interfaces (SAS® Visual Statistics, Model Studio, SAS® Visual Analytics, and a coding interface of SAS® Studio). This is a “how to” book with practical, real data examples in each chapter. This book also covers visualizations of clustering including a relatively new technique called t-SNE that can be used to better understand the underlying structure of the data before and after clustering.

While this book does *not* cover the theory of clustering and segmentation, there are a good number of available references at the end of each chapter where you can find appropriate additional materials that are relevant to each chapter.

Is This Book for You?

This book is designed for analysts, data miners, and data scientists who need to use the all in-memory platform of SAS Viya for the purposes of clustering and segmentation. I have not attempted to write a comprehensive book on segmentation, but this book is focused on the analytics and methods of SAS Viya actions, procedures, and even the SAS 9.4 procedures that can be used in conjunction with SAS Viya through SAS Studio. If you are a novice with clustering and segmentation, a chapter in the Appendix is designed for the basic understanding of how clustering works with distances of observations. If you already know clustering well, then this book will aid you in how to accomplish clustering and segmentation analytics using SAS Viya.

What Are the Prerequisites for This Book?

While it will be helpful if you are already familiar with clustering concepts, it isn’t mandatory as Appendix 2 provides a discussion on the basic concepts of distance and how that is used in many clustering algorithms. An understanding of basic analytics concepts such as linear regression, elementary probability, statistics, and machine learning will be helpful.

What Should You Know about the Examples?

Each example is real data that has been anonymized and is available for your use to understand how to apply each of the methods described in this book. The results that you obtain while executing each of the examples might differ from what is printed in this manuscript. The results might differ because of sampling proportions, distributed computing environments versus single computing, or other general options settings that can affect the results. One way to ensure that the results keep consistent in a distributed computing environment is to use the SAS Viya CAS node= restrictions that limit the processing to a single node. This will likely keep the results consistent; however, it may defeat the purpose of larger data sets where distributed computing is typically used to reduce execution times. Please see the "NWORKERS= Session Option" section in the *SAS® Cloud Analytic Services: User's Guide* concerning the CAS number of workers options available at: https://go.documentation.sas.com/?docsetId=casref&docsetTarget=n1v9x1q6ll09ypn0zknvo54rk9ya.htm&docsetVersion=v_001&locale=en.

Software Used to Develop the Book's Content

The software used in this book is SAS Viya version 3.5 and 4.0. (4.0 is the containerized deployment of SAS Viya and is labeled as 2020.1.x versions.)

Example Code and Data

This book includes tutorials for you to follow to gain hands-on experience with SAS. Each folder contains the data sets, code, and macros for each chapter except for Chapters 1 and 2. These chapters provide an introduction to clustering and segmentation with SAS and give you a flavor of the possible use cases. Chapters 3 through 6 all have examples, data, code where applicable, and any macros used in the code. You can access the example code and data for this book by linking to its author page at https://support.sas.com/collica.

We Want to Hear from You

SAS Press books are written *by* SAS Users *for* SAS Users. We welcome your participation in their development and your feedback on SAS Press books that you are using. Please visit sas.com/books to do the following:

- Sign up to review a book
- Recommend a topic
- Request information on how to become a SAS Press author
- Provide feedback on a book

About The Author

Randy Collica is a Principal Pre-Sales Solutions Architect at SAS supporting the retail, communications, consumer, and media industries. His research interests include segmentation, clustering, ensemble models, missing data and imputation, Bayesian techniques, and text mining for use in business and customer intelligence. He has authored and coauthored eleven articles and two books. He holds a US patent titled "System and Method of Combining Segmentation Data." He received a BS in electronic engineering from Northern Arizona University.

Learn more about this author by visiting his author page at http://support.sas.com/collica. There you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more.

Chapter 1: Introduction to Segmentation Using SAS

Introduction

In my previous book on segmentation and clustering (Collica 2017), I began with a description of what segmentation is, and I referred to Customer Relationship Management (CRM) in the context of marketing. CRM is the innate ability to understand as much about the customer in order to better serve the customer's needs, preferences, buying experience, and so on. SAS® Enterprise Miner™ and some SAS code were used throughout the book and now SAS offers a new addition to its analytics platform, namely SAS® Viya®.

SAS Viya is a high-performance computing architecture in which all computations are multi-threaded and contained entirely in-memory as each data set is contained in memory. In addition, the use of Massively Parallel Processing (MPP) processing is also possible as is using hardware acceleration using Graphics Processing Unit (GPU) devices. This allows much faster processing, and therefore algorithms that might have taken too long to be of practical use are now doable in a fraction of the time. This new architecture and the introduction of micro-services enables new algorithms and processes for data science while also allowing an open environment that supports SAS, Python, R, Lua, and Java. This book will describe how to use SAS Viya for clustering, segmentation, and visualization, while at the same time, introducing some new methods as well. My goal is to empower you with the knowledge of using SAS Viya for solving analytical business problems relating to clustering, segmentation, and visualization.

In the 1930s Chamberlin[2] had laid the foundation for segmentation by prioritizing the consumer over the producer by highlighting the significance of aligning products with the needs and desires of customers. Today really isn't too different except now the availability of different channels in which messaging and offers take place is much wider than ever before such as SMS text messages, telemarketing, direct mail, email, digital advertising, and personalized web offers.

Retailers can now install beacons in their brick-and-mortar stores. If a customer decides to accept the beacon service, the retailer using the beacon tracking system can know what aisle the customer is walking down in real time. For example, if the customer is in the shoe aisle, the retailer with pre-determined offers can then prompt an offer to the customer in real time! At check-out, that offer can be scanned at the register, and the marketing process can be a one-on-one communication by producer and consumer in real time at many locations simultaneously. This depiction is known as location-based marketing, and many organizations are aspiring to obtain this level of real-time marketing offer capability. But just the offer is only the start; the organization must have the infrastructure operating in order to capture the response, purchase, and record the transaction and preference directly into the customer data record. This is also true for the campaign management system's table schema for keeping accurate records and aligning specific offers and at times arbitrate among multiple offers.

The methods in this book will aid you in improving marketing effectiveness. By understanding who your customers are and what underlying homogeneous customer groups might be present within your data, you will have a greater insight about your customers. Armed with this new insight, you can more effectively develop offers to customers in specific segments that should have a higher propensity to purchase when such estimates are surfaced and acted upon. I am sometimes surprised at how often I find that customer segmentation efforts in industry are frequently lacking in the variety of data on their customer, patient, or account records. For example, if you only have web traffic data on your customers, then you are missing out on their demographics or firmographics if your customers are businesses. Likewise, if you have purchase transaction data and demographics or firmographics but you lack insights as to your customer's attitudinal information toward your organization, then again, you'll be lacking a complete 360-degree view of your customers.

Let's take an example of customer data where both behavioral and attitudinal attribute information exists on the same data set. A cluster segmentation of behavioral data is shown in Figure 1.1, which has seven clusters and some profile information. Figure 1.1 does not contain the attitudinal attributes in this model. However, in Figure 1.2 the attitudinal attribute was added with all other settings and attributes being the same from Figure 1.1.

Notice the changes to the channel_purchase attribute in the lower left quadrant (outlined in red) is vastly different in Figure 1.2 than in Figure 1.1. In this data set, a channel of "1" indicates purchase from a reseller only, a value of "2" is direct purchases only, and "3" is both direct and indirect reseller purchases. A value of "0" indicates neither and typically is return purchases. Adding just a single important attribute such as attitudinal information (values 1 to 5 as a nominal attribute) changes the cluster model significantly.

> **Key Message:** Behavioral and attitudinal data are typically much more influential in clustering for segmentation than demographic or firmographic data alone!

Figure 1.1: Cluster Segmentation without Attitudinal Data Attributes

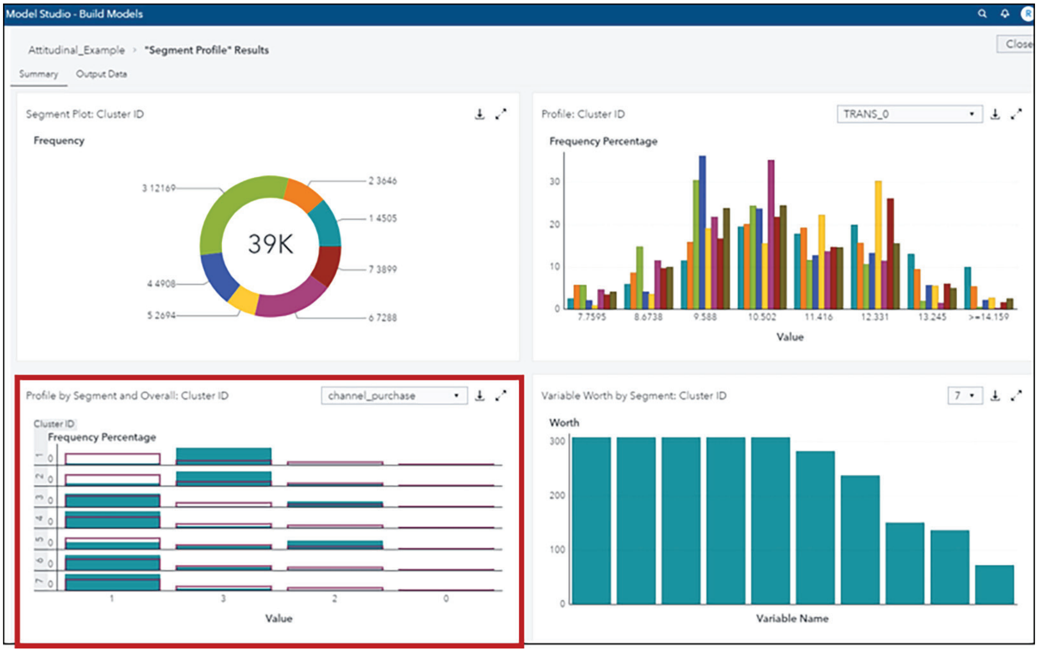
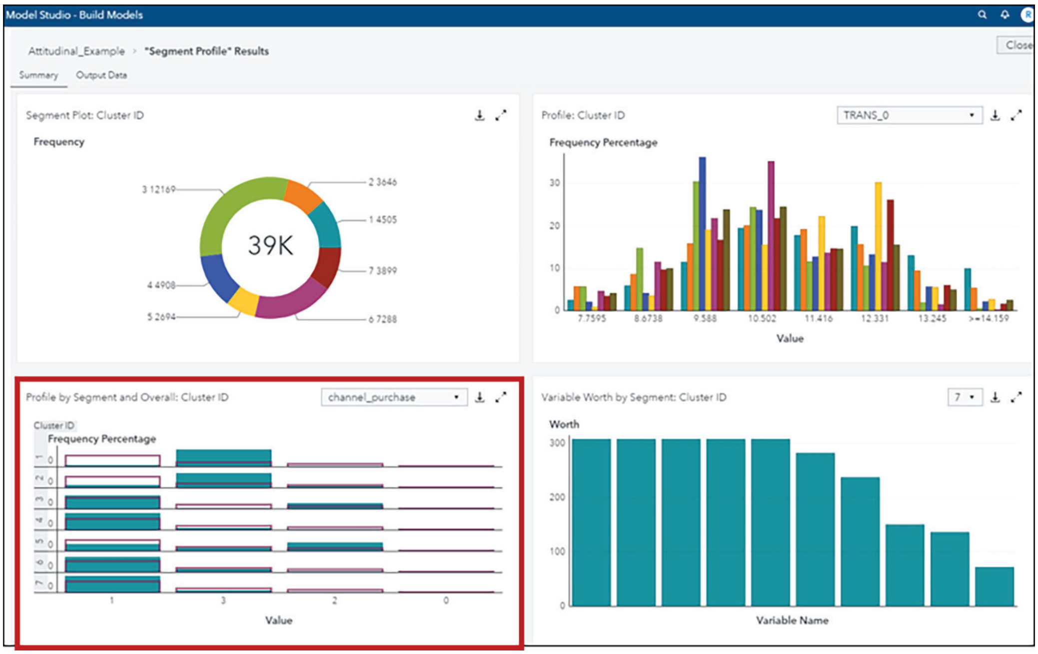


Figure 1.2: Cluster Segmentation with Attitudinal Data Attributes



Segmentation: Art, Science, or Both

In keeping with the science of segmentation, may I digress a bit to review the other side of segmentation, namely the artistic or "left-brain" side? When planning for a segmentation project, whether it be for marketing, exploration, sales, customer, or patron similarities, or even patient likeness, the key to developing a segmentation that meets your organization's needs will undoubtedly involve your domain expertise and expertise from others as well. A keen understanding of the objectives of the segmentation project will greatly help in what approach you should undertake, what data you will need or perhaps even acquire, and the timeline needed to complete the project. In most segmentation projects, a typical outcome objective will be when the derived segments are relatively homogeneous within each segment and yet distinctly different from other segments. The following examples help to explain how both domain experience and a data driven method can be combined for a desired outcome.

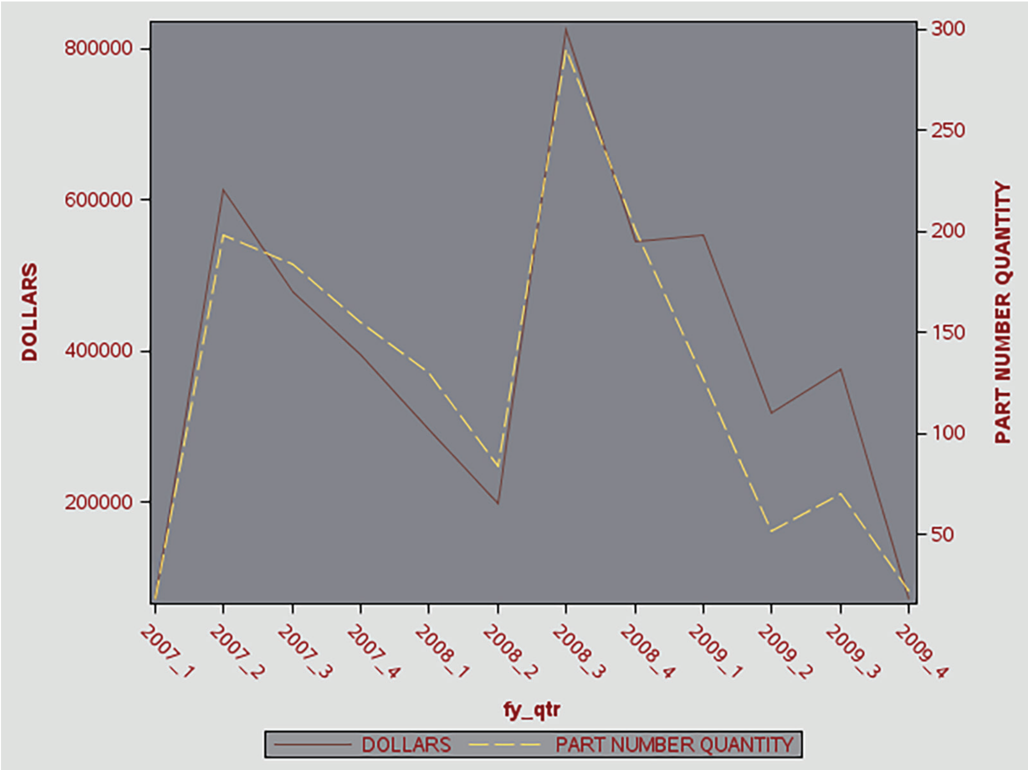
This section briefly describes two use cases in segmentation that involved segmentation methods that were for strategic usage rather than a specific tactical purpose. One use case was for a product marketing analysis and the other for a sales segmentation focus. These two examples are from genuine business situations where product and sales executives needed guidance and direction that was based on data. In both cases, however, the businesses had some preconceived ideas about how to approach and recommend certain attributes based on their historical knowledge of the available data. Also, in both cases, the historical data had changed, and therefore, the data-driven approach that I took changed the course of the analytics that were originally based on those preconceived ideas. This is where data and business domain expertise must work together to achieve the desired goals and objectives needed by each of these organizational groups – hence both art and science.

Use Case 1: Strategic Product Segmentation

Late in 2009, a senior product line manager at my previous employer came to me and asked for my assistance in a project that required re-setting the course for this product line. At that time in the high-tech industry, changes in computer technology were migrating to more commodity-based hardware, and data centers around the US and the world were taking note and making changes in their purchasing decisions and plans for how data centers operate. Because of this apparent change in direction, the senior management needed some data-driven insights to assist them in choosing a course of action to take in the future. They desired to have data-driven assistance in guiding where the newer purchases were going in the market; however, they also had some preconceived ideas as to what the purchasing behavior might look like for certain high-volume purchase customers.

The senior manager had told me that there were four accounts that he believed I should focus on as my "target" purchasing trend behavior and on which I should fashion the strategic analytics accordingly. I immediately proceeded to extract the historical data for these four customer accounts to review the trends. However, the trend that the product line manager thought was a "good" trend turned out not quite what he expected. In fact, the trend was the opposite of

Figure 1.3: Four Account Trends Averaged



his initial expectation – it was heading downward not upward! Figure 1.3 shows the average of the four account trends (names omitted for proprietary reasons). As you can observe, these trends were not exactly what was expected. Although the volume in revenues and quantities was reasonably high, the trend certainly wasn't what we desired to set the pace for in the future. At this point of the analysis, I had a serious decision to make as to how to approach the strategic segmentation project.

The basic question I wrestled with was should I use the data as originally planned as the "target" trend with which to measure all other accounts and therefore segment the customer base, or should I use something different? I thought about this dilemma for a while and concluded that this would be a great learning moment for data-driven methods. I decided to keep the "target" trend. However, I would look for other trends that were better suited for the goals and objectives of the product line. The company chose these four accounts because they thought the trend was heading upward based on their experience with these strategic accounts. Indeed, the trend downward wasn't going to be helpful for this analysis. I set out to find accounts that exhibited a more positive trend even though their level of revenue and quantity wasn't quite near the level of these accounts. My rationale was that the *right* trend was more important to this project than the current volume of the four accounts. For this project I used SAS® Enterprise Guide® and SAS Enterprise Miner.

The steps I used in the data prep and analytics are outlined in Table 1.1 below.

Table 1.1: Steps Used to Prepare Data for Strategic Product Segmentation

Step	Process Step Description	Brief Rationale
1	Selected product lines needed in the products data table.	Query only product lines needed in this analysis.
2	Using the product line codes from previous query, I then queried the product purchase transactions between 2007 and the end of 2009.	Queried product transactions purchased on dates of interest.
3	I then aggregated the total revenues and quantities by purchase channel (direct or indirect) by customer account.	Aggregated customer transactions.
4	I then queried the transactions for the four accounts that the product manager desired. I also selected a few targeted accounts that had the product line transactions with increasing purchases over time.	Better product purchase transactions for a target group to measure against.
5	I labeled the accounts with the target transactions as well as the four originally selected by the product manager.	Account labeling accomplished by using a SAS format for unique account IDs.
6	With customer purchase transactions labeled by account ID in step 5, I ran a transaction similarity measurement[3] that measured the distance between the target accounts and all others.	Measuring the average distance from all product transactions to desired target group.
7	Merged the average transaction similarity metric along with the labeled accounts with the customer firmographic data and market share estimates using predictive models previously developed.	Final merge of data sets.
8	Performed clustering on the final data set and noted variables that affected the segmentation and profiled the segments.	Profiled cluster segmentation.
9	Generated charts and diagrams and gave presentation to senior product line managers with recommendations for next steps and direction.	Final analytic insights and recommendations.

The process in step 7 measures a target transaction against all other transactions in both the magnitude and time unit dimensions (Chamberlin 1933). The procedure gives a distance metric that when clustered together gives transactions of similar shape both in magnitude and in time sequence. This in effect performed transaction clustering along with customer clustering based on other firmographic information such as market share, industry group, and so on. All the data prep and queries were done using SAS code in SAS Enterprise Guide. The similarity metric used the SIMILARITY procedure (a SAS/ETS® procedure – SAS Institute 2014), and the cluster segmentation was done in SAS Enterprise Miner.

The clustering results produced a segmentation of 10 segments, of which three were considered high priority for future growth potential for the product directions. The variables that made an impact in this segmentation were first chosen as potential candidates based on my analyst's knowledge of the data, which I was very familiar with, and the estimated market share potential for which I previously developed analytic predictive models (Collica 2010). Table 1.2 and Figure 1.4 show a table and chart, respectively, of the variables that influenced the 10 clusters and their profiles.

Key Message: Data-driven analysis that included the domain rules-based segment showed that the data-driven content was much more significant at driving the customer behavior segmentation than the rules-based segmentation.

Table 1.2: Key Variables Affecting Strategic Cluster Segmentation

Variable	Relative Importance
Company Segment	1.000
Industry Group	0.909
Log (ISS TAM)	0.780
RFM	0.778
Log (Similarity)	0.673
Log (Yrs Purch)	0.528
Orig Segment	0.321

Figure 1.4: Key Variables Affecting Strategic Cluster Segmentation

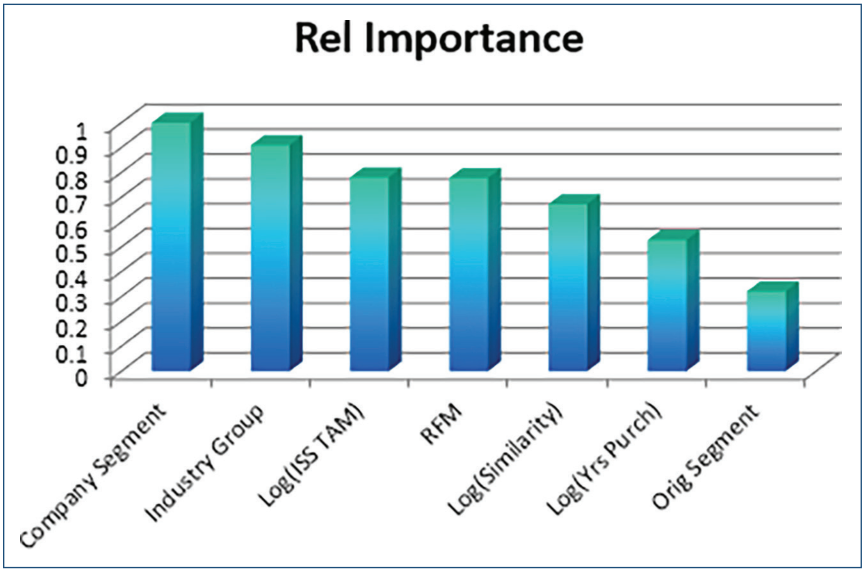
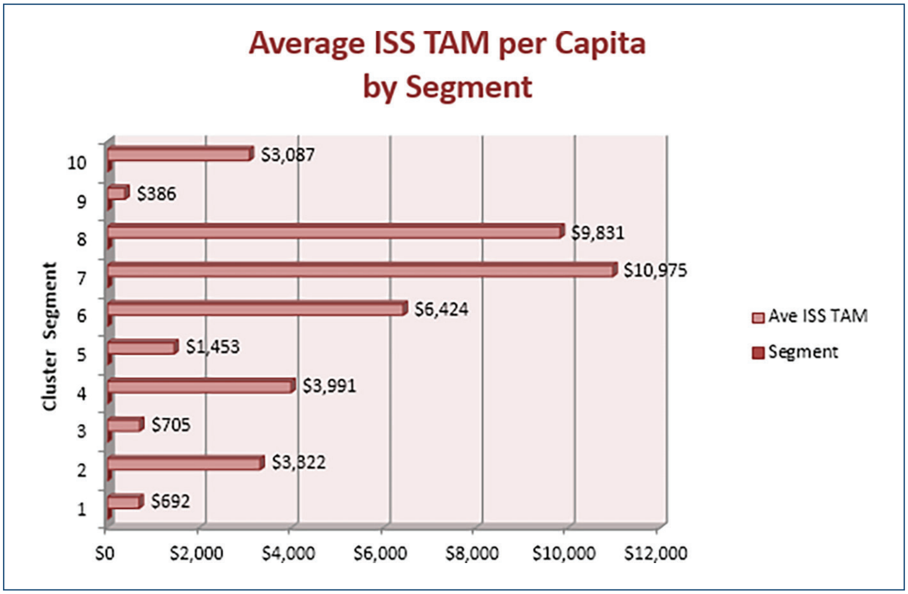


Figure 1.4 shows that Company Segment is the largest contributor, but the variable called Original Segment based on the manager's historical knowledge was the least important. The Company Segment variable was derived from the company's account reference file, which indicated whether the account was a corporate account, an enterprise or commercial account, or small and medium business (SMB) account, based on the definition that was applied as a business rule and was complied with by all countries. While the key message stated above is dominant for this analysis, it is not always true for all segmentation analytics. Of primary importance in this was the industry that the company was in and the estimated market called total addressable market (TAM). Recency, Frequency, Monetary value (RFM) is a typical segmentation that measures recency, frequency, and monetary value of purchases that is fully described in Chapter 4 of (Collica 2017). Once you have the estimated TAM at the customer account level, the estimates and other key attributes can be aggregated easily. The similarity metric also played an important factor. Figure 1.5 shows the value of the TAM per capita (total TAM divided by number of customers per segment) for the ISS product line group. Cluster segments 6–8 are considered the most valuable. Figure 1.5 shows the aggregate general relationship between the average similarity metric, ISS TAM, and the average number of years of customer purchase. This shows that there were non-linear relationships among these variables. The cluster segmentation did take this into account in the final clustering analysis. So the segments that show the highest average similarity and number of years purchase also had the highest average TAM.

The product-line management team was impressed by advanced analytical methods such as clustering of transactions into similar groups and using market share estimates. They also were aligned on the insights from the cluster segments that were of high-value and allowed more

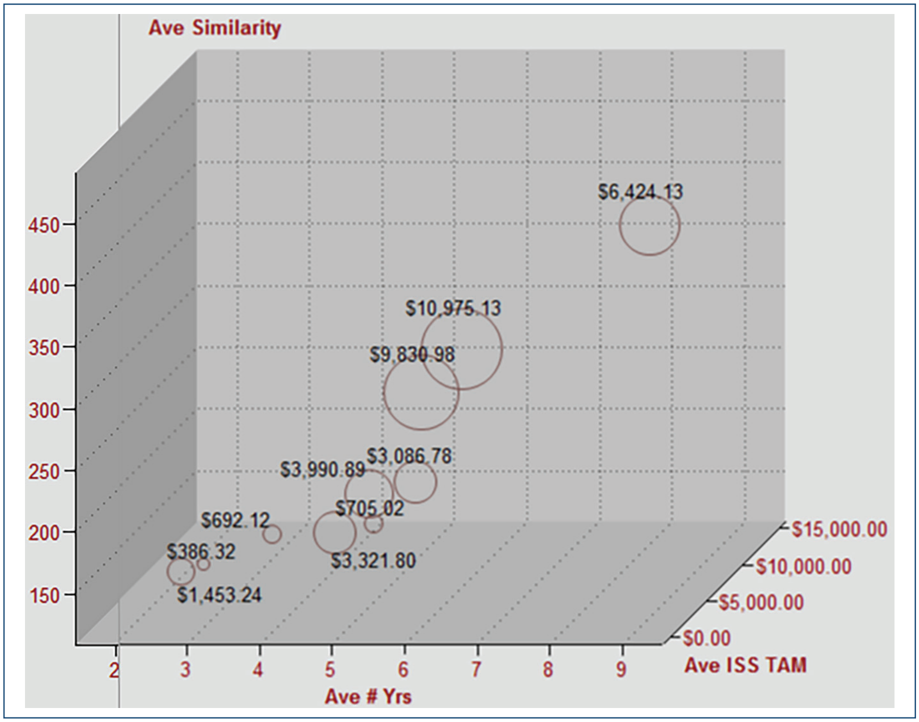
Figure 1.5: Total Addressable Market Estimates by Strategic Cluster Segment



strategic plans to be developed. Again, what differentiated this from a tactical segmentation was the fact that I used semi-supervised techniques. A supervised technique requires a training data set where the correct answers are in the data, whereas an unsupervised technique has no labeled training data. Although clustering is an unsupervised method, I gave it some general direction by using variables such as a metric that measured how close the target transaction shape was to all other transaction shapes as well as an estimate of the market share. This market estimate was developed by me almost 11 years earlier using SAS Enterprise Miner with a two-stage model (Collica 2010). The product line management using the three high-value clusters and other more mediocre clusters was able to make definite strategic plans for customer accounts and target them much better according to industry and size, and so on. The elements that you place into the segmentation will strongly influence whether the segmentation can be used strategically or for tactical purposes. Market indicators such as share of wallet (SOW) help direct the segmentation for a strategic purpose rather than a tactical one.

Key Message: The variables that you select for a segmentation will strongly influence whether the segmentation can be used strategically or tactically.

Figure 1.6: Plot Showing General Relationship between Ave ISS TAM, Similarity, and Yrs. Purchase



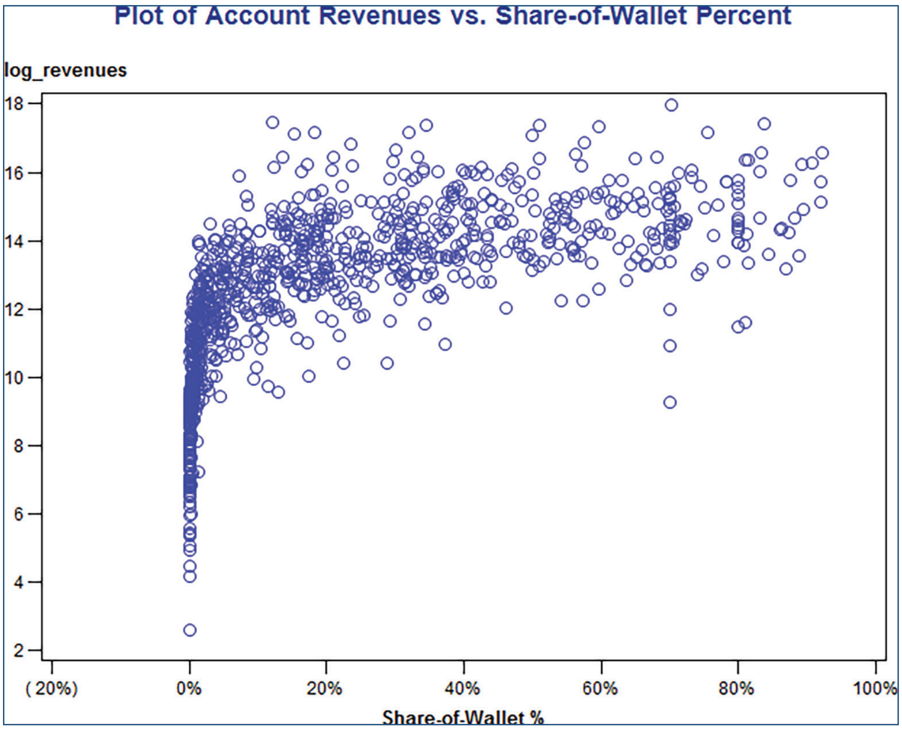
Use Case 2: Strategic Sales Segmentation

The key to this method is like Use Case 1 in that it requires an estimate of the revenue (or profit) share-of-wallet estimate in each account. When you can estimate the amount of total spending that the account can spend relating to the products and/or services that your organization can supply, the revenues that you generate from that account divided by the total estimated spending becomes the estimated share of wallet (SOW). This can be a very powerful metric if the estimate is reasonably accurate. Sales and marketing can use these estimates for strategic planning in areas such as the following:

- quota setting
- account prioritization
- product and messaging approaches
- segmenting customers by their SOW estimates

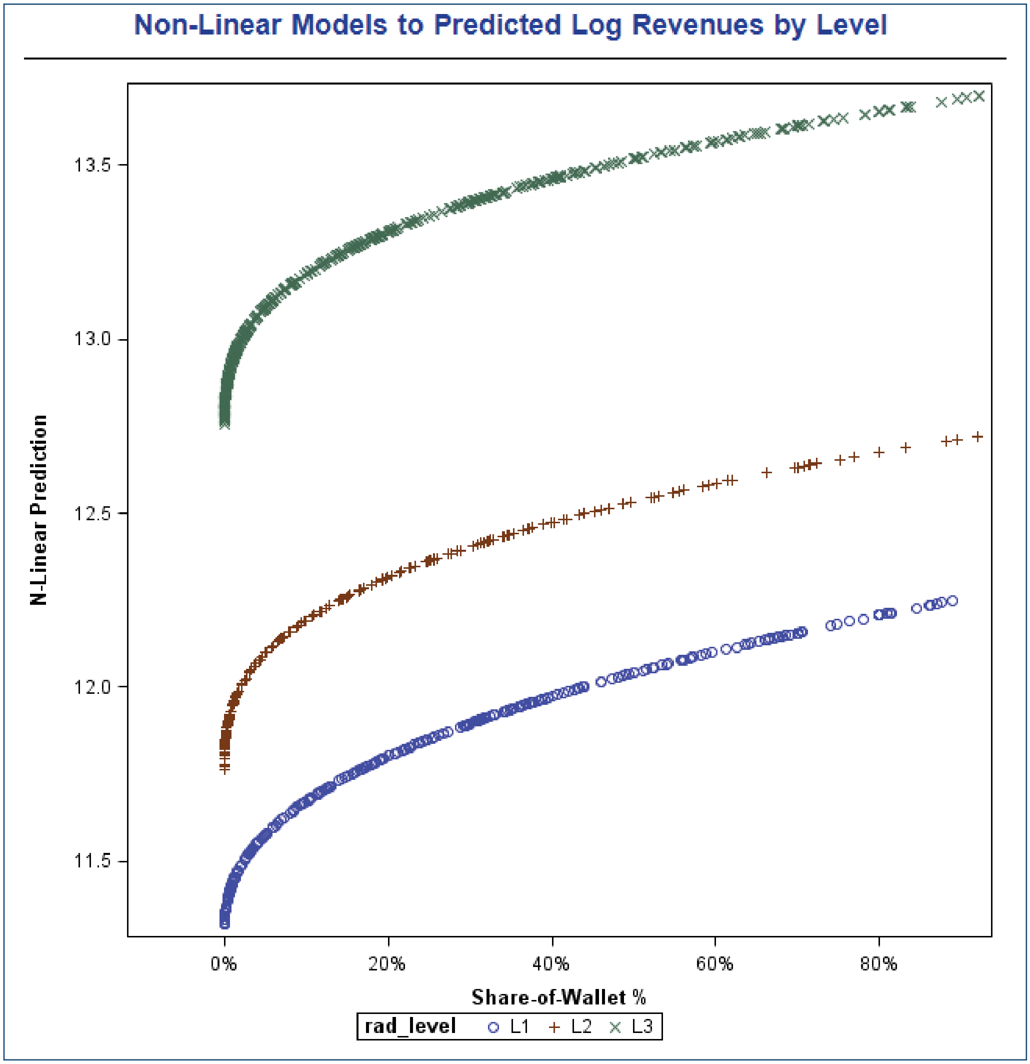
The business needed to segment the accounts so that their planning and goal setting process could be enhanced using the data-driven methodology to understand the potential versus their actual spending for IT products and services. The models that were developed in this SAS Global Forum paper were used in this particular business unit, and the planning process needed this segmentation and estimates at the account level, not just the total spending by industry like you can obtain from syndicated reports (Collica 2010). The plot in Figure 1.7 shows about 1,000

Figure 1.7: Account Revenues and SOW Percent



accounts with the logarithm of their latest year's revenues versus the estimated SOW percent. What this plot shows is the non-linear relationship between these metrics. However, it is very difficult to see any other pattern. One of the objectives that the sales management needed to accomplish by segmenting the sales accounts was to assist them in their sales planning and operations for the upcoming year. Historical segmentation methods relied heavily on a corporate segmentation based on historical revenues rather than current behavioral characteristics. Segmenting the accounts in Figure 1.7 with historical corporate segmentation methods didn't produce any usable analysis, so the segmentation proposed is to carve out some delineation of revenue and the SOW so that they could start the planning process. The segmentation in this case carved the SOW percent into three groups with splits at 10% and 40%. For the revenues, the two splits were $125,000 and $350,000. This produced nine segments. Now, when you fit a nonlinear model to the three different SOW levels, you obtain the plot shown in Figure 1.8.

Figure 1.8: Non-Linear Model Fit for Three Segment Levels



In Figure 1.8, you can more clearly observe the relationship for the different SOW segments. Each of the nine segments had a sales strategy that fit their SOW level and the total revenues that could be expected. Using other models in tactical campaigns, you could offer certain cross-sell/upsell products or services that better fit according to the expectations. The levels of R, A, and D represent Retain, Acquisition, and Develop. This strategy was used successfully in several of the major business units. This segmentation was also combined with other segmentations around the organization to improve the efficacy of marketing and could potentially be used in market research.

In the following chapters, you will learn how to use SAS Viya and SAS 9.4 together as one platform along with open source so that data scientists and analysts alike can use, collaborate, and deploy models in their organizations to gain insight and improve their businesses.

References

Chamberlin Edward. 1933. The Theory of Monopolistic Competition. Cambridge, Massachusetts: Harvard University Press.

Collica, Randy. 2010. “Estimating Potential IT Demand from Top to Bottom,” Proceedings of the SAS Global Forum 2010 Conference. Cary, NC: SAS Institute Inc. Paper no. 372.

Collica, Randall. S. 2017. Customer Segmentation and Clustering Using SAS® Enterprise Miner™, 3rd ed., Cary, NC.: SAS Institute Inc.

SAS Institute Inc. 2014. Chapter 24, “The Similarity Procedure,” SAS/ETS 13.2 User’s Guide, Cary, NC.