#### SAS733-2017

# Using Segmentation to Build More Powerful Models with SAS<sup>®</sup> Visual Analytics

Darius Baer, Ph.D., SAS Institute Inc.

## ABSTRACT

What will your customer do next? Customers behave differently; they are not all average. Segmenting your customers into different groups enables you to build more powerful and meaningful predictive models. You can use SAS® Visual Analytics to instantaneously visualize and build your segments identified by a decision tree or cluster analysis with respect to customer attributes. Then you can save the cluster/segment membership and use that as a separate predictor or as a group variable for building stratified predictive models. Dividing your customer population into segments is useful because what drives one group of people to exhibit a behavior may be quite different than what drives another group. By analyzing the segments separately, you are able to reduce the overall error variance or noise in the models. As a result, you improve the overall performance of the predictive models. This paper covers the building and use of segmentation in predictive models and demonstrates how SAS Visual Analytics, with its point-and-click functionality and in-memory capability, can be used for an easy and comprehensive understanding of your customers, as well as predicting what they are likely to do next.

## INTRODUCTION

Customers provide you with the ability to be profitable. How can you communicate with them and provide offers that maximize the relevance of your products and services for each customer? Because customers have different needs and wants, they have different reasons or drivers for buying your product or interacting with your company. Using analytic tools, you can efficiently and effectively group your customers according to their needs and wants. Then, you can communicate and market to them based on their different purchase and interaction behaviors as well as factors such as demographics. The better you are able to understand your customers' buying habits and lifestyle preferences, the more accurate your predictions of future buying behaviors will be.

The drivers for a behavior such as responding to an offer or buying a product can be very different from one customer to another. Some customers are motivated by price, others by convenience, and still others by customer service. These customers may be grouped by segmenting them according to customer attributes.

You need to treat your customers differently according to the segment to which they belong. Providing personalized offers and communication to each segment enables you to perform these functions:

- Increase share of customer wallet.
- Increase share of market by being more relevant in your business space.
- Increase relevant customer interactions while reducing irrelevant customer interactions.
- Improve customer satisfaction and thereby generate higher profits.

Because we are in the era of ever-expanding data and hyper-personalization, segmentation is now more effective than ever. As customers are different, you should prioritize segmentation as the foundation of customer insights across your company. There are a variety of segmentation strategies to stimulate customer preferences and to increase customer satisfaction. These are applicable to both current and prospective customers.

Marketers have used segmentation to provide a better relationship with their customers for a very long time. In fact, segmentation has been used since the introduction of customer relationship management (CRM) and database marketing. Within the context of Customer Segmentation Intelligence (CSI), there are a variety of attributes, including consumer demographics, geography, behavior, psychographics, events and cultural backgrounds. Over time, segmentation has proven its value across every imaginable industry, and brands continue to use this strategy throughout the stages of the customer journey:

- Acquisition
- Upsell/cross-sell
- Retention
- Win back

You can segment your customers and treat them based on their cluster attributes or you can proceed to building predictive models for each cluster based on the attributes most likely to motivate them to a desired behavior.

As a marketer, you also want to be able to predict or assess the likelihood that a customer will buy a product or service or engage in any other behavior related to your business. As stated before, customers have different reasons or drivers for responding to an offer or engaging in a behavior. You should not assume that all customers will behave the same. In fact, you are likely to get better predictions if you segment your customers prior to modeling their behavior, notwithstanding that you may be able to derive those differences within the predictive model without segmentation.

This paper will present both approaches: predictive modeling without segmentation and predictive modeling by segments.

For marketing and communication with your customers, you can use these strategies:

- Segment to obtain insights about your different types of customers.
- Build predictive models for specific types of customer interactions/offers.
- Segment and then build models within the segments.



## **Function**

#### Figure 1: Information Delivery Enablers

Data-driven marketers now have actionable advanced analytics available to make more powerful decisions within today's complex and interconnected business environments. Technology has provided in-memory computing systems that can process data and execute models in speeds that are orders of magnitude faster than just a decade ago. The functions of these technologies can deliver easily interpretable information in both tabular and graphic format. Using the in-memory facilities offers previously unheard of performance. And, best of all, the interface affords an ease-ofuse to allow the marketer or analyst to focus on the problem or issue at hand rather than figuring out how to solve the problem or write programs. All three factors are needed to enable successful information delivery. Figure 1 illustrates the information delivery process.

SAS Visual Analytics and SAS<sup>®</sup> Visual Statistics are interactive and visual solutions

from SAS that offer technology that meets all three criteria of function, performance, and ease-of-use. The tools represent an intersection of data management, interactive visualization, and big data analytics that empower the marketer to get answers to simple and complex analytic questions almost instantaneously. Both segmentation and predictive models are processed and presented so that the analyst can iteratively get the actionable output right away.

With SAS Visual Analytics and SAS Visual Statistics, you can visually group your customers based on common attributes. Having millions of customers does not slow down the process – you still see your results in a couple of seconds. Therefore, you can run as many scenarios as you like, changing which

attributes are included at will. Then you can go further with predictive modeling, using advanced analytical methods such as decision trees or regression. All of this is done without the need to write or manage computer code. However, if you need the scoring code produced by your analysis, it is easily exported at the push of a button. You have the ability to export score code for the clustering, decision trees, and regression visualizations. Score code enables you to place new and changed customers into current clusters and obtain likelihood scores for customers to respond to the campaign that you analyzed.

Export Score Code						
Do you want to export the score code for this model?						
options VALIDMEMNAME=EXTEND VALIDVARNAME=ANY;	4					
/**** User ID: Darius Baer ****/						
/**** Model Name: Clustering ****/						
SAS Code Generated by LASR Analytic Server for Clustering Date : 06Mar2017:16:10:12						
Cluster Method : KMEANS						
Number of Clusters : 8						
Convergence Criterion : 0.00						
*/						
array y 2 [9] temporary -						
array cid 2 [8] temporary ;						

Moving forward, we will briefly discuss the why and what of segmentation. Then we will discuss the how, specifically referencing SAS Visual Analytics and SAS Visual Statistics for visually creating, understanding, and using your segments. Finally, we will demonstrate the process of using the segments

to create predictive models. Your goal is to be as relevant as possible to your customers. Using both segmentation and predictive models together moves you closer to that goal.

## THE BUSINESS CASE FOR SEGMENTATION

By interacting with your entire customer universe through the same offers and communication, you satisfy only a small portion of them, resulting in smaller profits, irrelevant interactions, and dissatisfied customers. Using segmentation, you can build and manage stronger, longer, and more profitable customer lifetime relationships. Different customers are motivated by varying attributes, treatments, and interactions. Segmentation categorizes or classifies items or subjects into identifiable groups that share similar features. Features can be based on demographics, purchase behaviors, clickstream patterns, channel interactions, and so forth. You should build your segments according to your marketing strategies. If a company does not partition their customers into segments, they might be likely to treat all customers the same.



Figure 2: Customer Segmentation Intelligence (CSI) For Increasing Profits

We can use the visual image in figure 2 to demonstrate differences among customers. Notice in the left group that only the average customers inside the center dotted circle are happy. The company is treating all customers the same, as if they were all average. However, that treatment applies only to that small set in the center dotted region. The others are sad because their needs and wants are not being addressed. If you intelligently segment your customers into groups that can be treated differently and in accordance with their needs and wants, you should be able to make almost all customers happy as shown in the middle group. And your profits should grow as your new and existing customers become more loyal advocates for your brand, as in the group on the right.

Segmentation is both an art and a science. You can use analytics and statistical clustering as your science. However, understanding your business, your customers, and the interactions between the two provides the art. In the end, decision-makers within an organization are the final arbiters. If they do not

understand the analysis, nothing will ever change. There are people who are good at creative marketing strategy, and there are people who are good at marketing analytics. Using SAS Visual Analytics and SAS Visual Statistics serves as the conduit for people to toggle between the two and serves as the translator for both sides.

## WHAT IS THE BEST SEGMENTATION?

Part of the challenge in modern marketing is finding the golden mean between the two extremes of treating all customers the same and one-to-one personalization.

There are two major ways to increase customer profit. The first is by increasing market share - you get more customers. The second is by increasing share of wallet - you get more money from each customer. In both cases, segmentation enables you to treat customers differently based on the segment that each individual belongs to.

So, is there a "right" number of segments? We segment in order to manage our marketing initiatives



spanning millions of consumers. Our intent within the discipline of CRM is to nurture relationships with each customer, optimizing the individual's experience with the brand. It is extremely challenging to conduct one million discrete campaign objectives. So we segment as a compromise between marketing to each customer individually and treating all customers the same.

### Figure 3: How Many Segments?

As shown in Figure 3, you want to have

enough segments to treat customers according to their needs and wants without the effort becoming unmanageable. Furthermore, you might decide to have multiple segmentation models based on different dimensions:

- Segmentation strategy based on purchase history
- Segmentation strategy based on digital interactions and social media
- Segmentation strategy based on demographics

Of course, it is possible and worthwhile to create a unified segmentation strategy across all dimensions, sometimes referred to as ensemble segmentation.

You can execute segmentations using business rules (non-quantitative), analytical methods (quantitative), or some combination of the two. In this paper, we define the term segmentation as a process of grouping observations or customers. Thus, analytical methods that encompass both supervised (classification) and unsupervised (clustering) approaches are defined as a subset of segmentation that uses a statistical approach to analyze the data relationships. We demonstrate both supervised (decision trees) and unsupervised (k-means clustering). Then we demonstrate predictive modeling using decision trees and logistic regression.

## HOW SHOULD YOU USE SEGMENTATION?

There are two main business reasons for segmenting your customers:

- 1. Treat and manage customers in each segment differently.
- 2. Use segments in further analysis.

By segmenting customers (or prospects), the expectation is that different groups will respond uniquely to targeted communications, incentives, or offers. Ideally, marketers will want to interact with each segment in a manner that best matches the profile characteristics of that audience. Because different attributes

can help explain why customer groups behave differently, you can also run predictive analytic models within each segment. For example, you can predict the likelihood of responding to an offer to buy a product. However, since you might want to sell products to different segments, you would be better off building propensity models for each segment. This usually increases the signal-to-noise ratio, or predictive accuracy, of each model compared to building one model for all customers.

Ultimately, segmentation, like any analytical exercise within an organization, needs to prove its value. To save your segmentation projects from failing, you must be clear on how your organization will act differently with the results. See the SAS Global Forum 2016 paper "Enhanced Segmentation Using SAS® Visual Analytics and SAS® Visual Statistics." These are the four general methods to group or segment customers that were discussed in that paper:

- 1. Business rules
- 2. Quantiles
- 3. Supervised classification decision trees
- 4. Unsupervised clustering k-means clustering

In this paper, we will use only the supervised and unsupervised methods in SAS Visual Statistics. Although you may use business rules or quantiles for segmentation, we encourage the use of the more sophisticated analytic segmentation approaches. Rather than describe the details of the segmentation types which are explained elsewhere, we will go through a scenario where you will see how segmentation and predictive modeling go hand-in hand to produce marketing-ready results.

## **EXAMPLE SCENARIO FOR SEGMENTATION AND PREDICTIVE MODELING**

Suppose you work for a large firm that provides lodging that we will call Stay-With-Us. Your marketing department would like to provide an offer to increase occupancy. A previous offer was sent out to the 1,032,734 customers for whom Stay-With-Us has both data on previous rentals and demographic data. We also know whether each customer responded to the previous offer or not.

There were 60 variables available in the dataset. After preliminary examination, and for the purposes of demonstration, we have reduced the number of variables we will use in this example. Here are the variables/columns that we will work with and a short description of each.

Column Name	Column Description
member_id	Member ID
response	Response to the campaign
avg_days_btn_stays_c	Average days between stays
days_btn_stays_c	Number of days between stays
days_since_enroll	Number of days since program enrollment
dayssincelaststay	Number of days since the last stay
leadtime	Average number of days in advance booking is made
nbr_nights_c	Number of nights
num_stays_spring	Number of stays- Spring
points_earned	Number of points earned
proployalty	Loyalty to a brand within the portfolio
stay_c	Number of stays

#### Table 1: Variables Used in the Stay-With-Us Example

There was a 22.74% response to the campaign over the 1,032,734 customers. The marketing questions include:

- 1. What types of customers do we have?
- 2. If we create a similar campaign to encourage rentals, what attributes would best differentiate responders from non-responders?

The first question can be addressed by both supervised and unsupervised segmentation while the second requires a predictive model.

## SUPERVISED CLASSIFICATION

To use supervised segmentation requires you to identify a dependent (or target) variable of interest. This is usually done by using a decision tree model in which you explore all available attributes (or independent variables) to identify which are important in describing distinct groups, which are those groups that are most pure with respect to the dependent variable.

The objectives of a decision tree analysis are:

- Segmenting the population into groups that are as homogeneous as possible.
- Producing a branched structure of rules and groups, which is the model of the data. This is
  produced by successive splits of the marketable audience as the decision tree grows.
- Ending up with groups of similar customers for segmentation.
- Creating decision trees, which is a machine-learning method that splits customer data into smaller and smaller segments, each of which are increasingly "pure" (or unique).
- Using the decision tree to choose the most important variables for the supervised segmentation task, and to identify the audience segments based on specific values of those variables that correlate to the business objective.

So, by leveraging SAS Visual Statistics, we can ask: What drives response rate by the customers of Stay-With-Us Lodging?

- Decision tree analysis creates a data-driven segmentation, which can be used by marketers to better refine their customer campaigns.
- SAS Visual Analytics also provides the ability to explain the results to nontechnical stakeholders.

In our example with Stay-With-Us Lodging, if we use a decision tree to segment and response as the target variable, we can get a result with eight groups, as seen in Figure 4.



Figure 4: Decision Tree Segmentation

Figure 5 illustrates how SAS Visual Analytics shows the total group of 1,032,734 customers. The blue bars represent those who did not respond and the green bars represent those who did respond. You can see that the majority, almost 90%, are in the group for customers with less than 6.75 total stays. This is a final decision tree leaf or a supervised segment of the population, and it has a response average of 15.35%. The leaf on the left has a response average of 73%. It continued splitting into seven end leafs or segments. A supervised result is most often used as the final marketing deliverable. An unsupervised segmentation may be used by itself or as a preliminary analysis to a predictive model.



UNSUPERVISED CLUSTERING

Figure 5: Tree Details - First Split

Unsupervised clustering is best if you do not have a target variable of interest or if you simply want to find groups of customers that are not driven by pre-specified target characteristics. Do your customers naturally fall into different groups? Using SAS Visual Analytics, you can quickly ascertain if this is the case. From a marketing perspective, this can be useful if you do not have a specific behavior you want to profile, but rather are looking for a general understanding of who your customers really are. Finding natural groupings in the data is called unsupervised segmentation.

In unsupervised segmentation, you want to reduce the variation within each group while keeping the variation between groups as large as possible. K-means clustering analysis is commonly used for unsupervised segmentation.

In k-means clustering, the "means" are the centroids for each cluster and are represented by the averages of the values for each attribute used in the k-means clustering for the members in the cluster. In Figure 6, we compute the centroid for each cluster by averaging all the *x* values of the points in the cluster to form the *x* coordinate of the centroid, and



Figure 6: Cluster Matrix Visualization in SAS Visual Statistics

averaging all the *y* values to form the centroid's *y* coordinate. Figure 6 only shows two of the attributes or dimensions. Most k-means clustering use many more dimensions. Again, the centroid is the average of the values for each attribute of each member (or customer) in the cluster.

The k in k-means is simply the number of segments that you would like to find in the data. The beginning of a k-means analysis starts with a desired (or subjective) number of clusters k. So, in Figure 6, the analyst would have specified k=8, and the k-means clustering method would return the following:

- The eight cluster centroids when the clustering algorithm had completed iterating.
- Information on which of the data points (or customers) belong to each cluster.

The k-means algorithm for finding segments is simple and effective, and therefore is worth highlighting:

- 1. The algorithm starts by creating *k* initial cluster centers, usually randomly.
- 2. As more data is analyzed, the segments corresponding to these cluster centers are formed, by determining which is the closest center to each point or observation.
- 3. For each of these clusters, its center is recalculated iteratively as more data is assessed by finding the actual centroid of the points in the cluster.
- 4. The cluster centers typically shift while the process continues. Since cluster centers will shift, the algorithm continually recalculates which points belong to each cluster. After reassignments, the cluster centers might shift again and again.
- 5. The k-means procedure keeps iterating until there is no change in the cluster memberships, and the process terminates. Every observation that is a member of a cluster is closest to that cluster center than any other cluster center.

Although there is no guarantee that one k-means run results in a good segmentation analysis, there are pre-processing techniques to help. One best practice when executing a k-means clustering suggests that the population to be clustered should be sorted in descending order by the squared distance from the global mean over all the attributes used. With this approach, the first cluster begins with the observation or customer that is the most distant from the population mean. Each additional customer added is the next farthest from the population mean, and so forth. This approach often results in clusters (or segments) that are more homogenous (or pure). Equally, this reduces the chance that there will be multiple clusters near the population center, which is disadvantageous from a marketing perspective.

Because the attributes that are used for k-means clustering often have different scales (statistical variance), the data can be standardized to so that all variables have the same influence on the creation of the clusters. SAS Visual Analytics provides standardization as a check-box option. Even with standardization, some measures may have extreme standardized values (z-scores). These outliers can distort the cluster results. To accommodate that issue, you can winsorize the standardized values by capping the limits of the standardized values at plus or minus three standard deviations. This heuristic approach has been shown to bring about more useful clusters for marketing.

As to how many clusters or segments to use, once again, the business need and capability are more important than statistical correctness. We have found that between six and ten clusters usually provide sufficient differences while respecting the ability of the business to manage the different clusters by providing different offers and communications to each cluster. If you are interested in the statistical case for deciding on the "proper" number of clusters, you can review the Wikipedia article "Determining the number of clusters in a data set".

Finally, there are times when more than one clustering solution is needed, depending on the attributes important for each clustering exercise. Again, the business should direct the analyst to determine what is needed from a marketing perspective.

Let's return to the example scenario, which leverages the SAS Visual Statistics k-means clustering. Using unsupervised segmentation or clustering without a target variable with the Stay-With-Us Lodging data, we get a somewhat more interesting result. The variables used to cluster are: stay\_c, dayssincelaststay, points\_earned, proployalty, avg\_days\_btn\_stays\_c, leadtime, days\_since\_enroll, and num\_stays\_spring. We performed some pre-processing on the data prior to bringing the data into the SAS Visual Analytics environment. It is a common practice to standardize variables used in a clustering in order to equalize the scale over the different variables.

- We use the numeric attributes listed in the variable table, with the exception of the member ID and the response variable (which is our target).
- We standardize the data outside of SAS Visual Analytics in order to cap or winsorize our standardized values at plus or minus three standard deviations.
- We choose to create eight clusters. After some testing, this seemed sufficient and not too many.
- We cluster over 1 million customers and quickly achieve the eight-cluster result illustrated in Figure 7.



Figure 7: Visual Statistics Eight-Cluster Solution

Note these items in Figure 7:

- The top graphic, Cluster Matrix, shows that there is some separation for the clusters among the five attributes that are used in the graphic. The centroids for each cluster are shown by the small circles while the ellipses indicate the majority of values represented by each cluster for the specific attribute pairs.
- The bottom graphic, Parallel Coordinates, illustrates another way of looking at the differences among the clusters. The lines for each cluster are more visible for the values that cluster members have for each attribute. Thicker lines indicate more customers for that attribute value.
- The following two graphs single out customers who are very inactive (cluster 7 illustrated in Figure 8) versus those who rent often and have higher loyalty (clusters 4 and 6, illustrated in Figure 9).



Parallel Coordinates(1,285 polylines)

SAS Visual Statistics affords the ability to derive a cluster ID variable, which allows an analyst to augment the customer data by tagging individuals with segment assignments. We can then use other visual tools to understand our clustering results. We previously highlighted clusters 7, 4, and 6. Notice in Figure 10 that the inactive customers in cluster 7 account for around 1/3 of the total and are unlikely to respond to the offer, while the active customers in clusters 4 and 6 account for a small number of the total and are quite likely to respond to the offer.



Figure 10: Frequency and Response of Clusters

## **PREDICTIVE MODELING – DECISION TREE**

After clustering, we can use the cluster ID variable in predictive models. Previously, we built a decision tree to segment our customers. We discovered that most of the customers went into one leaf or segment. Looking at Figure 11, we can build the decision tree that is enhanced by including the cluster ID variable. To the left is a view of the total tree where green leaves indicate a response level greater than 50%, while blue indicates less than 50%. Next is the top node of the whole population of 1,032,734 customers, showing that the first split in the tree (most predictive) is the cluster ID variable. To the right is the first node of 746,893 customers consisting of clusters 0, 1, 3, and 7, of which 88.3% are non-responders. To the right of that is the node with 107,107 customers consisting of clusters 4 and 6, of which 76.7% are responder. Finally, there is the node with 178,734 customers consisting of clusters 2 and 5, of which 63.7% are non-responders.



Figure 11: Decision Tree With The Segment Variable Being Most Predictive Variable

You can see in Figure 12 that the response variable is predicted better by use of the cluster ID variable. Without the cluster ID is on the left. With the cluster ID is on the right. The lift indicates how much better the model predicts the target variable response than if you randomly selected the target, which would be according to the frequency of .2274. The top 5% of the scored customers are almost 4 times more likely to be responders. In other words, they would average an 85% response rate rather than 23%.



Figure 12: Lift Values for Decision Trees Without Cluster ID (left) Versus With Cluster ID (right)

## **PREDICTIVE MODELING – LOGISTIC REGRESSION**

As we suggested in the introduction, what drives one group of customers to behave in a desired way may be different than what drives other groups. Some segments are motivated by low prices, others by service, and still others by convenience. By building a logistic regression within each cluster, we can easily determine whether there are different factors that are important for the prediction of the response variable. The variables used to build the logistic regression are stay\_c, dayssincelaststay, points\_earned, proployalty, avg\_days\_btn\_stays\_c, days\_btn\_stays\_c, leadtime, days\_since\_enroll, and num\_stays\_spring. If we build a logistic regression over all the customers to predict the response, we obtain a result set that has all nine predictors as very significant.

You can see in Figure 13 that the r-square is .33 and the lift at 5% is almost 4 while the lift at 20% is just under 3. So, this is a good model. The top 3 predictors are the following:

- 1. Loyalty to a brand within the portfolio
- 2. Number of days between stays
- 3. Number of points earned



Dimensions	iteration History Con	vergence	FILSI	ausucs	ype	milest	Parameter	csum >
Parameter	Parameter				E[	z Val 🔻	]	Pr >  z
Number of days	between stays	0.0	01736	0.000	02	87.1583	9	<0.0001
Number of points	s earned	0.0	00018	2.30E	-7	80.0329	5	<0.0001
Number of stays	Number of stays				35	60.7570	1	<0.0001
Average number	Average number of days in advance booki				38	39.105	9	<0.0001
Number of days	Number of days since program enrollment				-6	33.4029	1	<0.0001
Number of stays	Number of stays- Spring			0.0025	13	30.3553	3	<0.0001
Number of nights	Number of nights			0.0003	87	9.02781	5	<0.0001
Average days be	-0.	00177	0.000	03	-59.599	3	<0.0001	
Loyalty to a brane	d within the portfolio	-1.	17439	0.0074	54	-157.543	3	<0.0001
Intercept		-1.	55451	0.0048	32	-321.72	3	<0.0001

Figure 13: Logistic Regression over all Customers

Loyalty has a negative relationship to responding, indicating that more brand loyal customers are less likely to respond. Conversely, the points earned and stays have a positive relationship, indicating that customers with more points and stays have a higher likelihood of responding. All of these make intuitive sense, because we might expect loyal brand customers to only want to stay at their brand rather than wanting to stay at another brand with the Stay-With-Us Lodging. On the other hand, customers who have earned lots of points and stayed with us a lot are continually incented to interact and would probably be more likely to be interested in the offer.

Now, we are ready to perform the logistic regression within each of the segments or clusters in order to ascertain whether the clusters may have different drivers for responding to the offer. You can see from Figure 14 that the R-square for the individual clusters is lower than the .3314 R-square over all the customers. This is because the variance from the individual variables that contributed to the separation of the clusters is not available for the regressions within each cluster.

Nonetheless, we can look at each cluster analysis and observe that the most important predictor variables are different among some of the clusters. Table 2 shows each segment with its response rate, frequency, and intercept as well as the top three predictors for the logistic regression in that segment. response C (event=1) Max-rescaled R-Square 0.1795 Observations Used: 58,084



Figure 14: Logistic Regression R-Square by Cluster

Segment	Response	Frequency	Intercept	Best Predictor	Second Predictor	Third Predictor
0	7.7%	182,297	-0.52	Number of days since the last stay – NEG	Number of stays – POS	Number of points earned – POS
1	11.2%	76,487	-1.66	Number of days since the last stay – NEG	Average days between stays – NEG	Number of days since program enrollment – POS
2	36.6%	120,650	-1.07	Number of points earned – POS	Number of days since the last stay – NEG	Number of days since program enrollment – POS
3	6.5%	175,724	-1.29	Number of days since program enrollment – NEG	Loyalty to a brand within the portfolio – NEG	Average days between stays – POS
4	6.7%	59,551	+0.66	Number of days since the last stay – NEG	Number of points earned – POS	Loyalty to a brand within the portfolio – NEG
5	35.8%	58,084	-0.82	Number of points earned – POS	Number of stays – POS	Number of days since the last stay – NEG
6	88.9%	47,556	+1.78	Number of points earned – POS	Number of days since the last stay – NEG	Loyalty to a brand within the portfolio – NEG
7	17.2%	312,385	-1.53	Number of days since program enrollment – NEG	Number of stays – POS	Avg # of days in advance booking is made – POS

Table 2: Segment	Attributes and	Predictors	from l	Logistic	Regression
------------------	----------------	------------	--------	----------	------------

Table 3 presents a more detailed perspective of the values for the top three predictors for specific variables. This table lists the most important predictors for the regression among the various clusters and the attribute averages within each response group for each cluster. You can see that clusters 4 and 6 are both the most active clusters as well as the most responsive to the campaign. Cluster 7 appears to be customers who have just become members because their number of stays is almost zero as well as the number of days since the last stay. Cluster 7 is made up of customers who have stayed very little, but are very loyal to only one brand within the portfolio. Following Tables 3 and 4, we will discuss these clusters in more detail through the use of heat maps.

Clus Num Subto	Resp Group tal: 0	Freq 182,297	Pct Group Resp 7.7%	Num of stays	Points earned 2,946	Days since last stay 80.3	Avg days between stays 19.9	Avg days in advance booking is made 8.7	Brand Loyalty 0.99
	0	168,248	0.0%	1.3	2,710	82.0	18.1	8.8	0.99
0	1	14,049	100.0%	2.3	5,772	59.0	42.1	8.2	0.98
Subto	tal: 1	76,487	11.2%	1.1	2506	143.0	703.1	11.3	0.97
4	0	67,891	0.0%	1.1	2332	146.5	714.6	11.0	0.97
1	1	8,596	100.0%	1.2	3885	115.8	612.5	13.4	0.95
Subto	tal: 2	120,650	36.6%	4.1	9,443	67.7	106.3	13.9	0.46
2	0	76,539	0.0%	3.8	7,658	71.9	114.1	13.1	0.47
2	1	44,111	100.0%	4.8	12,540	60.4	92.8	15.2	0.44
Subto	tal: 3	175,724	6.5%	1.4	2,889	276.2	47.7	9.7	0.94
3	0	164,302	0.0%	1.3	2,679	276.8	44.6	9.4	0.95
5	1	11,422	100.0%	2.1	5,904	267.3	92.1	14.9	0.82
Subto	tal: 4	59,551	67.0%	12.3	36,446	38.9	33.1	9.8	0.44
Δ	0	19,657	0.0%	11.5	30,674	50.4	35.0	7.9	0.51
- T	1	39,894	100.0%	12.7	39,290	33.3	32.2	10.8	0.41
Subto	tal: 5	58,084	35.8%	3.8	9,489	84.4	111.6	9.9	0.46
5	0	37,263	0.0%	3.0	6,284	94.8	124.9	9.2	0.49
	1	20,821	100.0%	5.4	15,225	65.6	87.8	11.1	0.39
Subto	tal: 6	47,556	88.9%	29.1	113,188	17.7	13.7	9.0	0.31
6	0	5,264	0.0%	26.9	83,806	26.3	14.7	7.2	0.39
0	1	42,292	100.0%	29.4	116,845	16.6	13.6	9.2	0.30
Subto	tal: 7	312,385	17.2%	0.0	35	0.1	0.2	5.0	0.00
7	0	258,745	0.0%	0.0	26	0.1	0.2	5.0	0.00
'	1	53,640	100.0%	0.1	79	0.2	0.4	5.0	0.01

 Table 3: Predictor Values by Cluster and Response Group

Comparing the attributes from Table 3, we are now able to create a description for each of the eight segments for our Stay-With-Us Lodging customers which is displayed in Table 4.

		Percent	
Cluster		Response	
Number	Frequency	in Group	Description
0	182,297	7.7%	Infrequent, but brand loyal
1	76,487	11.2%	Not recent, very infrequent, brand loyal
2	120,650	36.6%	Reasonably recent, Not frequent
3	175,724	6.5%	Not recent at all, infrequent, brand loyal
4	59,551	67.0%	Lots of stays, frequent, recent
5	58,084	35.8%	Infrequent, but recent
6	47,556	88.9%	Most stays, very frequent, very recent
7	312,385	17.2%	New customer

 Table 4: Stay-With-Us Cluster Important Attributes

It is only by segmenting our population and then running a predictive model on each group that we are able to observe these differences. The Stay-With-Us marketing department should treat customers differently based on which attributes are most important in each segment's predictive model. As well, the marketing department should consider providing different offers and communication to each segment.

New customers (cluster 7) need to be motivated to stay one more time, and, hopefully, soon. The offer might be price or possibly, a service that will make the customer feel special. For our brand loyal customers (clusters 0, 1, and 3) who have not stayed very often, we might offer them a discount on a different brand. For our very loyal and frequent customers (clusters 4 and 6), we might provide a special service or double points. For our intermediate customers (clusters 2 and 5), we might give them a special one year membership in a loyalty class above their current grade.

You may notice that there is a recency, frequency, and monetary aspect to our customers which is often referred to as RFM. By looking at Table 3, you can categorize the eight clusters as having different RFM values. Then, you can communicate with them appropriately.

## **BUBBLE PLOTS FOR IMPORTANT VARIABLES FROM REGRESSION**

In this section, we look at a bubble plot in Figure 15 to observe the campaign response with respect to "number of days since last stay" and "number of points earned". We can see that Clusters 4 and 6 which have the most points earned are also the most likely to respond to this campaign. In fact, the lower the number of points, the less likely the members of a cluster are to respond to the campaign. The Stay-With-Us marketing team should consider developing campaigns for the lower response clusters to get a higher response rate. Cluster 7 which has no points and zero days since last stay must be on their first stay. The motivation for them is simply to get them to stay with us the next time they need lodging.



Number of days since the last stay by Number of points earned sized by response grouped by 8 Cluster ID

Figure15: Bubble Plot of Clusters by Number of Days since Last Stay & Number of Points Earned

## HEAT MAPS FOR IMPORTANT PREDICTORS FOR SPECIFIC SEGMENTS.

In this section, we look at heat maps to understand better the important predictors for specific segments. This visual presentation improves the interpretation of the segmentation and predictive modeling results for the marketing analysts.



Figure16: Heat Maps of Clusters 4 and 6

In clusters 4 and 6 displayed in Figure 16, a small number of customers have as many as 1.4 million points. Because the number is so small, we filtered the data to 700,000 for the variable, "number of points earned." For the top performing clusters (clusters 4 and 6), the most important predictors of response are "number of days since the last stay" and "number of points earned." Looking at the heat maps in Figure 16, you can see that customers in those segments who have stayed most recently and are high point earners are the most likely to respond. You can also see that most of the customers in those segments have a lower number of points and have a last stay that is more recent. A marketer who wants to get low responders from these segments to respond might provide incentives to stay in the near future or provide more points for staying.





On the other hand, we can look at the largest segment (number 7), displayed in Figure 17, in which, of the 312,385 customers, 309,811 have zero stays and booked an average of 6 days in advance. We can look at the other 2,574 customers, but that seems too small a percentage of this cluster to bother. Nonetheless, we can still see a pattern from the predictive model. The top three predictors include "number of days since program enrollment", "number of stays", and "avg # of days in advance booking is made". Because we have really no influence on the "number of days since program enrollment", let's look at a heat map of "number of stays" and "avg # of days in advance booking is made". As both of those increase, so does the response rate. Perhaps we can incent this group to book sooner and more often. Maybe this group is price sensitive. A further analysis might help us determine whether providing a discount for booking far in advance might be a good motivator. It also might increase the number of stays, or the loyalty of this group. As the group is, for the most part, brand new, we should consider that with any of our offers or communications.





Lastly, let's look at segment 0 which has a very low response rate of 7.7%. These customers do not seem to rent from Stay-With-Us Lodging very much at all. In fact, they average 1.4 stays and the left heat map indicate the majority have had either one or two stays. The response rate of this group is higher when they have more stays and when the number of days since the last stay is lower. Perhaps a different offer would be more appropriate for this group to get them to think of Stay-With-Us Lodging when they travel. Because they are about 18% of our customer base, it is worthwhile to consider their contribution to the bottom line. The most important predictors are "number of days since the last stay" and "number of stays". If we think of them as frequent customers, we will probably be less successful in our communications and offers. We should think of them as what they are, which is almost new customers who still need to be persuaded that our brand is the best for them.

What we have discovered in our brief investigation is that for different segments of our customer population, it is best to use different offers and communications. We often observe that what drives the customers in one cluster to respond to a communication or offer is often different from those in another cluster. A cluster of young, low-income people might be price sensitive, while an older and more salaried group might be interested in higher-quality products.

When your business process will accommodate, you might consider clustering by stores rather than by customers. If your purchase patterns vary for different types of stores or categories of customers, then you should consider group clustering rather than individual clustering. Then, you can build your models within each cluster.

## CONCLUSION

Segmentation enables you to treat your customers relevantly according to their characteristics. In this way, you can improve your ability to meet your customers' needs and wants. Because it is so easy to use statistical segmentation with SAS Visual Statistics, we recommend that you use either the decision tree or the clustering visualization. If you want to segment based on a target variable, then use the decision tree. If you want to get an unsupervised segmentation with no target variable, then use the clustering visualization. Determine what you want to accomplish and then use that goal to select the method. If you also have certain business rules that must be followed, you can incorporate them either into more variables or into filtering approaches. Because SAS Visual Statistics is so fast, you can use it to explore your data as well as determine which analytic approaches work the best. However, you must always stay focused on the end goal, which is what the marketing team is looking for.

The marketing team wants to communicate and interact with the customers in ways that are most appropriate to them. Segmentation delivers groups that differ based on the input attributes. Predictive models within the groups enables you to see and act on those attributes that are most significant to the model. The predictive model addresses the question, "What drives my customer to interact with my company the way they do?" As you become more familiar with your groups of customers, you may find that it is more profitable to communicate differently with each group and provide them with offers that are most consistent with what motivates them to be your customer. If you want to motivate the Stay-With-Us customer to make another reservation soon, then you must provide an offer that is consistent with how that customer wants to be treated. Running various test campaigns within your segments enables you to determine what will work and what will not. On an even more advanced level, you might be able to move customers from a less profitable group to a more profitable one through your communications and offers.

In summary, by segmenting and building predictive models within clusters, we are better able to understand our customers, thereby providing better interactions, and, ultimately, increasing customer satisfaction and profitability. Throughout this paper, we have focused on both the science as well as the art of segmentation and modeling. After building your models, you can easily score your customers within SAS Visual Analytics as well as export the score code to use elsewhere.

SAS Visual Analytics and SAS Visual Statistics deliver an interactive and collaborative technology to produce quick solutions while being curious and innovative. Marketers are dreamers who are always looking for new data sources (including clickstream, social, Internet of Things wearables, and so on), which require the ability to scale to very large amounts of information. What is compelling about SAS Visual Analytics and SAS Visual Statistics is their ability to perform sophisticated analysis, and to produce visualizations that support data-driven storytelling, ensuring that brands understand and take action on insights.

With the availability of visualization and the ability to run your analyses in an interactive, point-and-click interface, you can easily use sophisticated segmentation approaches of supervised classification and unsupervised clustering. You can always add in business rules and quantile approaches if you find them useful. However, not having statistical abilities does not restrict you from using the SAS Visual Analytics and SAS Visual Statistics interface to process your data. SAS Visual Analytics and SAS Visual Statistics offer you the technology to create market-driven segments, enabling a powerful marketing strategy that is comprehensive and nimble. As you incorporate algorithmically-derived segmentation and, in some cases, predictive modeling and scoring by segment into your marketing strategies, you can quickly take results from one campaign and influence future campaigns with increased intelligence and personalization.

Customer Segmentation Intelligence encourages you to go beyond segmentation and predictive modeling by integrating the two into a more cohesive customer interaction strategy. You goal is to treat all your customers as they would like to be treated, while focusing on corporate profitability. Realizing that you will likely have different offers and communication for different segments requires you to both segment your population and to analyze your predictive models within each segment. By following this path, you will provide your marketing team with the most powerful solution that is both dynamic and data-centered for your customers.

## REFERENCES

- Baer, Darius S. 2012. "CSI: Customer Segmentation Intelligence for Increasing Profits." Proceedings of the SAS Global Forum 2012 Conference. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings12/103-2012.pdf.
- Baer, Darius S., Ph.D., and Grover, Suneel, MBA, MS. "Enhanced Segmentation Using SAS® Visual Analytics and SAS® Visual Statistics." *Proceedings of the SAS Global Forum 2016 Conference*. Cary, NC: SAS Institute Inc. Available at <u>http://support.sas.com/resources/papers/proceedings16/SAS6222-</u> 2016.pdf.
- 3. SAS Institute Inc. 2014. SAS® Visual Statistics 7.1: User's Guide. Cary, NC: SAS Institute Inc. . Available at <a href="https://support.sas.com/documentation/onlinedoc/vs/7.1/vsug.pdf">https://support.sas.com/documentation/onlinedoc/vs/7.1/vsug.pdf</a>.

## **RECOMMENDED READING**

SAS Institute Inc. 2014. SAS® Visual Statistics 7.1: User's Guide

## **CONTACT INFORMATION**

Your comments and questions are valued and encouraged. Contact the author at:

Darius Baer SAS Institute, Inc. 420 Edison Place Superior, CO 80027 919-531-9510 303-204-7015 darius.baer@sas.com https://www.linkedin.com/in/darius-baer-614a8b www.sas.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.