



Who are my best customers?

Using SPSS to get the most out of your marketing database

Understanding
your
customers
is the key
to developing
successful
marketing
programs

Who are my best customers? This is a question that essentially every business professional — every marketer, sales manager, product developer and service specialist — would be delighted to answer.

And, in today's turbulent marketplace it continues to get more difficult and more expensive to reach, attract and retain customers. Due to these pressures, more organizations are using database marketing to maximize the value of their existing customers. Turning customer data into knowledge and information to act on is a powerful tool and a necessary element of business survival.

Understanding the unique characteristics of your customers gives you valuable insight. Knowledge about your most and least profitable customers, their purchase patterns, buying behaviors and demographic profile are key to developing successful marketing programs. A better understanding of what your customers look like helps you develop customer loyalty, retention and rewards programs, up-sell and cross-sell programs, and target marketing programs. Informed decisions are also the basis for successful advertising, promotions, direct mail campaigns and other marketing communications.

Many ways exist to examine characteristics that define your best customers, and just as many ways exist to measure those characteristics. This white paper explains one way to perform a customer analysis using SPSS.

In this case study, the marketing database includes 2,000 customers with the following data:

- date when customer first became a customer
- purchase history by dollar value of orders
- response rate to different offers
- household income level
- region
- gender, and other demographic variables

The goal is to look at the amount of money spent over time to identify different segments of customers by demographic group. In this paper we'll use various data analysis techniques, from basic to sophisticated, to extract actionable information from our database.

The insight you yield with even the most elementary procedures can have profound implications on how well you understand your customers. It's important not to underestimate them. Combining knowledge of your business with a flexible and powerful analysis tool is the best way to get the most out of your data.

Initial exploration: What is my typical customer like?

We begin by exploring the different variables in our database to answer questions such as:

- In which regions do our customers reside? How are our customers distributed across the three sales regions?
- What is the average income level of our customers?
- How long have our customers been customers?
- What is the average response to our different promotions? How many people responded to Offer 1?
- How much money do our customers spend?

SPSS offers several methods to quickly obtain the answers to your questions



	Frequency	Percent	Valid Percent	Cumulative Percent
Valid East	707	34.2	41.1	41.1
Middle	611	29.5	35.5	76.6
West	402	19.4	33.4	100.0
Total	1720	83.1	100.0	
Missing System	350	16.9		
Missing				
Total	350	16.9		
Total	2070	100.0		

Chart 1 and Table 1. The SPSS table and chart, automatically created with complete labels, reveal most customers (34 percent) live in the Eastern region.

SPSS offers several methods to quickly obtain the answers to these questions. SPSS Frequencies and Descriptives procedures are very good at providing a first look at our data, and presumably, more ideas on the kinds of analyses we will perform.

Analyzing where customers reside help us determine territories. SPSS Frequencies provide a table of counts and percents by category along with a visual representation of the data in a bar, histogram or pie chart. SPSS automatically presents the results as a table and chart complete with labels.

From this analysis we learn what may prove to be important. From the pie chart and results in Table 1, we see the largest portion of our customer base (34 percent) lives in the Eastern region, and the smallest proportion of customers (19 percent) reside in the West. And, 16.9 percent of our database have no region listed.

SPSS flags missing data for special treatment. It is useful to know when and why information is missing. For example, you might want to distinguish between data missing because they don't apply and data missing because they are unavailable. In Table 1, "percent" includes the missing data, "valid percent" excludes it from the calculations, for a fast side-by-side comparison of how the missing data affect the results.

To get information on household income, we examine basic summary statistics, such as the mean, minimum and maximum values. Interval, or continuous variables, such as income measured in dollars or age measured in years, are best first examined with descriptive statistics. The SPSS Descriptives procedure gives us a set of summary statistics. We see from Table 2 that the average household income of the 2,000 customers in our database is approximately \$61,000, and that most incomes range from about \$50,000 to \$72,000.

	Minimum	Maximum	Mean	Std. Deviation
HH Income	\$38,552.00	\$95,571.00	\$61,298.29	\$11,886.75

Table 2. The SPSS Descriptives procedure provides a quick summary statistics showing average household income is approximately \$61,000.

For the most accurate customer lifetime value, use a predictive model

To answer the question “How long have our customers been customers?” we must manipulate a field in our database and then count the number of customers in each period. Since the database contains the date we entered the customer into the database, we first compute a new variable: length of time as a customer. By using one of the many time functions available in SPSS, we easily transform the date into the length of time, in years, since we

entered the customer. After computing this new variable, we can request a frequency chart (Table 3) of the length of time a customer has been a customer.

From Table 3 we learn about 29 percent of our customers have been in the database for more than 10 years, and just over half have been with us for seven years. Next, we ask “Who spends the most money?” Best customers are typically defined as the most profitable customers, or the customers who spend the most money with your organization.

# of Cases	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 1	180	8.2	8.2	8.2
2	172	8.3	8.7	17.2
3	147	7.1	7.4	24.6
4	144	7.0	7.3	31.9
5	123	6.0	6.7	38.5
6	121	6.0	6.7	44.6
7	145	7.0	7.3	51.9
8	120	6.0	6.8	58.7
9	112	5.6	6.8	64.4
10	126	6.0	6.8	71.2
11	107	5.7	6.9	77.1
12	120	6.0	6.9	83.6
13	100	5.7	6.9	89.5
14	110	5.5	6.8	94.4
15	111	5.6	6.8	98.0
Total	1984	95.8	100.0	
Missing System	86	4.2		
Total	90	4.2		
Total	2074	100.0		

Table 3. An SPSS frequency chart indicates that 51 percent of our customers have been so for more than seven years.

For the most accurate customer lifetime value, a predictive model combines previous purchases and behaviors and a forecast of future purchases. In this example, we begin with the total value of the orders placed by each customer.

First, we create a new variable, total order value (in dollars), by summing the value of each order (Value1, Value2 and so on) in our database. Since total value is a continuous variable, a histogram is the most efficient way to graphically display the results.

Promotion analysis is another important step toward understanding customers

In a histogram, each bar represents a range of data. From the histogram in Chart 2 we see the majority of customers spent \$500 or less, and that fewer people spent more and more money. The average amount spent by customers is \$1,360 and a very few customers spent in excess of \$7,000.

So far, we know a typical customer:

- lives in the East
- has a household income of \$61,000
- has been a customer for seven years
- spends \$1,360 on our products and services

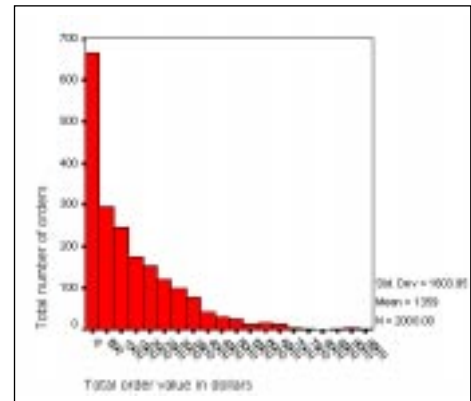


Chart 2. From the histogram, we see the majority of customers spent \$500 or less, and fewer people spent more money.

How did customers respond to the different offers?

Performing promotion analysis is another important step toward understanding customers. Evaluating marketing programs and offers helps identify what worked and what did not. It helps pinpoint when and why certain programs were successful, so you can duplicate your success and learn from your failures.

Offer 1 response				
		Frequency	Percent	Cumulative Percent
Valid	Did not respond	1110	55.0	55.0
	Responded	890	45.0	100.0
Total		2000	100.0	
Missing	System Missing	10	.5	
	Total	10	.5	
Total		2010	100.0	

Table 4. Almost 45 percent, or 890 people in the customer database, responded to Offer 1.

To answer the questions “How many people responded to each offer?” and “What is the average response to our different promotions?” we run an SPSS Frequency on each offer response and SPSS Descriptives on order value for the four offers.

shows 39 percent response to Offer 2, 37.4 percent response to Offer 3 and 17.4 percent response to Offer 4.

In Table 4, we see 890, or almost 45 percent of people in the customer database responded to Offer 1. Similar analysis for the other offers

This raises a new question: Were there unique characteristics in Offer 4 that made it more successful in getting people to respond? In other words, is this result significant?

Description Statistics				
	Minimum	Maximum	Mean	Std. Deviation
Order value (offer 1)	\$ 0.00	\$3,280.23	\$376.64	\$602.32
Order value (offer 2)	\$ 0.00	\$3,477.31	\$312.38	\$566.38
Order value (offer 3)	\$ 0.00	\$3,377.18	\$293.98	\$536.48
Order value (offer 4)	\$ 0.00	\$3,339.68	\$175.63	\$496.73

The information on purchase history (Table 5) reveals the average value for Offer 3, \$294, is lower than the other offers. Whether this difference is significant will be determined by further analysis.

Table 5. The analysis on purchase history reveals the average value for Offer 3, \$294, is lower than the other offers.

Further analysis: How do my customers differ?
How are they similar?

Now that we have a basic understanding of our customers and the success of various offers, we leverage the power of analysis by looking at two or more variables at once. SPSS helps find underlying relationships that are difficult to see otherwise. For example, we already know how our customers are distributed across the regions and how many people responded to Offer 1. Next we'll look into how people responded to Offer 1 based upon the region. We'll investigate the answers to these questions:

- What is the average customer lifetime by each region?
- How did people respond to Offer 1 based on the region?

SPSS makes it easy to compare different groups of data. SPSS' Crosstabs, Comparison of Means, Clustered bar charts and boxplots present results clearly; Chi-square statistics, Analysis of Variance (ANOVA) and SPSS CHAID identify when results and findings are statistically significant. This is important because, when you know what is meaningful, you don't waste your efforts.

Next we explore the question, "What is the average customer lifetime by each region?" A powerful statistical chart, the boxplot, displays both the mean and distribution of the data together. From the boxplot in Chart 3, it is easy to see the average length of time for the Western region is greater than the other regions.

Comparison of Means provides summary statistics for a joint distribution. The report in Table 6 (contains the same information as the boxplot, but in table format) shows while the overall average length of time in the database is 7.49 years, people in the West have a longer average tenure than do those in the Eastern or Middle region. Is this a significant finding?

Statistical significance tells you if the differences you see are random, or if they are sufficiently large to justify further consideration. If the differences are random, it means the results are what would reasonably be expected to happen. That is, none of the variables had a significant influence or impact on the results.

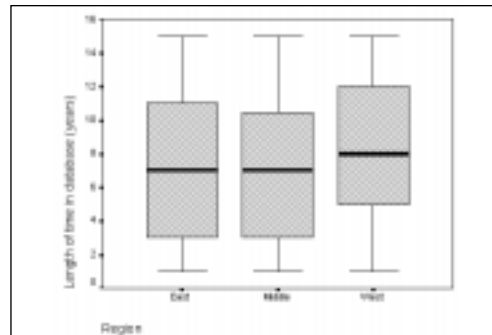


Chart 3. The boxplot displays both the mean and distribution of the data together. It is easy to see the average length of time for the Western region is greater than the other regions.

SPSS makes it easy to compare different groups of data

Report			
Length of time in database (years)			
East	Mean		7.38
	N		702
	Std. Deviation		4.40
Middle	Mean		6.96
	N		601
	Std. Deviation		4.27
West	Mean		8.48
	N		398
	Std. Deviation		4.31
Total	Mean		7.49
	N		1708
	Std. Deviation		4.37

Table 6. This Comparison of Means report shows while the overall average length of time in the database is 7.49 years, people in the West have a longer tenure than those in the East or Middle.

If the differences are statistically significant, it means they were higher than expected to occur, and indicates the potential influence of some additional non-random factor. When statistical significance exists, it is a strong indication for further exploration.

The ANOVA report in Table 7 shows the differences between region and length of time *are* statistically significant. Since the significance is .000, or less than .05, we can conclude the differences in means are likely significant: the overall distribution of average length and region is probably not due to random causes, but to something else. Examples of possible causes are: we first opened a regional office in the West, more need for the product in different areas, or a certain product feature was introduced successfully to one region. This is where it is also important to know your business to leverage data to support your hunches.

		Sum of Squares	df	Mean Square	F	Sig.
Length of time in database (years * region)	Between Groups	588.490	1	588.490	15.188	.000
	Within Groups	31985.954	1795	17.819		
Total		32574.444	1796			

Table 7. The ANOVA report shows the differences we see are statistically significant, a strong indication for further exploration.

When statistical significance exists, it is a strong indication for further exploration

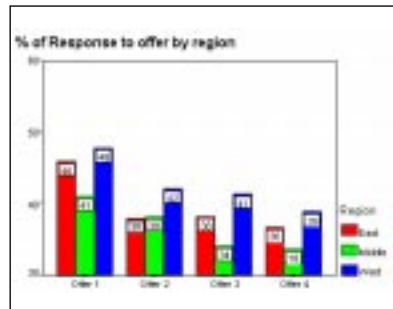


Chart 4. The SPSS clustered bar chart provides a quick and clear way to present response patterns by region.

		Offer 1 response			
		Did not respond	Responded	Total	
Region - East	Count	391	318	709	
	% of Region	55.3%	44.7%	100.0%	
	% of Offer 1	41.0%	41.3%	41.1%	
Middle	Count	364	247	611	
	% of Region	59.6%	40.4%	100.0%	
	% of Offer 1	39.1%	32.1%	35.6%	
West	Count	198	263	461	
	% of Region	49.5%	50.5%	100.0%	
	% of Offer 1	20.9%	26.5%	23.4%	
Total		954	788	1742	
		% of Region	55.5%	44.5%	100.0%
		% of Offer 1	106.8%	189.0%	100.0%

Table 8. While only 26.5 percent of the people who responded to Offer 1 were from the West, over half (50.5 percent) of the Westerners responded to the Offer.

Next we continue our analysis of offer response. SPSS provides a quick way to present the information for all four offers together, graphically using the clustered bar chart. Chart 4 provides a summary of response patterns by region. We see the Middle region tends to under-order relative to the other two, particularly the West. This is a finding we could not have guessed by looking at the frequency distribution of region, which showed us the Western region contained the fewest people.

To find out if this is significant, we can further explore the results of individual offers by region. To answer the question “How did people respond to Offer 1 based upon the region?” we perform an SPSS crosstab on Offer 1 and region. Table 8 shows 41.3 percent of the people who responded to Offer 1 were from the East. While only 26.5 percent of the people who responded to Offer 1 were from the West, over half (50.5 percent) of the Westerners responded to the Offer. To understand if region determines the likelihood of response to Offer 1, we compare the percentages in the ‘% of Region’ and find that 45 percent of people

from the Eastern region responded to this offer, and that 40 percent of people in the Middle responded. Based on this information, we conclude the West is a good region for an offer such as Offer 1. However, while it appears the percentages are different, that is insufficient reason to start duplicating Offer 1 in the Western region. First, we must determine if these percentages are statistically significant. Here, the Chi-square statistic indicates if statistical significance exists.

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	9.970 ^a	2	.007
Likelihood Ratio	9.954	2	.007
Linear-by-Linear Association	1.999	1	.157
N of Valid Cases	1720		

^a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 179.33.

Table 9. A Chi-square of .007 for the region and Offer 1 indicates the difference between regions is significant.

Table 9 contains Chi-square information for the region and Offer 1. Convention holds that the Pearson Chi-square statistic should be less than .05 for the exhibited differences to be statistically significant (at the 95 percent confidence level). In this case, the Chi-square is .007, and is therefore significant. There could be a specific, identifiable reason that made Offer 1 more successful in the Western region, such as the copy spoke more directly to their needs, or the media type was better matched to get and keep their attention.

By identifying what made the campaign successful in the West, we can leverage that knowledge in future offers to this region. We also may choose to explore any relationships that underlie region.

The Chi-square statistic tells you if the differences you are seeing are random

Which customers spend the most money?

Another way to look at purchase history is to assess total amount spent, rather than just the money spent on individual orders. Perhaps a relationship between total money spent and region will reveal some insights.

A one-way ANOVA gives you specific information about the significance of the differences in average value that you may see.

The first thing that one-way ANOVA provides is a table of Descriptive Statistics. Table 10 shows the average total amount spent from all four offers by region vary widely. In the Middle region, the average amount spent was \$1,206, and in the East, \$1,391, while in the West, the average was over \$1,600.

Dependent Variable: Total Orders		Groupings		95% Confidence Interval		
Chi-Square	LL Percent	Mean Difference (I-J)	Std. Error	df	Lower Bound	Upper Bound
East	Middle	\$ 195.00	\$ 88.71	112	-\$ 61.02	\$ 461.02
	West	-\$ 234.26	\$ 108.23	109	-\$ 472.36	\$ 103.84
Middle	East	\$ 194.00	\$ 88.71	112	-\$ 61.02	\$ 461.02
	West	-\$ 112.50	\$ 108.23	109	-\$ 324.93	\$ 99.93
West	East	\$ 235.00	\$ 108.23	109	\$ 67.97	\$ 402.03
	Middle	\$ 411.26	\$ 108.23	109	\$ 153.80	\$ 668.72

* The mean difference is significant at the .05 level.

Table 10. The one-way ANOVA shows the difference between spending levels in the East and West are not statistically significant; the difference between the Western and Middle regions are significant.

The final piece of the report shows the average difference exhibited between the spending levels in the East and the West are *not* statistically significant. On the other hand, it shows the difference between the Western and the Middle regions *are* significant. You can use this meaningful information to further identify how and why these regions differ, and develop targeted marketing plans to leverage the differences. For example, a different marketing and sales mix, different offer, or special bundle of products and services may

work better in the Middle region. The marketing programs in the West should be repeated in the West for even greater success.

Predicting the total amount spent

Predictive models are powerful tools to help target your prospects and optimize marketing resources. They help answer questions such as “How much will a household spend given their income?”

Predictive models are powerful tools to help target your prospects and optimize marketing resources



Chart 5. The scatterplot shows the shape of the relationship between these two variables.

		HH Income	Total value of orders 1-4
Pearson	HH Income	1.000	.608**
Correlation	Total value of orders 1-4	.608**	1.000
Sig. (2-tailed)	HH Income		.000
	Total value of orders 1-4	.000	
N	HH Income	1984	1984
	Total value of orders 1-4	1984	2000

** Correlation is significant at the 0.01 level (2-tailed).

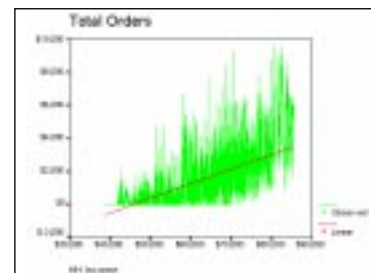
Table 11. The correlation coefficient shows a strong relationship of 60.8 percent revealing that as household income increases, the total money spent on our products increases.

In many statistical studies, the goal is to establish a relationship, expressed via an equation, for predicting typical values of one variable given the value of another. SPSS offers several procedures for establishing relationships and defining predictive models from scatterplots and correlations, to linear and logistic regression analysis, to CHAID analysis. And, with SPSS’ tutorial, step-by-step instructions and “What’s this?” help, you don’t have to be a statistician to perform these procedures.

Chart 5 shows the shape of the relationship between these two variables. The scatterplot is the right chart to display the joint distribution of two continuous, or interval variables.

MODEL: MOD_1							
Independent: MINCOME							
Dependent	Min	Rsq	d.f.	F	Sigt	ss	t1
TOTVALB	LN	.370	1982	1162.26	.000	4075.3	.0667

Table 12 and Chart 6. A linear regression defines the relationship between household income and total money spent. The more money earned, the more they spend.



The correlation coefficient of 60.8 percent, displayed in Table 11 indicates a strong relationship between household income and total money spent. Regression analysis further defines the relationship with a model, as shown in Table 12 and Chart 6. This relationship means that as household income increases, the total money spent on our products increases. We could use this finding to better forecast sales and improve our market efforts.

Example programs include: targeting higher income households with more products and services, or developing customer retention programs that help keep the higher income households happy, long-term customers, while matching marketing resources to the potential revenue of the segment. So far, we have seen a relationship between a customer's region and their likelihood of spending money with us. Additionally, we have seen that income is positively related to total money spent.

Segmenting customers for more profitable and successful marketing

CHAID identifies the unique segments within the data, so you can get the best results from your marketing programs

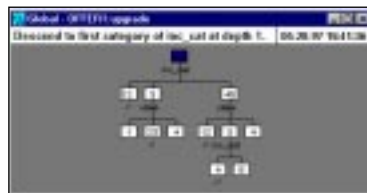


Chart 7. SPSS CHAID presents a model of which combinations are most likely to respond to Offer 1.

Database marketers often use a technique called Chi-squared Automatic Interaction Detection, or CHAID. Rather than tell us whether the relationship between two variables appears significant, CHAID tells us what combinations of characteristics from several variables are most likely to result in an outcome (for this example, response to an offer). We put region, product class category and categorized income into a CHAID model to find out which combinations are most likely to respond to Offer 1. SPSS CHAID automatically builds a tree diagram of the results, as in Chart 7.

Chart 8 shows the detail of the top branch, which shows the variable with the most significant influence on the response to Offer 1. Income was found to be the highest predictor (which corresponds to the earlier regression findings). In this case, CHAID goes beyond the regression example to explore further interactions.

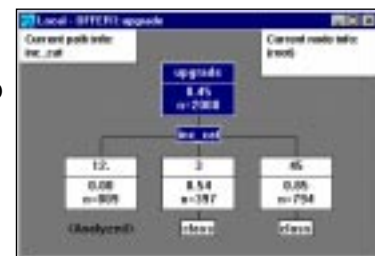


Chart 8. The top branch of the CHAID diagram reveals income as the highest predictor of response.

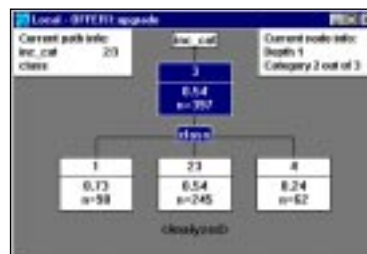


Chart 9. CHAID identifies a unique segment with income in category 3 (\$57,750 to \$65,000) and product class in category 1.

In Chart 9, the details of the next level of branches reveals that if income is in category 3 (\$57,750 to \$65,000) and if product class is category 1, there is a 73 percent response rate. CHAID identifies the unique segments within the database, so you can leverage the combinations of characteristics to get the best results from your marketing programs.

We found households with income of \$57,750 to \$65,000 (who purchased from product class 1) are more likely to purchase from Offer 1. A CHAID analysis with additional variables may lead to additional findings; for example, we may find that while in general, the Middle did not respond well to

our offers, women of another particular income group did respond well, and thus may be a fruitful target for another direct mail campaign.

Taking action

SPSS helped
us learn more
about
underlying
patterns

SPSS allowed us to quickly assess the averages and distributions of our data to learn some important things about our typical customers: they tend to be longer-term customers, from the Eastern region, have not responded well to Offer 3 on the whole, and are likely to have higher incomes. Understanding the profile of our typical customer helps provide better insight for future marketing efforts. By comparing multiple characteristics and groups, SPSS helped us learn more about underlying patterns: not only was Offer 3 the least lucrative for us, but it was particularly unproductive in the Middle region, a region which tended to respond less well than did the other two. And, customers in the Middle region had the lowest average income, helping to explain their relatively low response to our offers. By identifying these groups of customers, we can target marketing and customer retention programs.

Finally, using powerful SPSS predictive modeling and segmentation techniques to identify relationships, we developed a model that describes the relationship between income and total money spent to help predict future sales. We also identified unique customer segments by their likelihood to respond to Offer 1. Using segmentation results based on predicted response is the key to developing profitable marketing programs. When segment characteristics are matched with individual customers and prospects, you can duplicate successful programs and revise or eliminate unprofitable programs to get the best results.

As a result of this analysis, we can make the following plans:

- Build a new customer retention program for best customers in the segment defined by high-income, Western region, long-time customers, who purchase in product class 1
- Develop and test a new bundle of products and services to better target the needs of the Middle region, lower income customers and prospects
- Repeat sales development of Western regions in Middle and Eastern regions to build long-time customers
- Duplicate Offer 1 to prospects in the Western region
- Match the funds of future marketing campaigns to the predicted segment profitability (based initially on household income)

By performing more tasks, you could pursue more interrelationships. For the purposes of this paper, however, we have shown that SPSS gives you a host of analysis options, and you do not need to be a statistician, or even employ the most sophisticated techniques in SPSS, to learn valuable information with real business implications.

About SPSS

SPSS Inc. is a multinational software products company that provides statistical product and service solutions. The company's mission is to drive the widespread use of statistics. SPSS products and services are used worldwide in corporate, academic and government settings for all types of research and data analysis.

The company's four lines of business are: business analysis (including survey research, marketing and sales analysis and data mining); scientific research; quality improvement; and process management. Headquartered in Chicago, SPSS has more than 30 offices and 60 distributors serving countries around the world.

Contacting SPSS

To place an order or to get more information, call your nearest SPSS office or visit our World Wide Web site at <http://www.spss.com>

SPSS Inc.	+1.312.329.2400	SPSS Ireland	+353.1.66.13788
United States and Canada	Toll-free: 1.800.543.2185	SPSS Israel Ltd.	+972.9.526700
SPSS Bay Area	+1.415.453.6700	SPSS Italia srl	+39.51.252573
SPSS Federal Systems (U.S.)	+1.703.527.6777	SPSS Japan Inc.	+81.3.5466.5511
SPSS Argentina srl.	+541.816.4086	SPSS Korea	+82.2.552.9415
SPSS Asia Pacific Pte. Ltd.	+65.3922.738	SPSS Latin America	+1.312.494.3226
SPSS Australasia Pty. Ltd.	+61.2.9954.5660 Toll-free: 1800.024.836	SPSS Malaysia Sdn Bhd	+603.704.5877
SPSS Belgium	+32.162.389.82	SPSS Mexico Sa de CV	+52.5.575.3091
SPSS Benelux	+31.183.636711	SPSS Middle East and Southeast Asia	+971.4.525536
SPSS Central and Eastern Europe	+44.(0)1483.719200	SPSS Newton	+1.617.965.6755
SPSS East Mediterranean and Africa	+972.9.526700	SPSS Scandinavia AB	+46.8.102610
SPSS France SARL	+33.1.4699.9670	SPSS Schweiz AG	+41.1.266.90.30
SPSS Germany	+49.89.4890740	SPSS Singapore Pte.	+65.2991238
SPSS Hellas SA	+30.1.7251925	SPSS Taiwan Corp.	+886.2.5771100
SPSS Hispano- portuguesa S.L.	+34.1.447.37.00	SPSS UK Ltd.	+44.1483.719200

SPSS is a registered trademark and the other SPSS products named are trademarks of SPSS Inc. All other names are trademarks of their respective owners.