DATA UNDERSTANDING AND DATA PREPARATION PHASES

Clothing Store Data Set

For this case study we meld together the data understanding and data preparation phases, since what we learn in each phase immediately affects our actions in the other phase. The *clothing-store* data set contains information about 28,799 customers in the following 51 fields:

- Customer ID: unique, encrypted customer identification
- Zip code
- Number of purchase visits
- Total net sales
- Average amount spent per visit
- Amount spent at each of four different franchises (four variables)
- Amount spent in the past month, the past three months, and the past six months
- Amount spent the same period last year
- Gross margin percentage
- Number of marketing promotions on file
- Number of days the customer has been on file
- Number of days between purchases
- Markdown percentage on customer purchases
- Number of different product classes purchased
- Number of coupons used by the customer
- Total number of individual items purchased by the customer
- Number of stores the customer shopped at
- Number of promotions mailed in the past year
- Number of promotions responded to in the past year
- Promotion response rate for the past year
- Product uniformity (low score = diverse spending patterns)
- Lifetime average time between visits
- Microvision lifestyle cluster type
- Percent of returns
- Flag: credit card user
- Flag: valid phone number on file
- Flag: Web shopper

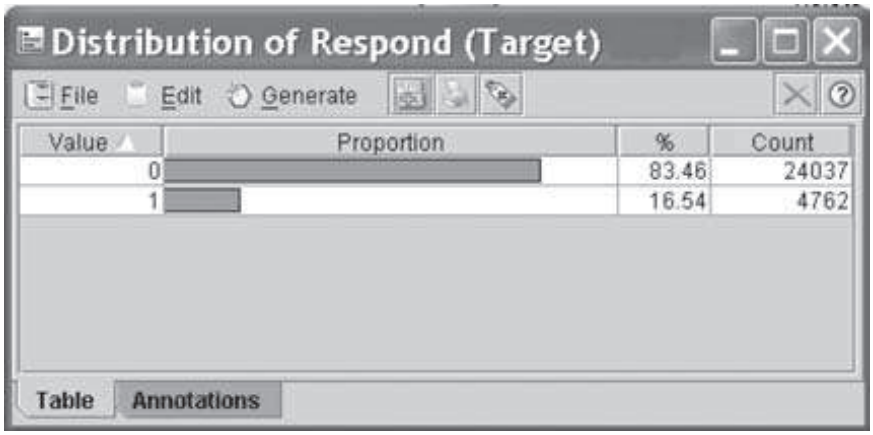


Figure 7.2 Most customers are nonresponders.

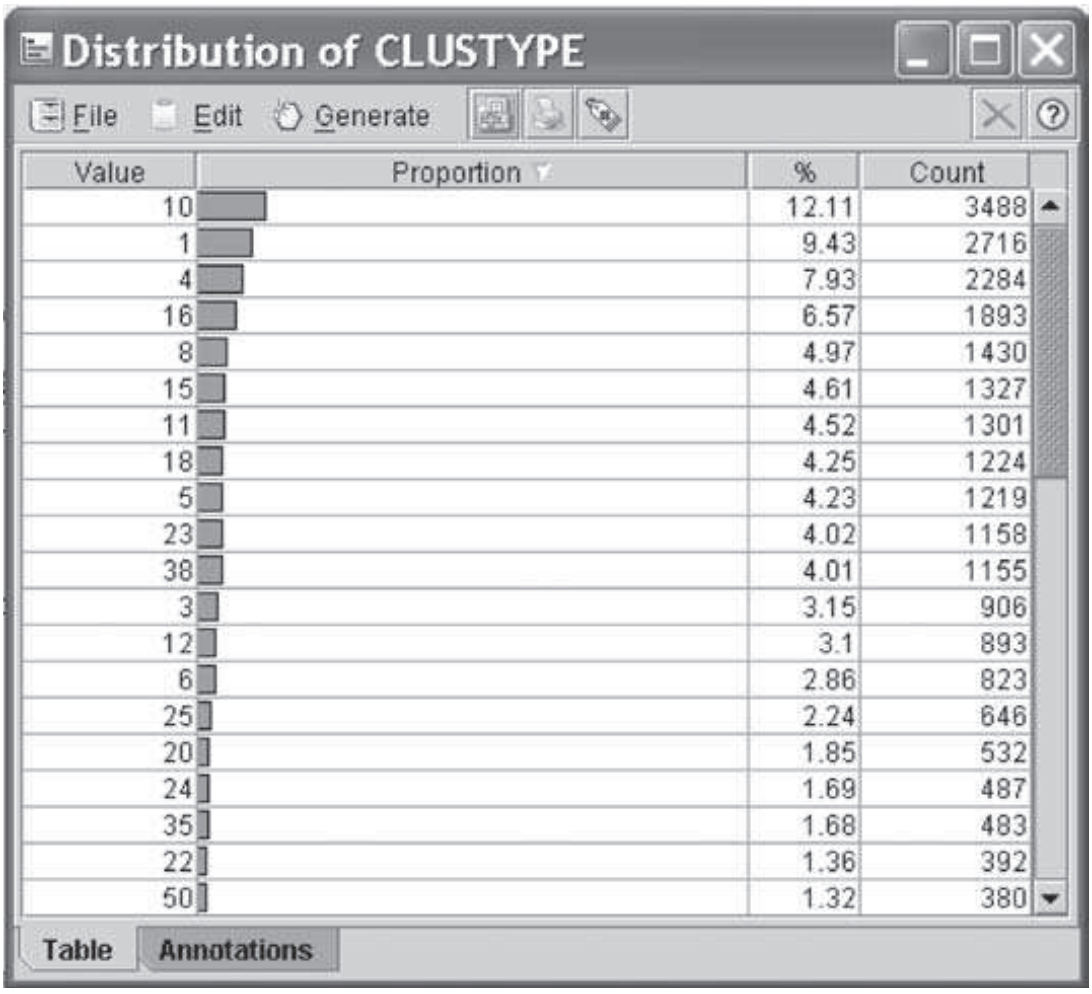
- 15 variables providing the percentages spent by the customer on specific classes of clothing, including sweaters, knit tops, knit dresses, blouses, jackets, career pants, casual pants, shirts, dresses, suits, outerwear, jewelry, fashion, legwear, and the collectibles line; also a variable showing the brand of choice (encrypted)
- Target variable: response to promotion

These data are based on a direct mail marketing campaign conducted last year. We use this information to develop classification models for this year's marketing campaign. In the data understanding phase, we become more familiar with the data set using exploratory data analysis (EDA) and graphical and descriptive statistical methods for learning about data. First, what is the proportion of responders to the direct mail marketing promotion? Figure 7.2 shows that only 4762 of the 28,799 customers, or 16.54%, responded to last year's marketing campaign (1 indicates response, 0 indicates nonresponse.) Since the proportion of responders is so small, we may decide to apply balancing to the data prior to modeling.

One of the variables, the Microvision lifestyle cluster type, contains the market segmentation category for each customer as defined by Claritas Demographics [4]. There are 50 segmentation categories, labeled 1 to 50; the distribution of the most prevalent 20 cluster types over the customer database is given in Figure 7.3.

The six most common lifestyle cluster types in our data set are:

1. *Cluster 10: Home Sweet Home*—families, medium-high income and education, managers/professionals, technical/sales
2. *Cluster 1: Upper Crust*—metropolitan families, very high income and education, homeowners, manager/professionals
3. *Cluster 4: Midlife Success*—families, very high education, high income, managers/professionals, technical/sales
4. *Cluster 16: Country Home Families*—large families, rural areas, medium education, medium income, precision/crafts
5. *Cluster 8: Movers and Shakers*—singles, couples, students, and recent graduates, high education and income, managers/professionals, technical/sales
6. *Cluster 15: Great Beginnings*—young, singles and couples, medium-high education, medium income, some renters, managers/professionals, technical/sales



Value	Proportion	%	Count
10		12.11	3488
1		9.43	2716
4		7.93	2284
16		6.57	1893
8		4.97	1430
15		4.61	1327
11		4.52	1301
18		4.25	1224
5		4.23	1219
23		4.02	1158
38		4.01	1155
3		3.15	906
12		3.1	893
6		2.86	823
25		2.24	646
20		1.85	532
24		1.69	487
35		1.68	483
22		1.36	392
50		1.32	380

Figure 7.3 The 20 most prevalent Microvision lifestyle cluster types.

Overall, the clothing store seems to attract a prosperous clientele with fairly high income and education. Cluster 1, *Upper Crust*, represents the wealthiest of the 50 cluster types and is the second most prevalent category among our customers.

Moving to other variables, we turn to the customer ID. Since this field is unique to every customer and is encrypted, it can contain no information that is helpful for our task of predicting which customers are most likely to respond to the direct mail marketing promotion. It is therefore omitted from further analysis.

The zip code can potentially contain information useful in this task. Although ostensibly numeric, zip codes actually represent a categorization of the client database by geographic locality. However, for the present problem, we set this field aside and concentrate on the remaining variables.