CHAPTER 3

EXPLORATORY DATA ANALYSIS

HYPOTHESIS TESTING VERSUS EXPLORATORY DATA ANALYSIS

When approaching a data mining problem, a data mining analyst may already have some a priori hypotheses that he or she would like to test regarding the relationships between the variables. For example, suppose that cell-phone executives are interested in whether a recent increase in the fee structure has led to a decrease in market share. In this case, the analyst would *test* the *hypothesis* that market share has decreased and would therefore use *hypothesis-testing* procedures.

A myriad of statistical hypothesis testing procedures are available through the traditional statistical analysis literature, including methods for testing the following hypotheses:

- The *Z*-test for the population mean
- The *t*-test for the population mean
- The *Z*-test for the population proportion
- The *Z*-test for the difference in means for two populations

Discovering Knowledge in Data: An Introduction to Data Mining, By Daniel T. Larose
ISBN 0-471-66657-2

- The *t*-test for the difference in means for two populations
- The *t*-test for paired samples
- The *Z*-test for the difference in population proportions
- The χ^2 goodness-of-fit test for multinomial populations
- The χ^2-test for independence among categorical variables
- The analysis of variance *F*-test
- The *t*-test for the slope of the regression line

There are many other hypothesis tests throughout the statistical literature, for most conceivable situations, including time-series analysis, quality control tests, and nonparametric tests.

However, analysts do not always have a priori notions of the expected relationships among the variables. Especially when confronted with large unknown databases, analysts often prefer to use *exploratory data analysis* (EDA) or *graphical data analysis*. EDA allows the analyst to:

- Delve into the data set
- Examine the interrelationships among the attributes
- Identify interesting subsets of the observations
- Develop an initial idea of possible associations between the attributes and the target variable, if any

GETTING TO KNOW THE DATA SET

Simple (or not-so-simple) graphs, plots, and tables often uncover important relationships that could indicate fecund areas for further investigation. In Chapter 3 we use exploratory methods to delve into the *churn* data set[1] from the UCI Repository of Machine Learning Databases at the University of California, Irvine. The data set is also available at the book series Web site. In this chapter we begin by using the Clementine data mining software package from SPSS, Inc.

To begin, it is often best simply to take a look at the field values for some of the records. Figure 3.1 gives the results of using Clementine's table node for the *churn* data set, showing the attribute values for the first 10 records. *Churn*, also called *attrition*, is a term used to indicate a customer leaving the service of one company in favor of another company. The data set contains 20 variables worth of information about 3333 customers, along with an indication of whether or not that customer churned (left the company). The variables are as follows:

- *State:* categorical, for the 50 states and the District of Columbia
- *Account length:* integer-valued, how long account has been active
- *Area code:* categorical
- *Phone number:* essentially a surrogate for customer ID

- *International Plan:* dichotomous categorical, yes or no
- *VoiceMail Plan:* dichotomous categorical, yes or no
- *Number of voice mail messages:* integer-valued
- *Total day minutes:* continuous, minutes customer used service during the day
- *Total day calls:* integer-valued

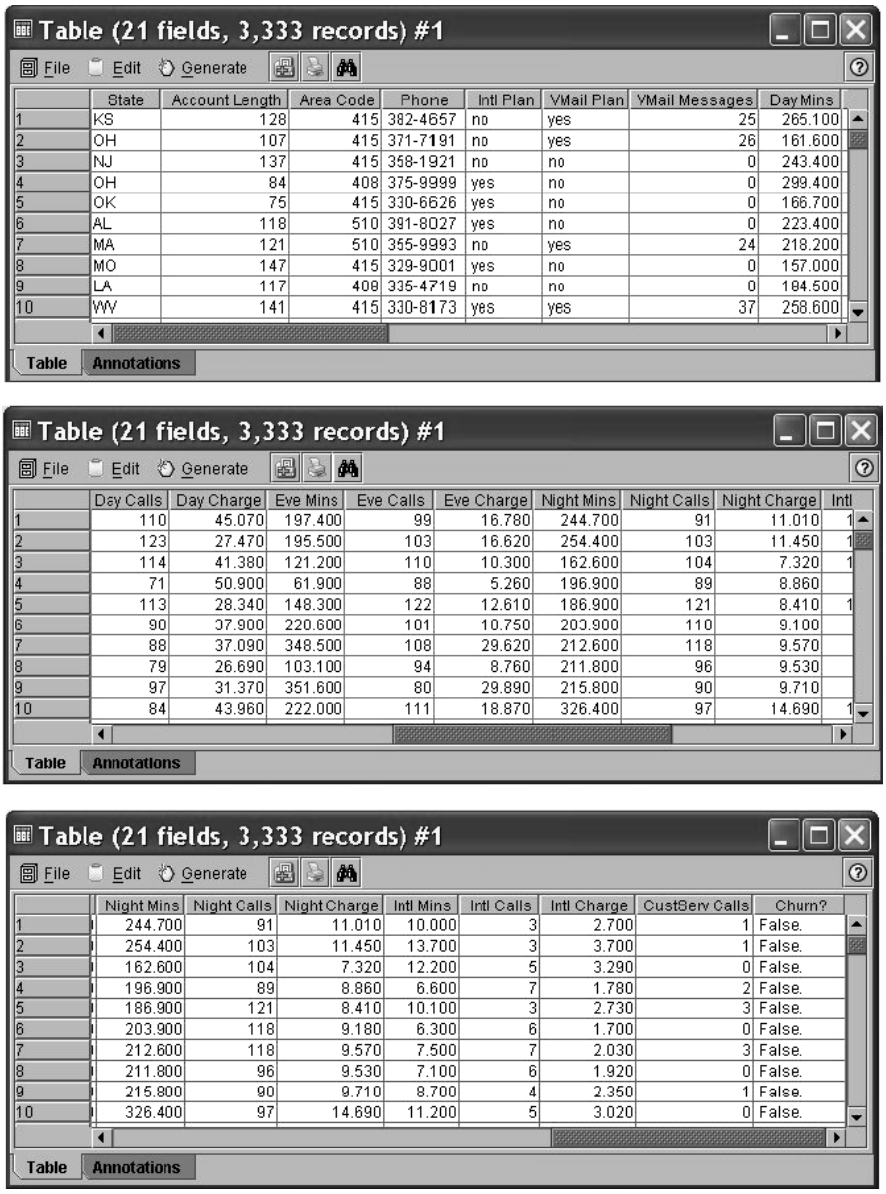

Table (21 fields, 3,333 records) #1

File Edit Generate

	State	Account Length	Area Code	Phone	Intl Plan	VMail Plan	VMail Messages	Day Mins
1	KS	128	415	382-4657	no	yes	25	265.100
2	OH	107	415	371-7191	no	yes	26	161.600
3	NJ	137	415	358-1921	no	no	0	243.400
4	OH	84	408	375-9999	yes	no	0	299.400
5	OK	75	415	330-6626	yes	no	0	166.700
6	AL	118	510	391-8027	yes	no	0	223.400
7	MA	121	510	355-9993	no	yes	24	218.200
8	MO	147	415	329-9001	yes	no	0	157.000
9	LA	117	408	335-4719	no	no	0	184.500
10	WV	141	415	330-8173	yes	yes	37	258.600

Table (21 fields, 3,333 records) #1

File Edit Generate

	Day Calls	Day Charge	Eve Mins	Eve Calls	Eve Charge	Night Mins	Night Calls	Night Charge	Intl
1	110	45.070	197.400	99	16.780	244.700	91	11.010	1
2	123	27.470	195.500	103	16.620	254.400	103	11.450	1
3	114	41.380	121.200	110	10.300	162.600	104	7.320	1
4	71	50.900	61.900	88	5.260	196.900	89	8.860	
5	113	28.340	148.300	122	12.610	186.900	121	8.410	1
6	98	37.900	220.600	101	18.750	203.900	118	9.180	
7	88	37.090	348.500	108	29.620	212.600	118	9.570	
8	79	26.690	103.100	94	8.760	211.800	96	9.530	
9	97	31.370	351.600	80	29.890	215.800	90	9.710	
10	84	43.960	222.000	111	18.870	326.400	97	14.690	1

Table (21 fields, 3,333 records) #1

File Edit Generate

	Night Mins	Night Calls	Night Charge	Intl Mins	Intl Calls	Intl Charge	CustServ Calls	Churn?
1	244.700	91	11.010	10.000	3	2.700	1	False.
2	254.400	103	11.450	13.700	3	3.700	1	False.
3	162.600	104	7.320	12.200	5	3.290	0	False.
4	196.900	89	8.860	6.600	7	1.780	2	False.
5	186.900	121	8.410	10.100	3	2.730	3	False.
6	203.900	118	9.180	6.300	6	1.700	0	False.
7	212.600	118	9.570	7.500	7	2.030	3	False.
8	211.800	96	9.530	7.100	6	1.920	0	False.
9	215.800	90	9.710	8.700	4	2.350	1	False.
10	326.400	97	14.690	11.200	5	3.020	0	False.

Figure 3.1 Field values of the first 10 records in the *churn* data set.

- *Total day charge:* continuous, perhaps based on foregoing two variables
- *Total evening minutes:* continuous, minutes customer used service during the evening
- *Total evening calls:* integer-valued
- *Total evening charge:* continuous, perhaps based on foregoing two variables
- *Total night minutes:* continuous, minutes customer used service during the night
- *Total night calls:* integer-valued
- *Total night charge:* continuous, perhaps based on foregoing two variables
- *Total international minutes:* continuous, minutes customer used service to make international calls
- *Total international calls:* integer-valued
- *Total international charge:* continuous, perhaps based on foregoing two variables
- *Number of calls to customer service:* integer-valued

DEALING WITH CORRELATED VARIABLES

One should take care to avoid feeding correlated variables to one's data mining and statistical models. At best, using correlated variables will overemphasize one data component; at worst, using correlated variables will cause the model to become unstable and deliver unreliable results.

The data set contains three variables: *minutes*, *calls*, and *charge*. The data description indicates that the *charge* variable may be a function of *minutes* and *calls*, with the result that the variables would be correlated. We investigate using the *matrix plot* shown in Figure 3.2, which is a matrix of scatter plots for a set of

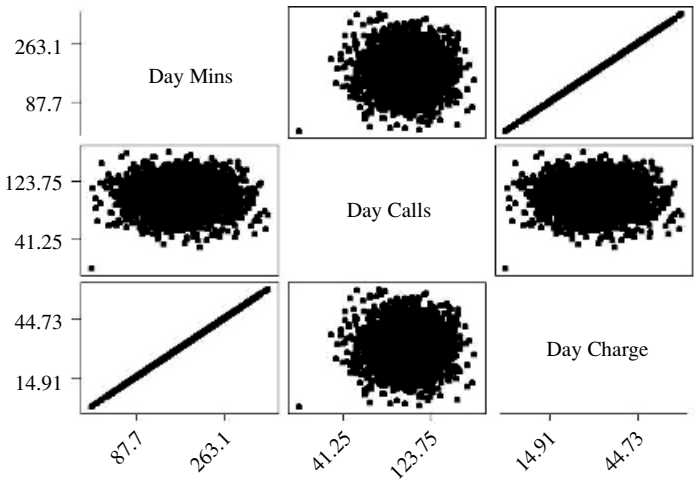

Figure 3.2 Matrix plot of *day minutes*, *day calls*, and *day charge*.

Regression Analysis: Day charge versus Day Mins

```
The regression equation is
Day Charge = 0.000613 + 0.170 Day Mins

Predictor        Coef     SE Coef           T        P
Constant    0.0006134   0.0001711        3.59    0.000
Day Mins     0.170000    0.000001   186644.31    0.000

S = 0.002864    R-Sq = 100.0%    R-Sq(adj) = 100.0%
```

Figure 3.3 Minitab regression output for *day charge* versus *day minutes*.

numeric variables. The matrix plot comes courtesy of Minitab, a widely used statistical package.

There does not seem to be any relationship between *day minutes* and *day calls* or between *day calls* and *day charge*. This we find to be rather odd, as one may have expected that as the number of calls increased, the number of minutes would tend to increase (and similarly for charge), resulting in a positive correlation between these fields. However, the graphical evidence does not support this, nor do the correlations, which are $r = 0.07$ for both relationships (from Minitab, not shown).

On the other hand, there is a perfect linear relationship between *day minutes* and *day charge*, indicating that *day charge* is a simple linear function of *day minutes* only. Using Minitab's regression tool (Figure 3.3), we find that we may express this function as the estimated regression equation: "*Day charge* equals 0.000613 plus 0.17 times *day minutes*." This is essentially a flat-rate model, billing 17 cents per minute for day use. Note from Figure 3.3 that the *R*-squared statistic is precisely 1, indicating a perfect linear relationship.

Since *day charge* is correlated perfectly with *day minutes*, we should eliminate one of the two variables. We do so, choosing arbitrarily to eliminate *day charge* and retain *day minutes*. Investigation of the *evening*, *night*, and *international* components reflected similar findings, and we thus also eliminate *evening charge*, *night charge*, and *international charge*. Note that had we proceeded to the modeling phase without first uncovering these correlations, our data mining and statistical models may have returned incoherent results, due in the multiple regression domain, for example, to multicollinearity. We have therefore reduced the number of predictors from 20 to 16 by eliminating redundant variables. A further benefit of doing so is that the dimensionality of the solution space is reduced, so that certain data mining algorithms may more efficiently find the globally optimal solution.

EXPLORING CATEGORICAL VARIABLES

One of the primary reasons for performing exploratory data analysis is to investigate the variables, look at histograms of the numeric variables, examine the distributions of the categorical variables, and explore the relationships among sets of variables. On the other hand, our overall objective for the data mining project as a whole (not just the EDA phase) is to develop a model of the type of customer likely to churn

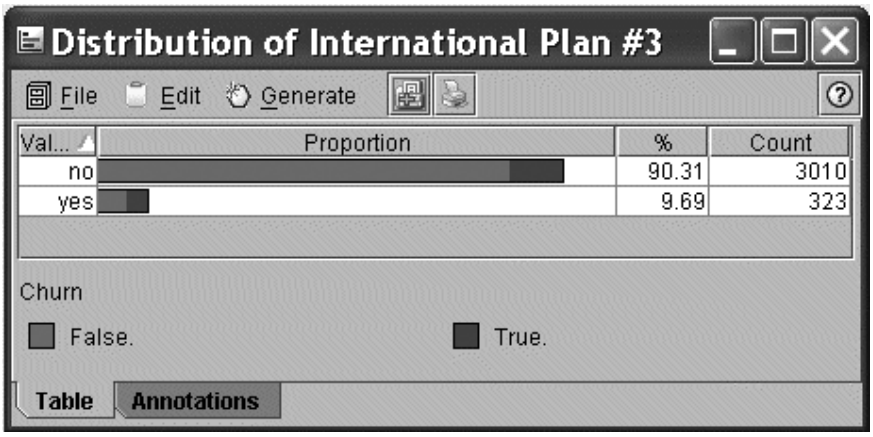

Figure 3.4 Comparison bar chart of churn proportions by International Plan participation.

(jump from your company's service to another company's service). Today's software packages allow us to become familiar with the variables while beginning to see which variables are associated with churn. In this way we can explore the data while keeping an eye on our overall goal. We begin by considering the categorical variables.

For example, Figure 3.4 shows a comparison of the proportion of churners (red) and nonchurners (blue) among customers who either had selected the International Plan (yes, 9.69% of customers) or had not selected it (no, 90.31% of customers). The graphic appears to indicate that a greater proportion of International Plan holders are churning, but it is difficult to be sure.

To increase the contrast and better discern whether the proportions differ, we can ask the software (in this case, Clementine) to provide same-size bars for each category. In Figure 3.5 we see a graph of the very same information as in Figure 3.4, except that the bar for the *yes* category has been stretched out to be the same length as the bar for the *no* category. This allows us to better discern whether the churn proportions differ

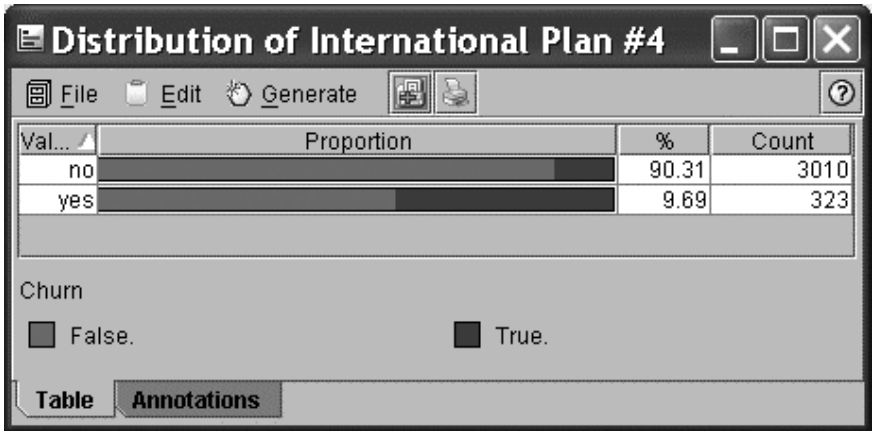

Figure 3.5 Comparison bar chart of churn proportions by International Plan participation, with equal bar length.

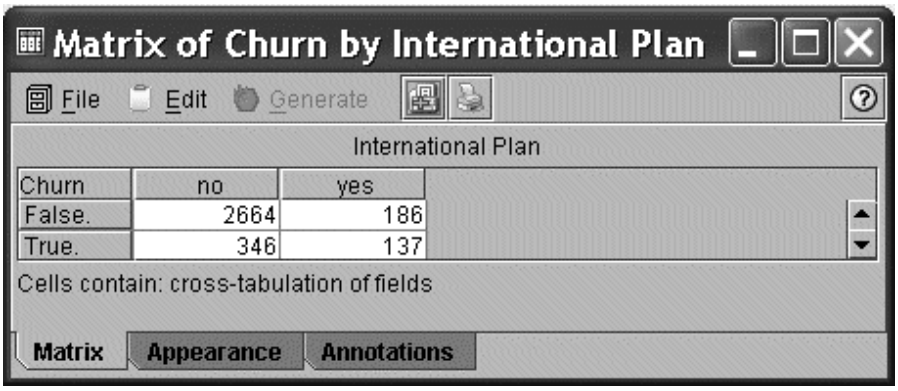

Figure 3.6 Cross-tabulation of International Plan with churn.

among the categories. Clearly, those who have selected the International Plan have a greater chance of leaving the company's service than do those who do not have the International Plan.

The graphics tell us that International Plan holders tend to churn more frequently, but they do not *quantify* the relationship. To quantify the relationship between International Plan holding and churning, we may use cross-tabulations, since both variables are categorical. Figure 3.6 shows Clementine's cross-tabulation. Note that the counts in the first column add up to the total number of nonselectors of the International Plan from Figure 3.4, $2664 + 346 = 3010$; similarly for the second column. The first row in Figure 3.6 shows the counts of those who did not churn, while the second row shows the counts of those who did churn. So the data set contains $346 + 137 = 483$ churners compared to $2664 + 186 = 2850$ nonchurners; that is, $483/(483 + 2850) = 14.5\%$ of the customers in this data set are churners.

Note that $137/(137 + 186) = 42.4\%$ of the International Plan holders churned, compared with only $346/(346 + 2664) = 11.5\%$ of those without the International Plan. Customers selecting the International Plan are more than three times as likely to leave the company's service than those without the plan.

This EDA on the International Plan has indicated that:

1. Perhaps we should investigate what it is about the International Plan that is inducing customers to leave!
2. We should expect that whatever data mining algorithms we use to predict churn, the model will probably include whether or not the customer selected the International Plan.

Let us now turn to the VoiceMail Plan. Figure 3.7 shows in a bar graph with equalized lengths that those who do not have the VoiceMail Plan are more likely to churn than those who do have the plan. (The numbers in the graph indicate proportions and counts of those who do and do not have the VoiceMail Plan, without reference to churning.)

Again, we may quantify this finding by using cross-tabulations, as in Figure 3.8. First of all, $842 + 80 = 922$ customers have the VoiceMail Plan, while $2008 + 403 = 2411$ do not. We then find that $403/2411 = 16.7\%$ of those without the VoiceMail Plan

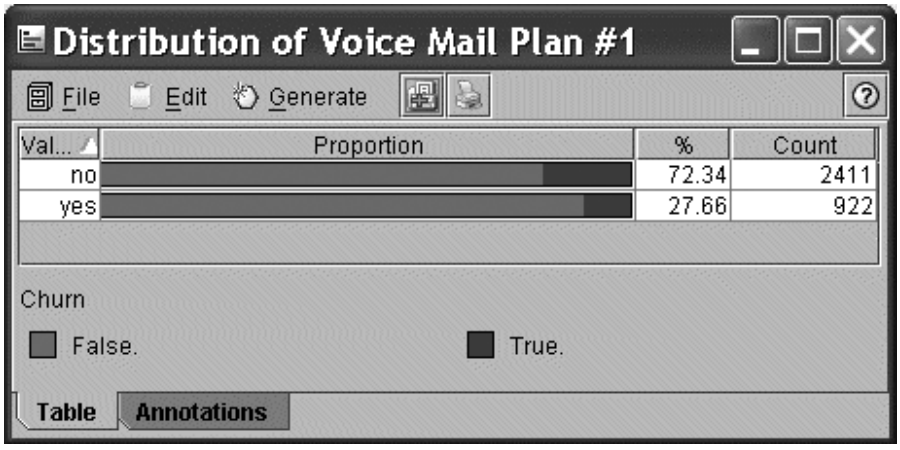

Figure 3.7 Those without the VoiceMail Plan are more likely to churn.

are churners, compared to $80/922 = 8.7\%$ of customers who do have the VoiceMail Plan. Thus, customers without the VoiceMail Plan are nearly twice as likely to churn as customers with the plan.

This EDA on the VoiceMail Plan has indicated that:

1. Perhaps we should enhance the VoiceMail Plan further or make it easier for customers to join it, as an instrument for increasing customer loyalty.
2. We should expect that whatever data mining algorithms we use to predict churn, the model will probably include whether or not the customer selected the VoiceMail Plan. Our confidence in this expectation is perhaps not quite as the high as that for the International Plan.

We may also explore the *two-way interactions* among categorical variables with respect to churn. For example, Figure 3.9 shows a pair of horizontal bar charts for

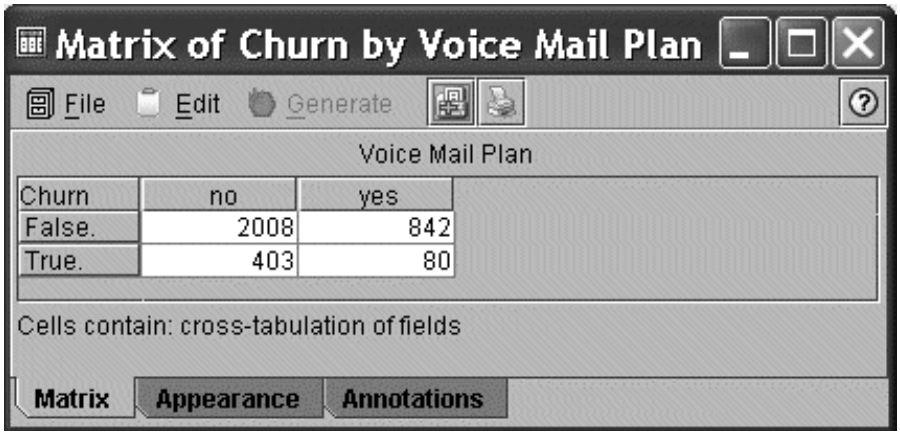

Figure 3.8 Cross-tabulation of VoiceMail Plan with churn.

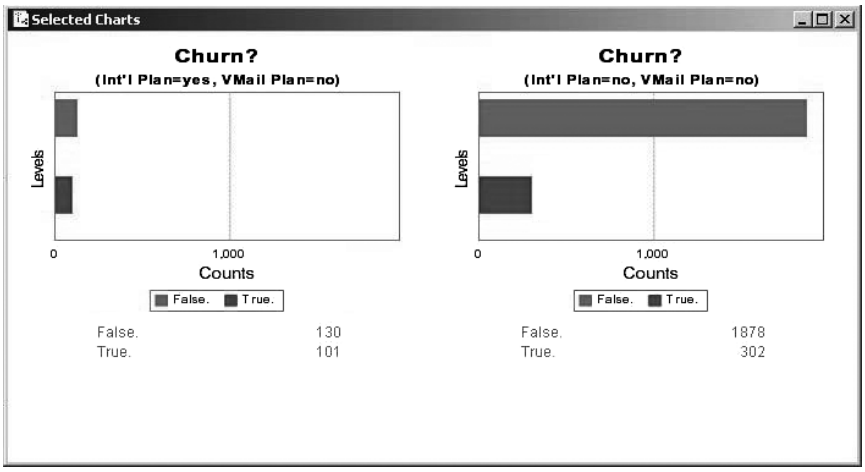

Figure 3.9 Bar charts of customers who churned, without VoiceMail Plan, subsetted by International Plan selection.

customers who did not select the VoiceMail Plan (*Vmail Plan* = *no*). The bar chart on the right contains customers who did not select the International Plan either, while the bar chart on the left contains customers who did select the International Plan.

Note that there are many more customers who have neither plan ($1878 + 302 = 2180$) than have the International Plan only ($130 + 101 = 231$). More important, among customers without the VoiceMail Plan, the proportion of churners is greater for those who do have the International Plan ($101/231 = 44\%$) than for those who don't ($302/2180 = 14\%$).

Next, Figure 3.10 shows a pair of horizontal bar charts for customers who did select the VoiceMail Plan (*Vmail Plan* = *yes*). There are many more customers who

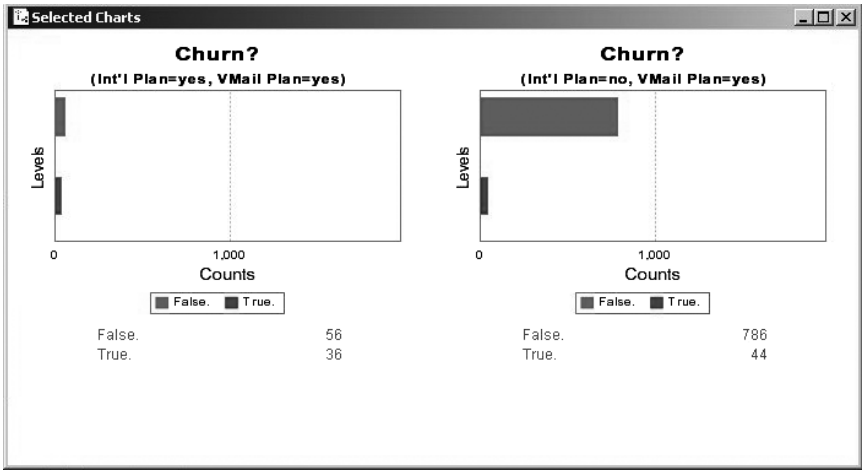

Figure 3.10 Bar charts of customers who churned, with VoiceMail Plan, subsetted by International Plan selection.

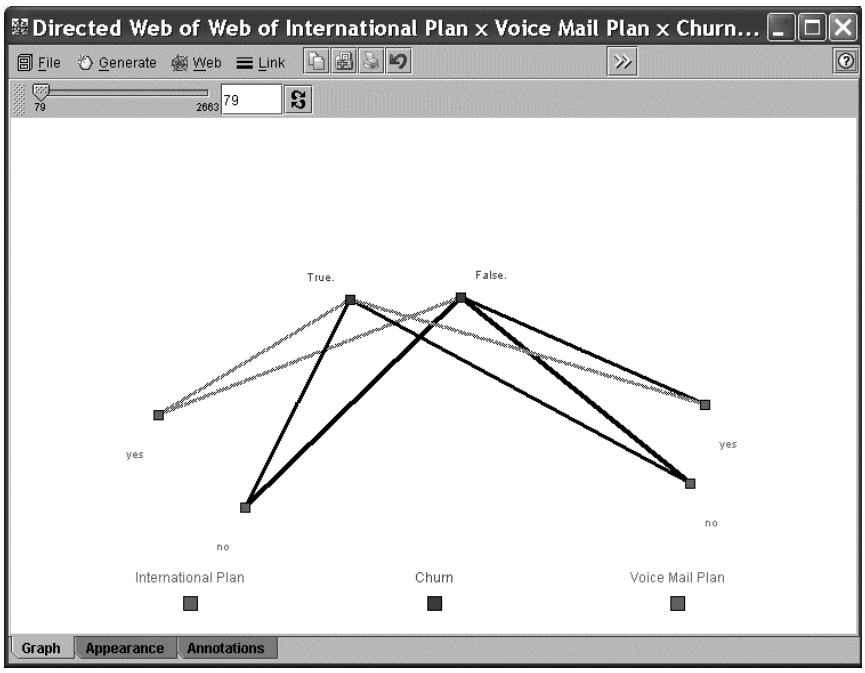

Figure 3.11 Directed web graph supports earlier findings.

have the VoiceMail Plan only ($786 + 44 = 830$) than those who have both plans ($56 + 36 = 92$). Again, however, among customers with the VoiceMail Plan, the proportion of churners is much greater for those who also select the International Plan ($36/92 = 39\%$) than for those who don't ($44/830 = 5\%$). Note that there is no interaction among the categorical variables. That is, International Plan holders have greater churn regardless of whether or not they are VoiceMail Plan adopters.

Finally, Figure 3.11 shows a Clementine *directed web graph* of the relationships between International Plan holders, VoiceMail Plan holders, and churners. Compare the edges (lines) connecting the VoiceMail Plan = Yes nodes to the Churn = True and Churn = False nodes. The edge connecting to the Churn = False node is heavier, indicating that a greater proportion of VoiceMail Plan holders will choose not to churn. This supports our earlier findings.

USING EDA TO UNCOVER ANOMALOUS FIELDS

Exploratory data analysis will sometimes uncover strange or anomalous records or fields which the earlier data cleaning phase may have missed. Consider, for example, the *area code* field in the present data set. Although the area codes contain numerals, they can also be used as categorical variables, since they can classify customers according to geographical location. We are intrigued by the fact that the area code field contains only three different values for all the records—408, 415, and 510—all three of which are in California, as shown by Figure 3.12.

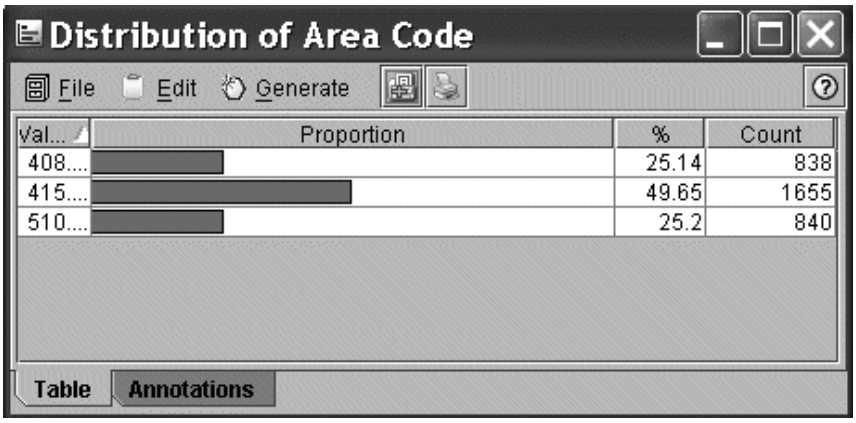

Figure 3.12 Only three area codes for all records.

Now, this would not be anomalous if the records indicated that the customers all lived in California. However, as shown in the cross-tabulation in Figure 3.13 (only up to Florida, to save space), the three area codes seem to be distributed more or less evenly across all the states and the District of Columbia. It is possible that domain experts might be able to explain this type of behavior, but it is also possible that the field just contains bad data.

We should therefore be wary of this area code field, perhaps going so far as not to include it as input to the data mining models in the next phase. On the other hand, it may be the *state* field that is in error. Either way, further communication

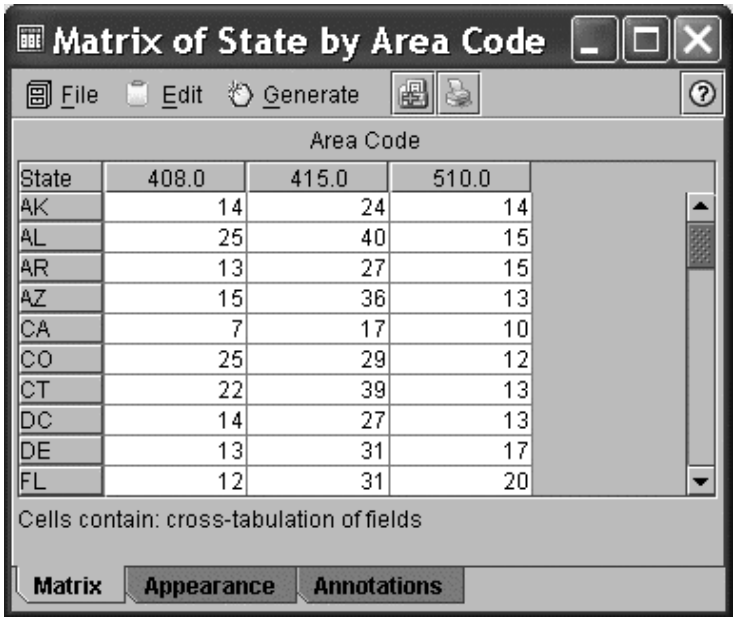

Figure 3.13 Anomaly: three area codes distributed across all 50 states.

with someone familiar with the data history, or a domain expert, is called for before inclusion of these variables in the data mining models.

EXPLORING NUMERICAL VARIABLES

Next, we turn to an exploration of the numerical predictive variables. We begin with numerical summary measures, including minimum and maximum; measures of center, such as mean, median, and mode; and measures of variability, such as standard deviation. Figure 3.14 shows these summary measures for some of our numerical variables. We see, for example, that the minimum *account length* is one month, the maximum is 243 months, and the mean and median are about the same, at around 101 months, which is an indication of symmetry. Notice that several variables show this evidence of symmetry, including *all* the *minutes*, *charge*, and *call* fields.

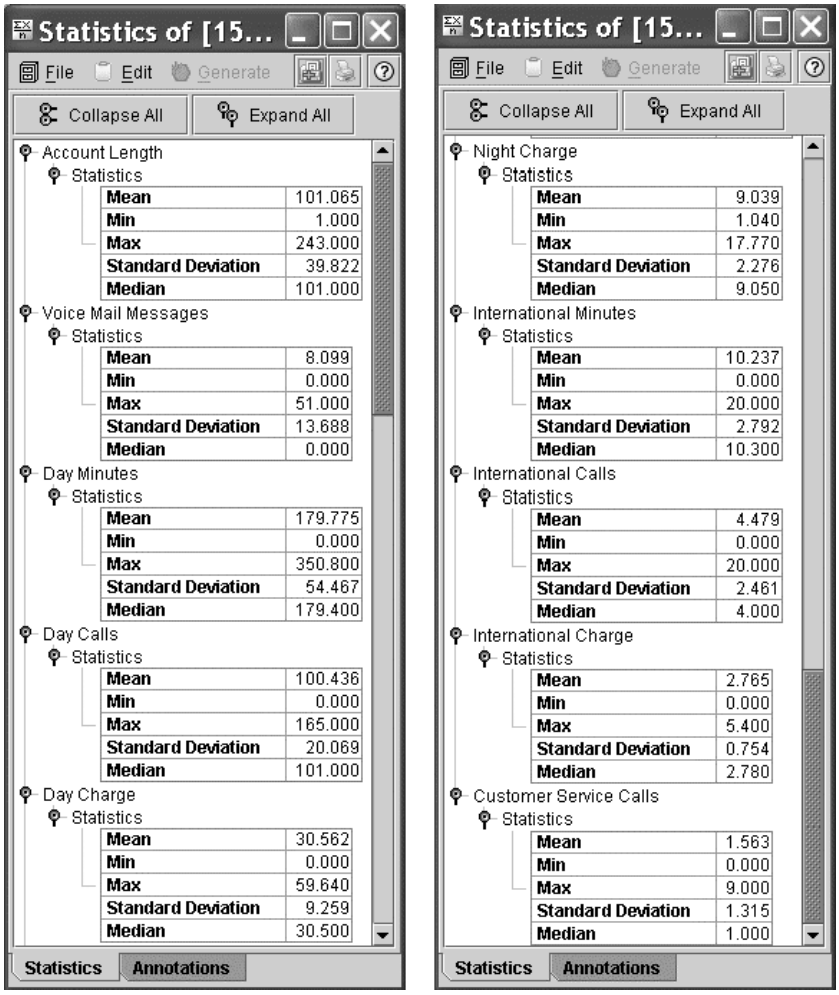

Figure 3.14 Summary statistics for several numerical variables.

Fields not showing evidence of symmetry include *voice mail messages* and *customer service calls*. The median for *voice mail messages* is zero, indicating that at least half of all customers had no voice mail messages. This results, of course, from fewer than half of the customers selecting the VoiceMail Plan, as we saw above. The mean of *customer service calls* (1.563) is greater than the median (1.0), indicating some right-skewness, as also indicated by the maximum number of customer service calls being nine.

As mentioned earlier, retaining correlated variables will, at best, overemphasize a certain predictive component at the expense of others, and at worst, cause instability in the model, leading to potentially nonsensical results. Therefore, we need to check for the correlation among our numerical variables. Figure 3.15 shows the correlations for two of the variables, *customer service calls* and *day charge*, with all of the other numerical variables. Note that all the correlations are shown as weak (this categorization is user-definable), except for the correlation between *day charge* and *day minutes*,

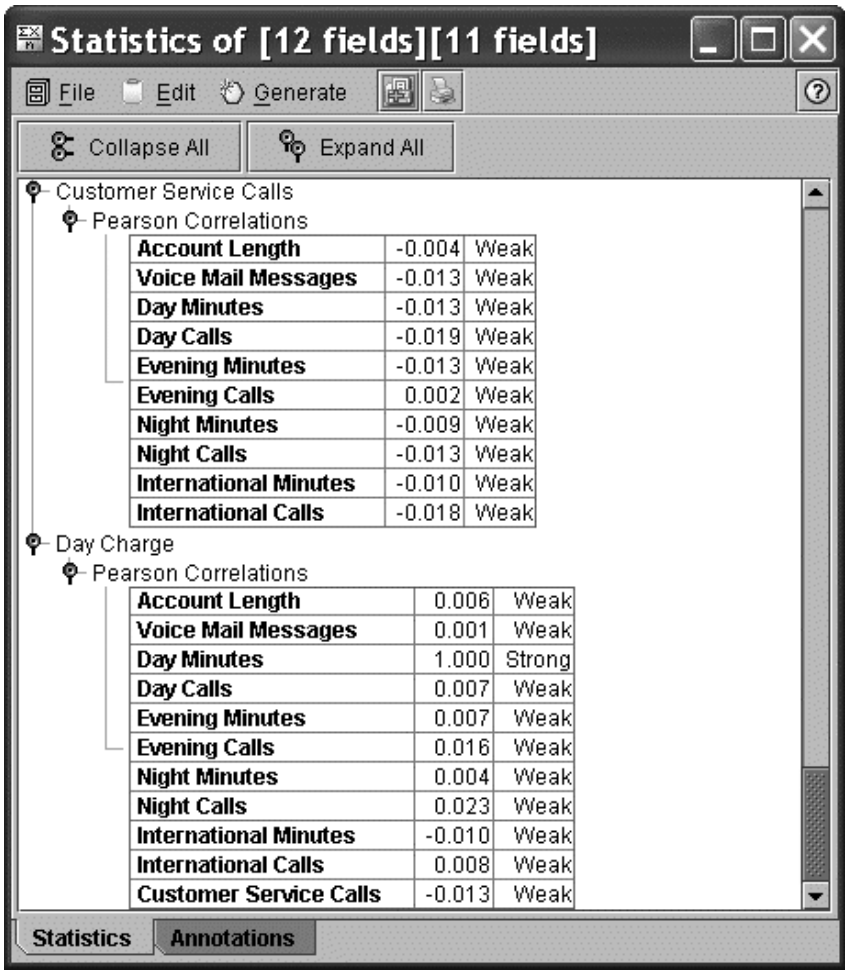

Figure 3.15 Correlations for *customer service calls* and *day charge*.

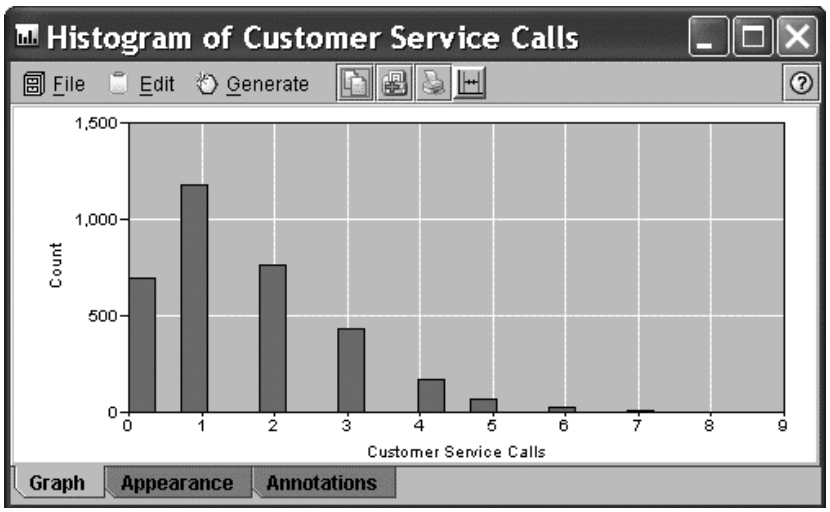

Figure 3.16 Histogram of customer service calls.

which is $r = 1.0$, the perfect linear relationship we discussed above. We checked for all pairwise correlations, and found all weak correlations once the *charge* fields were removed (not shown).

We turn next to graphical analysis of our numerical variables. We show three examples of histograms, which are useful for getting an overall look at the distribution of numerical variables, for the variable *customer service calls*. Figure 3.16 is a histogram of customer service calls, with no overlay, indicating that the distribution is right-skewed, with a mode at one call.

However, this gives us no indication of any relationship with churn, for which we must turn to Figure 3.17, the same histogram of customer service calls, this time with

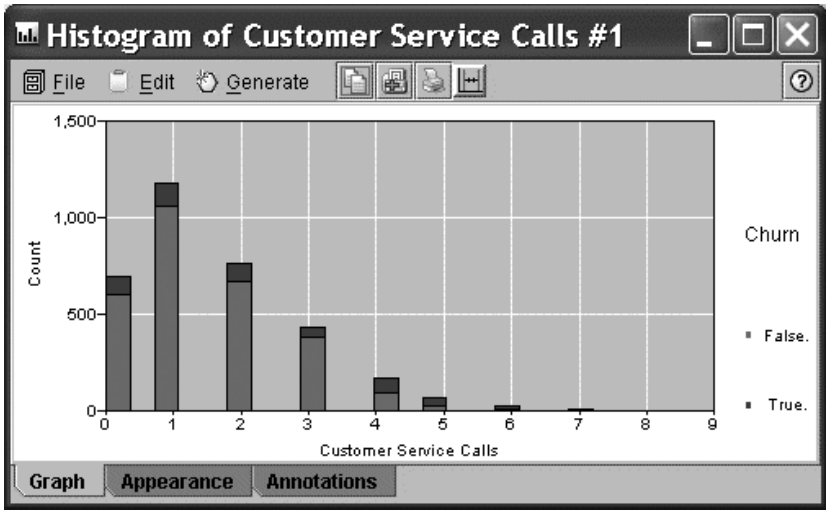

Figure 3.17 Histogram of customer service calls, with churn overlay.

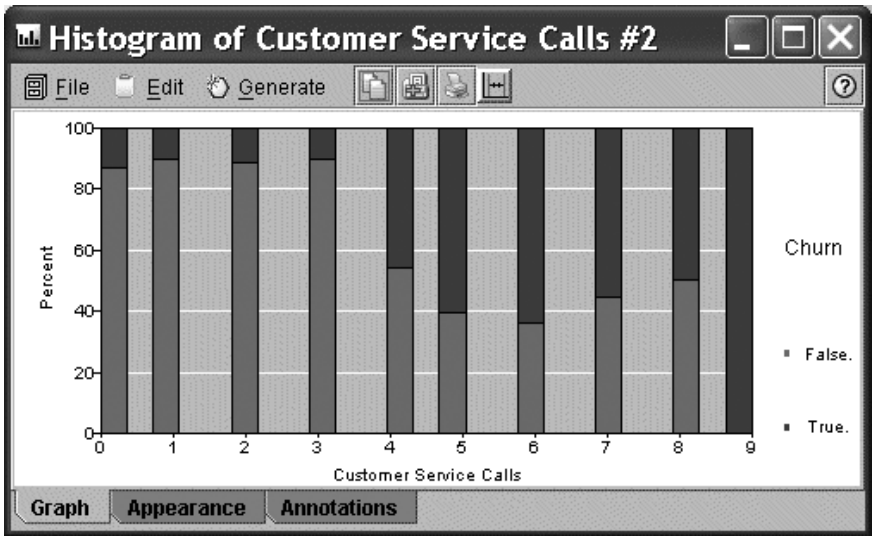

Figure 3.18 Normalized histogram of customer service calls, with churn overlay.

churn overlay. Figure 3.17 hints that the proportion of churn may be greater for higher numbers of customer service calls, but it is difficult to discern this result unequivocally. We therefore turn to a *normalized histogram*, where every rectangle has the same height and width, as shown in Figure 3.18. Note that the *proportions* of churners versus nonchurners in Figure 3.18 is exactly the same as in Figure 3.17; it is just that "stretching out" the rectangles that have low counts enables better definition and contrast. The pattern now becomes crystal clear. Customers who have called customer service three or fewer times have a markedly lower churn rate (dark part of the rectangle) than that of customers who have called customer service four or more times.

This EDA on the customer service calls has indicated that:

1. We should track carefully the number of customer service calls made by each customer. By the third call, specialized incentives should be offered to retain customer loyalty.
2. We should expect that whatever data mining algorithms we use to predict churn, the model will probably include the number of customer service calls made by the customer.

Examining Figure 3.19, we see that the normalized histogram of *day minutes* indicates that very high day users tend to churn at a higher rate. Therefore:

1. We should carefully track the number of day minutes used by each customer. As the number of day minutes passes 200, we should consider special incentives.
2. We should investigate why heavy day users are tempted to leave.
3. We should expect that our eventual data mining model will include *day minutes* as a predictor of churn.

Figure 3.20 shows a slight tendency for customers with higher *evening minutes* to churn. Based solely on the graphical evidence, however, we cannot conclude

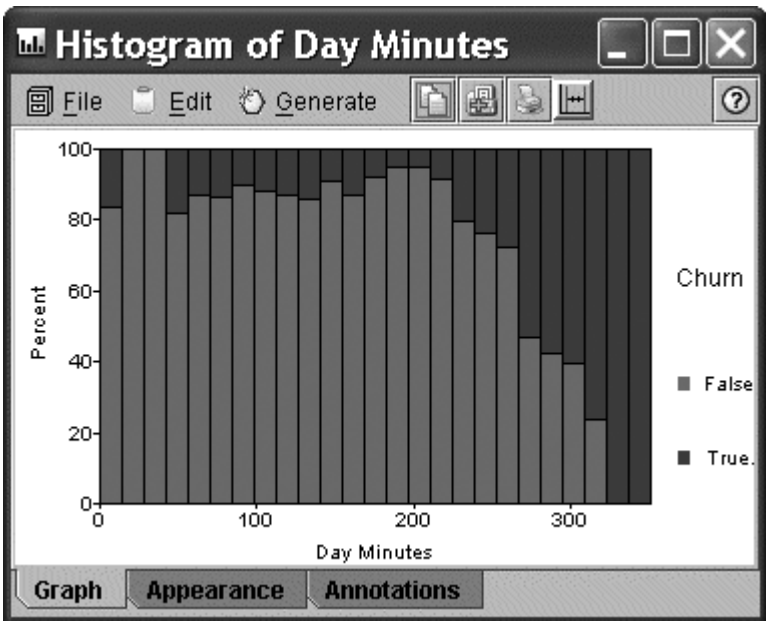

Figure 3.19 Customers with high day minutes tend to churn at a higher rate.

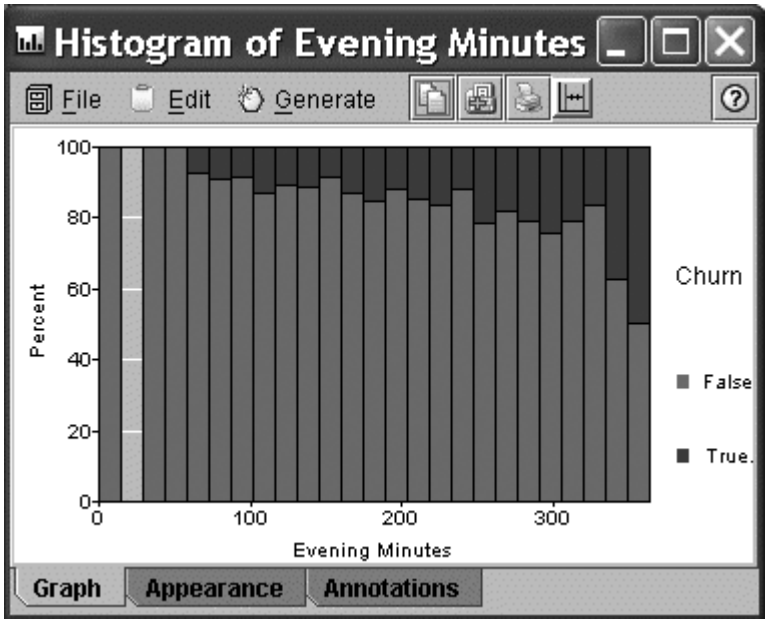

Figure 3.20 Slight tendency for customers with higher evening minutes to churn at a higher rate.

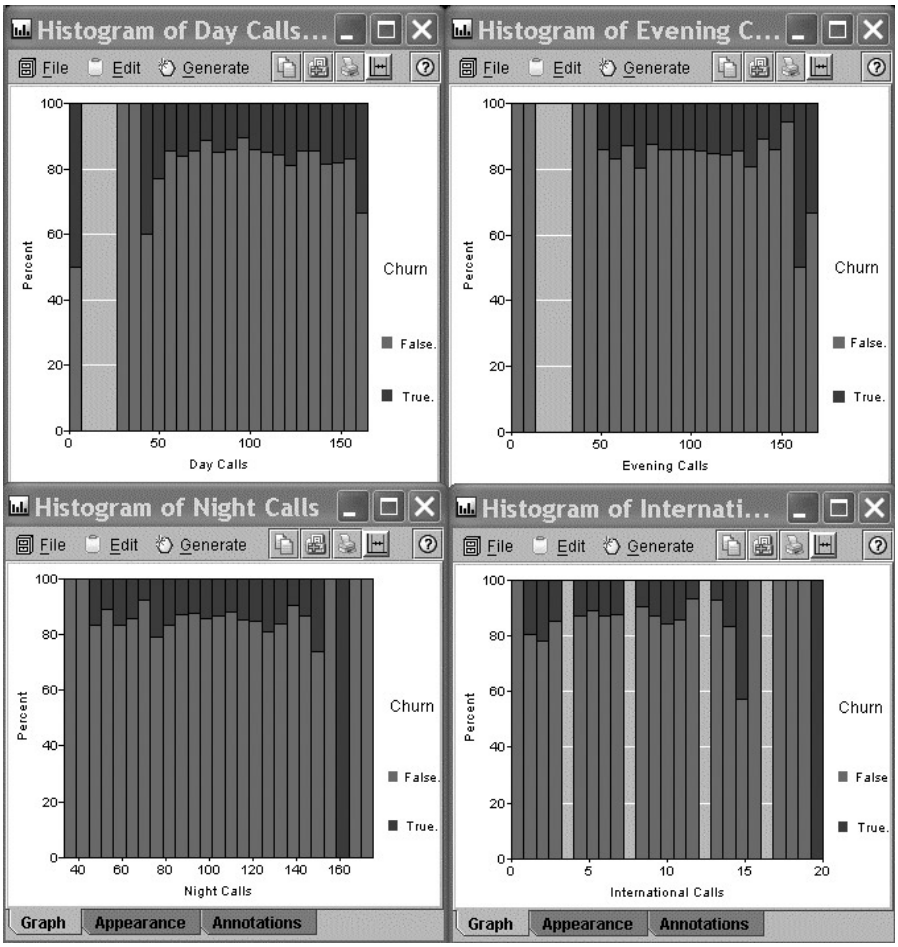

Figure 3.21 No association of churn with *day calls*, *evening calls*, *night calls*, or *international calls.*

beyond a reasonable doubt that such an effect exists. Therefore, we shall hold off on formulating policy recommendations on evening cell-phone use until our data mining models offer firmer evidence that the putative effect is in fact present.

Finally, Figures 3.21 and 3.22 indicate that there is no obvious association between churn and any of the remaining numerical variables in the data set. Figure 3.21 shows histograms of the four *calls* variables, *day*, *evening*, *night*, and *international calls*, with a churn overlay. Figure 3.22 shows histograms of *night minutes*, *international minutes*, *account length,* and *voice mail messages*, with a churn overlay. The high variability in churn proportions in the right tails of some of the histograms reflects the small sample counts in those regions.

Based on the lack of evident association between *churn* and the variables in Figures 3.21 and 3.22, we will not necessarily expect the data mining models to

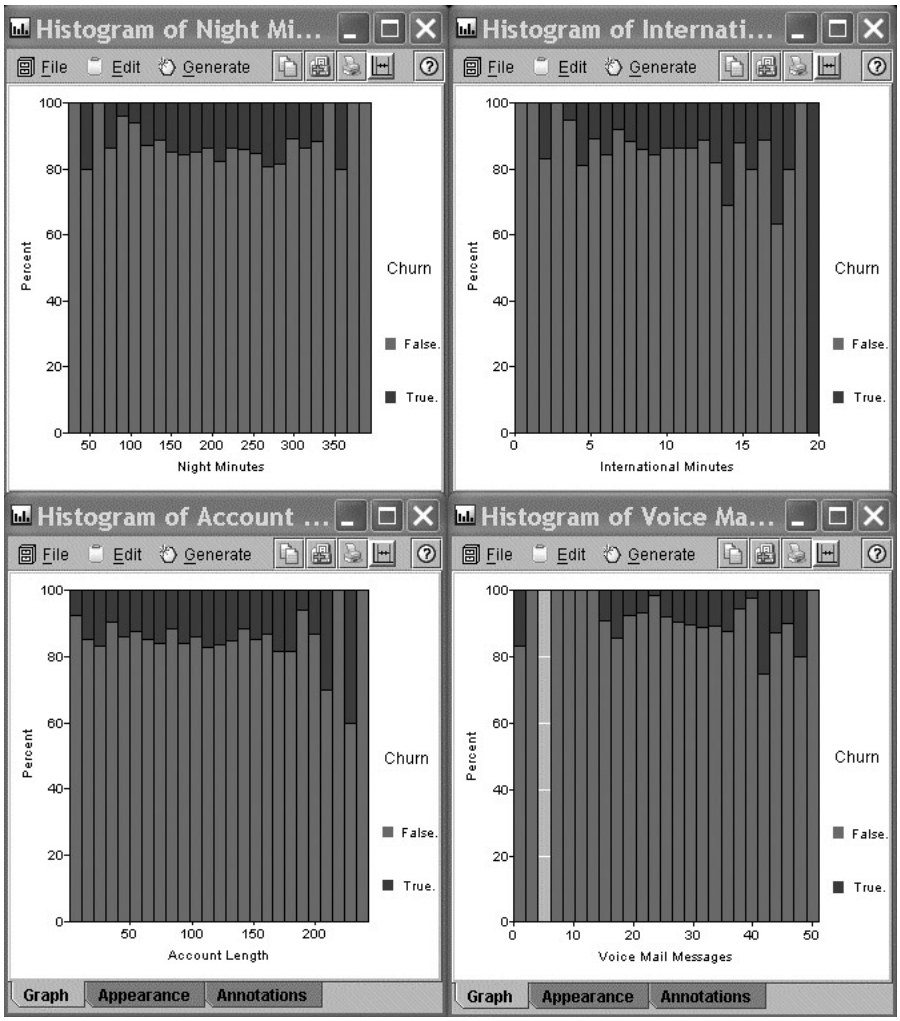

Figure 3.22 No association of churn with *night minutes, international minutes, account length*, or *voice mail messages*.

uncover valuable predictive information using these variables. We should, nevertheless, retain them as input variables for the data mining models. The reason for retaining these variables is that actionable associations may still exist for identifiable subsets of the records, and they may be involved in higher-dimension associations and interactions. In any case, unless there is a good reason (such as strong correlation) for eliminating the variable prior to modeling, we should probably allow the modeling process to identify which variables are predictive and which are not.

An exception to this situation is if there are so many fields that algorithm performance is degraded. In this case, one may consider setting aside temporarily variables

TABLE 3.1 Summary of Exploratory Findings Thus Far

Variable	Disposition
State	Anomalous. Omitted from model.
Account length	No obvious relation with churn, but retained.
Area code	Anomalous. Omitted from model.
Phone number	Surrogate for ID. Omitted from model.
International Plan	Predictive of churn. Retained.
VoiceMail Plan	Predictive of churn. Retained.
Number of voice mail messages	No obvious relation with churn, but retained.
Total day minutes	Predictive of churn. Retained.
Total day calls	No obvious relation with churn, but retained.
Total day charge	Function of *minutes*. Omitted from model.
Total evening minutes	May be predictive of churn. Retained.
Total evening calls	No obvious relation with churn, but retained.
Total evening charge	Function of *minutes*. Omitted from model.
Total night minutes	No obvious relation with churn, but retained.
Total night calls	No obvious relation with churn, but retained.
Total night charge	Function of *minutes*. Omitted from model.
Total international minutes	No obvious relation with churn, but retained.
Total international calls	No obvious relation with churn, but retained.
Total international charge	Function of *minutes*. Omitted from model.
Customer service calls	Predictive of churn. Retained.

with no obvious association with the target, until analysis with more promising variables is undertaken. Also in this case, dimension-reduction techniques should be applied, such as principal components analysis [2].

Table 3.1 summarizes our exploratory findings so far. We have examined each of the variables and have taken a preliminary look at their relationship with *churn*.

EXPLORING MULTIVARIATE RELATIONSHIPS

We turn next to an examination of possible multivariate associations of numerical variables with churn, using two- and three-dimensional scatter plots. Figure 3.23 is a scatter plot of *customer service calls* versus *day minutes* (note Clementine's incorrect reversing of this order in the plot title; the y-variable should always be the first named). Consider the partition shown in the scatter plot, which indicates a high-churn area in the upper left section of the graph and another high-churn area in the right of the graph. The high-churn area in the upper left section of the graph consists of customers who have a combination of a high number of customer service calls and a low number of day minutes used. Note that this group of customers could not have been identified had we restricted ourselves to univariate exploration (exploring variable by single variable). This is because of the *interaction* between the variables.

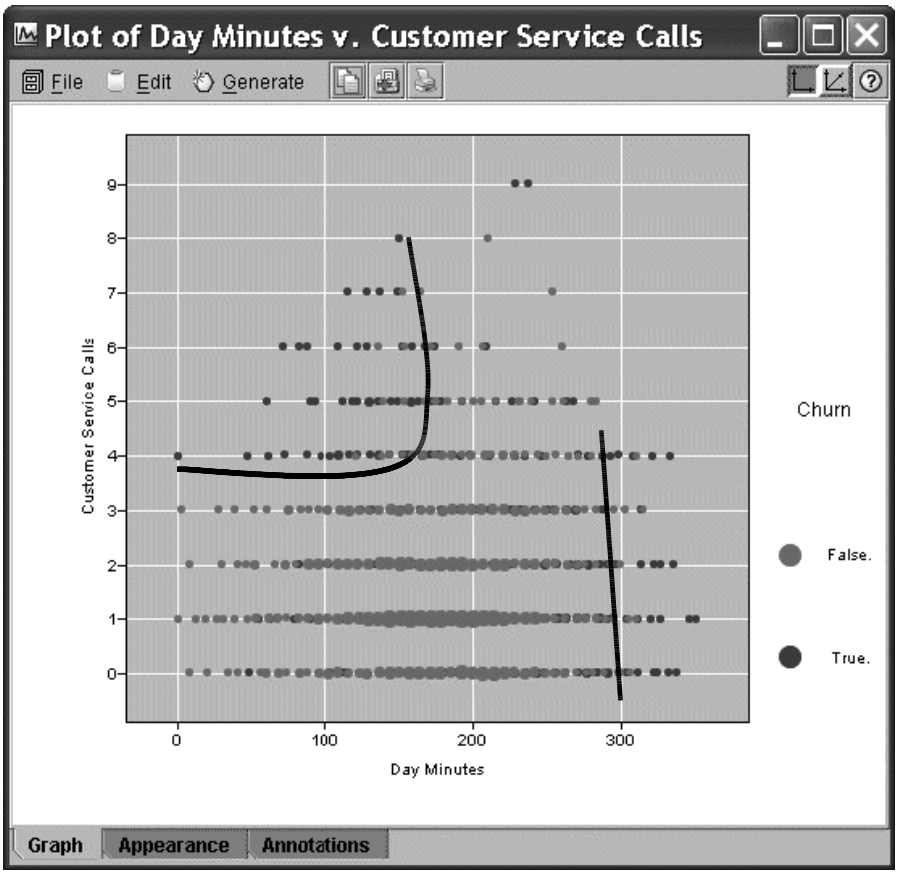

Figure 3.23 Scatter plot of *customer service calls* versus *day minutes*.

In general, customers with higher numbers of customer service calls tend to churn at a higher rate, as we learned earlier in the univariate analysis. However, Figure 3.23 shows that of these customers with high numbers of customer service calls, those who also have high day minutes are somewhat "protected" from this high churn rate. The customers in the upper right of the scatter plot exhibit a lower churn rate than that of those in the upper left.

Contrast this situation with the other high-churn area on the right (to the right of the straight line). Here, a higher churn rate is shown for those with high day minutes, *regardless* of the number of customer service calls, as indicated by the near-verticality of the partition line. In other words, these high-churn customers are the same ones as those identified in the univariate histogram in Figure 3.19.

Sometimes, three-dimensional scatter plots can be helpful as well. Figure 3.24 is an example of a plot of day minutes versus evening minutes versus customer service calls, with a churn overlay. The scroll buttons on the sides rotate the display so that the points may be examined in a three-dimensional environment.

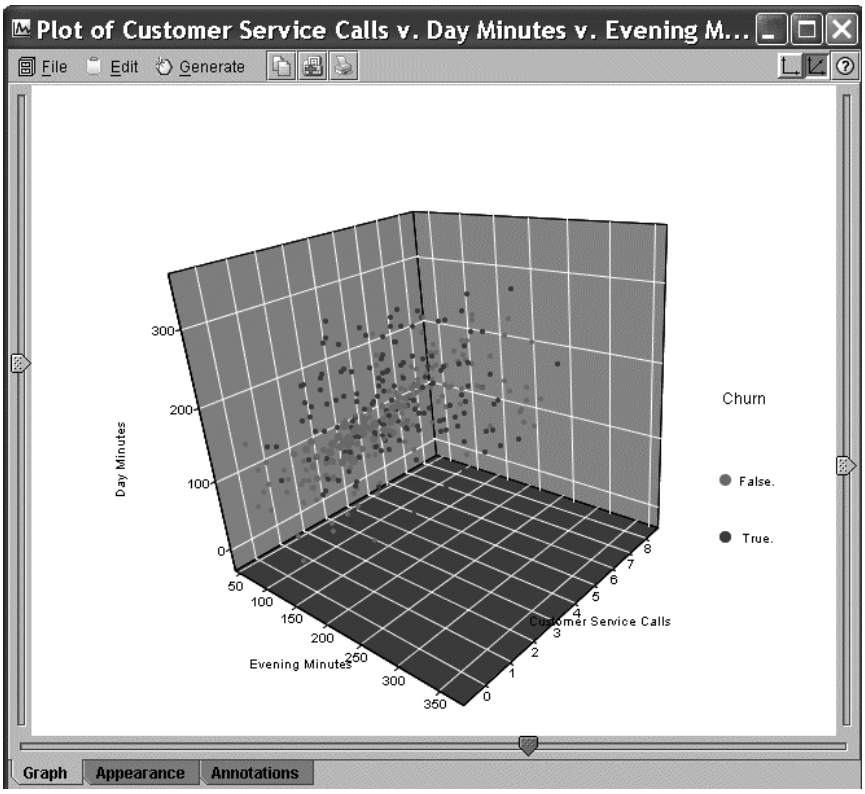

Figure 3.24 Three-dimensional scatter plot of *day minutes* versus *evening minutes* versus *customer service calls*, with a churn overlay.

SELECTING INTERESTING SUBSETS OF THE DATA FOR FURTHER INVESTIGATION

We may use scatter plots (or histograms) to identify interesting subsets of the data, in order to study these subsets more closely. In Figure 3.25 we see that customers with high *day minutes* and high *evening minutes* are more likely to churn. But how can we quantify this? Clementine allows the user to click and drag a select box around data points of interest, and select them for further investigation. Here we selected the records within the rectangular box in the upper right. (A better method would be to allow the user to select polygons besides rectangles.)

The *churn* distribution for this subset of records is shown in Figure 3.26. It turns out that over 43% of the customers who have both high *day minutes* and high *evening minutes* are churners. This is approximately three times the churn rate of the overall customer base in the data set. Therefore, it is recommended that we consider how we can develop strategies for keeping our heavy-use customers happy so that they do not leave the company's service, perhaps through discounting the higher levels of minutes used.

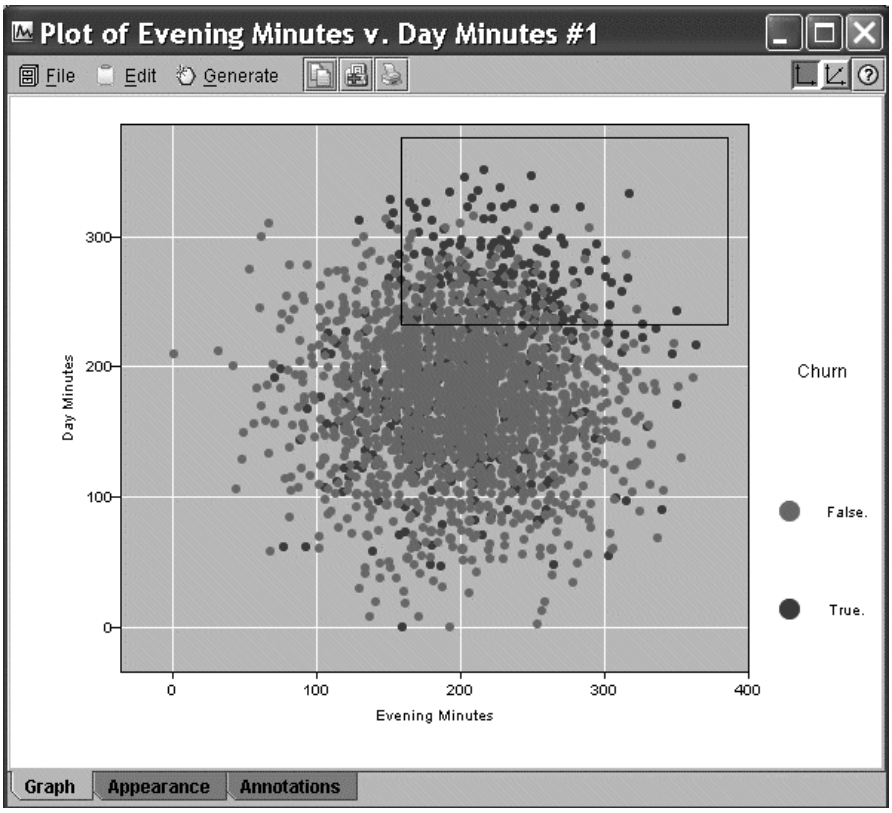

Figure 3.25 Selecting an interesting subset of records for further investigation.

BINNING

Binning (also called *banding*) refers to the categorization of numerical or categorical variables into a manageable set of classes which are convenient for analysis. For example, the number of *day minutes* could be categorized (binned) into three classes: *low*, *medium*, and *high*. The categorical variable *state* could be binned into

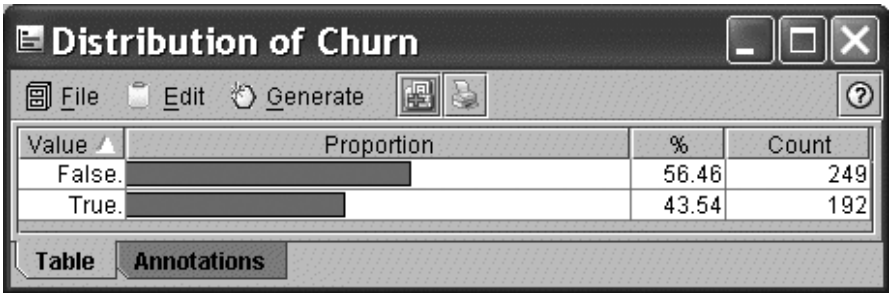

Figure 3.26 Over 43% of customers with high day and evening minutes churn.

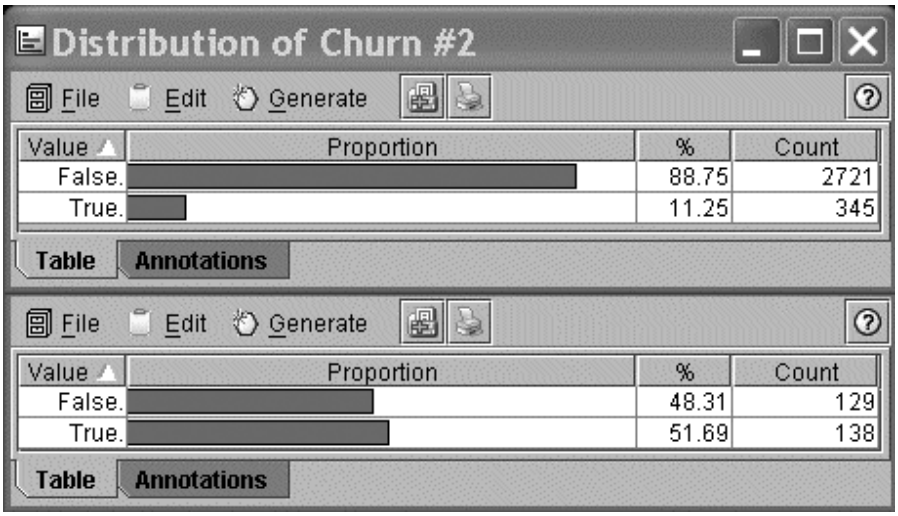

Figure 3.27 Churn rate for customers with low (top) and high (bottom) customer service calls.

a new variable, *region*, where California, Oregon, Washington, Alaska, and Hawaii would be put in the *Pacific* category, and so on. Properly speaking, binning is a data preparation activity as well as an exploratory activity.

There are various strategies for binning numerical variables. One approach is to make the classes of equal width, analogous to equal-width histograms. Another approach is to try to equalize the number of records in each class. You may consider yet another approach, which attempts to partition the data set into identifiable groups of records, which, with respect to the target variable, have behavior similar to that for other records in the same class.

For example, recall Figure 3.18, where we saw that customers with fewer than four calls to customer service had a lower churn rate than that of customers who had four or more calls to customer service. We may therefore decide to bin the *customer service calls* variable into two classes, *low* and *high*. Figure 3.27 shows that the churn rate for customers with a low number of calls to customer service is 11.25%, whereas the churn rate for customers with a high number of calls to customer service is 51.69%, more than four times higher.

SUMMARY

Let us consider some of the insights we have gained into the *churn* data set through the use of exploratory data analysis.

- The four *charge* fields are linear functions of the *minute* fields, and should be omitted.
- The *area code* field and/or the *state* field are anomalous, and should be omitted until further clarification is obtained.

- The correlations among the remaining predictor variables are weak, allowing us to retain them all for any data mining model.

Insights with respect to *churn*:

- Customers with the International Plan tend to churn more frequently.
- Customers with the VoiceMail Plan tend to churn less frequently.
- Customers with four or more *customer service calls* churn more than four times as often as do the other customers.
- Customers with high *day minutes* and *evening minutes* tend to churn at a higher rate than do the other customers.
- Customers with both high *day minutes* and high *evening minutes* churn about three times more than do the other customers.
- Customers with low *day minutes* and high *customer service calls* churn at a higher rate than that of the other customers.
- There is no obvious association of *churn* with the variables *day calls*, *evening calls*, *night calls*, *international calls*, *night minutes*, *international minutes*, *account length*, or *voice mail messages*.

Note that we have not applied any data mining algorithms yet on this data set, such as decision tree or neural network algorithms. Yet we have gained considerable insight into the attributes that are associated with customers leaving the company, simply by careful application of exploratory data analysis. These insights can easily be formulated into actionable recommendations, so that the company can take action to lower the churn rate among its customer base.

REFERENCES

1. C. L. Blake and C. J. Merz, *Churn Data Set*, UCI Repository of Machine Learning Databases, `http://www.ics.uci.edu/~mlearn/MLRepository.html`. University of California, Department of Information and Computer Science, Irvine, CA, 1998.
2. Daniel Larose, *Data Mining Methods and Models*, Wiley-Interscience, Hoboken, NJ (to appear 2005).

EXERCISES

1. Describe the possible consequences of allowing correlated variables to remain in the model.

a. How can we determine whether correlation exists among our variables?

b. What steps can we take to remedy the situation? Apart from the methods described in the text, think of some creative ways of dealing with correlated variables.

c. How might we investigate correlation among categorical variables?

2. For each of the following descriptive methods, state whether it may be applied to categorical data, continuous numerical data, or both.

a. Bar charts

b. Histograms

c. Summary statistics