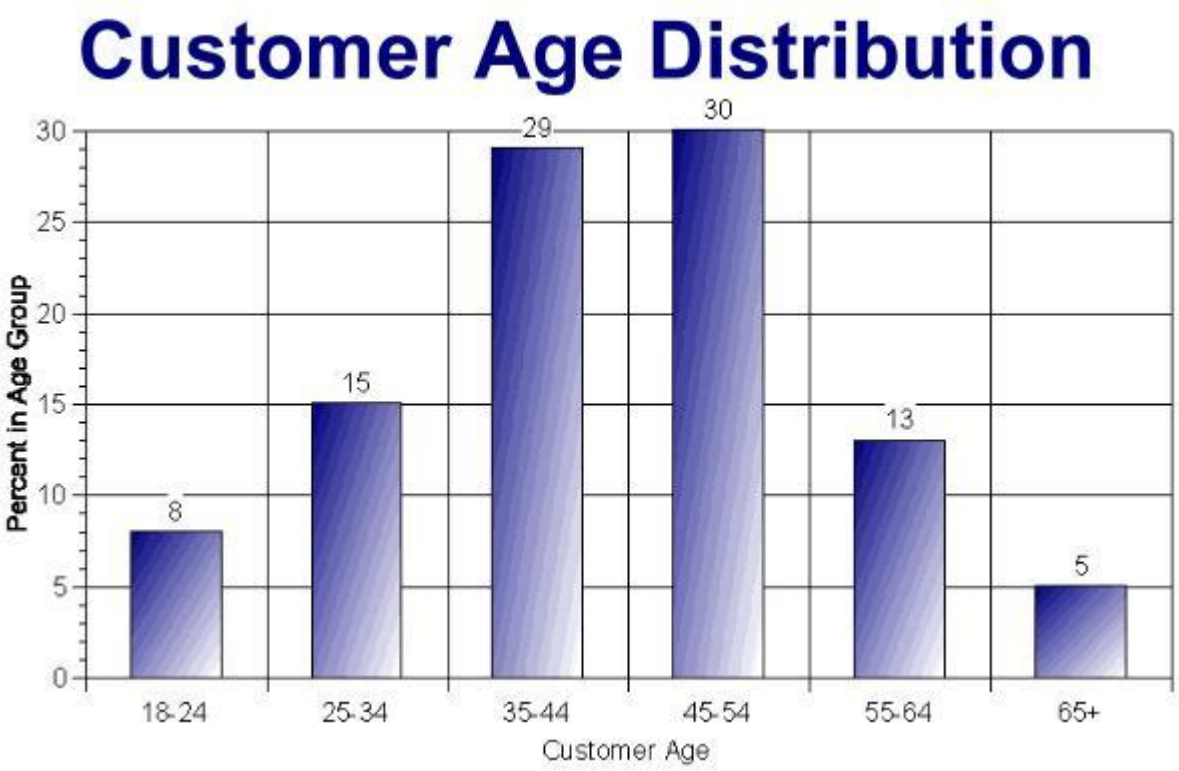
Basic Introduction to Analytic Techniques

The following discussion compares various general approaches to data analysis, ranging from simple single-variable profiling and two-way crosstabulations (crosstabs) to multivariate techniques such as regression and CHAID.

Simple Profiling

The simplest way to examine data is to profile individual variables from a database. This is accomplished with frequency distributions. The following bar graph shows a simple profile of the age distribution of customers:



We can see that the two largest age groups, 35-44 and 45-54, account for nearly 60% of customers; and that the single largest group, ages 45-54, accounts for 30% of all customers. And of course, we can produce similar graphs for other variables, such as income, family size,

home value, education, and so forth.

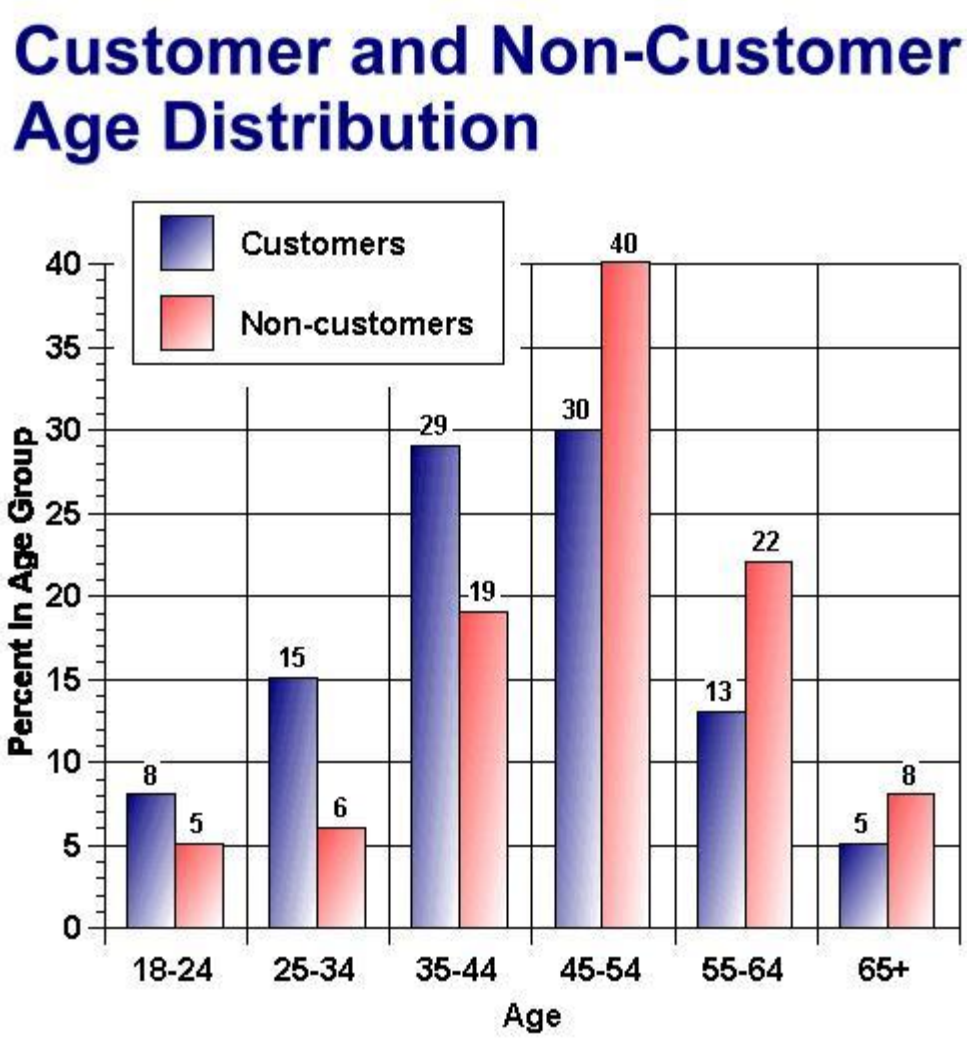
While such univariate (i.e., one variable at a time) profiles may be useful for simple tasks such as verifying the integrity of variables in a database and getting a basic idea of the range and frequency of the variable's values, it is unwise to base targeting definitions or other strategic decisions on them. In this case, what we really want to know is this: How do customers differ from non-customers?

Crosstabulation

In fact, the main goal of nearly all data mining should be to identify valid and reliable patterns which are predictive of similarities within, and differences between, sub-groups represented in the database. Simple profiling of only one population (e.g., customers; high-value customers; bad credit risks; people who are taking a new, experimental drug; etc.) has severe limitations, and can even lead to a wrong or dangerous conclusion.

This is most apparent, of course, in situations such as clinical drug trials, where the inclusion of a control or placebo group, as well as one or more treatment groups, is a standard procedure. However, this also applies to virtually any data mining situation where we are trying to find useful patterns of information, such as target marketing, manufacturing defect analysis, credit or fraud analysis, response-to-promotion analysis, etc.

We almost always want to know how the target market differs from the non-target; how the loyal or high-spending customer differs from the non-loyal or low-spending customer or the non-customer; how the good credit risk differs from the poor one; or, in general, how the more successful or preferred outcome differs from the less successful or less preferred outcome. The next bar graph, which is based on a crosstabulation of two variables (age and customer status), demonstrates how much more useful this understanding of differences really is:



Now we can see that although the 45-54 age group contains nearly one-third of our customers, it contains fully 40% of the non-customers. For targeting purposes, if we are trying to identify demographic characteristics of non-customers who resemble our current franchise (who have already demonstrated an affinity for our brand), then it may be better to focus on the 35-44 age group and, secondarily, on the 25-34 age group, since these latter two age groups both show much higher penetration of customers within their age group than the 45-54 age group, while also accounting for 44% of all customers. This is a very different conclusion from that drawn from the previous simple profiling, which would have been misleading.

But crosstabs or simple spreadsheets are also somewhat limited in their information value and actionability. In reality, we usually want to examine several variables at once, especially when we are trying to explain or predict the influences of some variables on "outcome" variable of interest. E.g., if we want to try to untangle the influences of several demographic characteristics on response to a promotional effort, so that we can understand what demographics have the greatest influence, then we need to move up one level of analytic complexity to multivariate techniques such as regression modeling.

Regression

Linear regression and **logistic regression** are two examples of multivariate modeling techniques, of which there are also many others. Linear regression is useful when we are using several variables to predict the values of a continuous dependent variable, such as customer value in dollars. We can include various predictor variables such as age, income, family size, education, etc., and then use linear regression to tell us the unique influence of each predictor on customer dollar value, controlling for the influence of all the other predictors.

Thus, individual cross-tabs may show that both age and income seem quite important predictors of customer value, and that education is slightly important, but family size is not important. But linear regression can show us the unique importance of each predictor by examining all of the predictors' simultaneous influences on the dependent variable (customer value). So regression is a more powerful way of sorting out multiple influences than the eyeballing of output from a series of two-way crosstabs.

Regression generates exact coefficients for each predictor, and shows us what proportion of the variability of the dependent variable (customer value) is uniquely explained by each individual predictor. This makes it possible to build a predictive model which has predictor coefficients that can be used to "score" the records in a prospect file. This means that we can rank prospects in terms of what their likely value would be if they became customers, based on their demographic characteristics. Armed with this knowledge, our targeting will be much more successful, less costly and more profitable.

When the dependent variable we are trying to predict has only two values (e.g., responder vs. non-responder to a promotional mailing), then we can use a special type of regression called Binary Logistic Regression. As with Linear Regression, we generate coefficients for each predictor, and these coefficients can be used to score a prospect file to determine the best prospects for a promotional mailing. There are other variations of regression-type techniques, such as Curvilinear Regression and Loglinear Analysis, but we will not go into them here. Our intent is simply to point out that multivariate analysis is a much more powerful and useful data mining technique than either univariate or bivariate analysis.

Regression-type techniques also have limitations, however. For example, it is difficult to understand the influence of complex interactions among predictor variables on the dependent variable. (E.g., does income influence customer value differently for older prospects than it does for younger ones?) In order to clarify these potentially complex interactive relationships, we would typically use other methods, either instead of regression or in addition to it. Examples of these techniques are listed on our **Marketing Analytics page**. These might include techniques such as Classification and Regression Trees (**CART**), Chi-square Automatic Interaction Detection (**CHAID**), **Neural Networks** or **Genetic Programming**, etc. We will not attempt to cover all these other techniques here. Instead, we will select one of the more easily understood techniques (CHAID) for additional discussion. Please visit our **CHAID case study** page to see an application of CHAID predictive segmentation modeling.