7 Analytic File Creation

One of the common misconceptions in applied business analytics is that having the requisite data readily available marks the beginning of data crunching. Under most circumstances, having the data in-hand is merely the beginning of the process of creating an analyzable data file, rather than the beginning of data analyses. The reason for that is that, as discussed in Chapter 1, most data used in marketing analytics (as well as other aspects of business) is a by-product of the electronic transaction processing—in other words, the data layout, formatting and other characteristics reflect the inner-workings of those systems, rather than the demands of the ensuing analyses. Hence, before meaningful analysis can be initiated, an often significant amount of data engineering has to take place.

Taking the steps necessary to convert not-directly-analyzable raw data into, a still raw, but analyzable data file constitutes the creation of an analytic file. The extent of the required re-engineering will vary depending on the interplay between the characteristics of the original data and the informational demands at hand, but there are a number of specific considerations that needs to be addressed almost universally—those are discussed in the ensuing chapter.

Data Gathering

Data capture can be either *incidental* or *purposeful*. Data is collected incidentally when it is a by-product of ongoing business operations. Its capture is, to a large degree, a result of technological progress. For instance, stores equipped with barcode scanners capture product purchase data because the process of barcode scanning generates an electronic record, each time it occurs. Similarly, an online purchase creates a transaction record, just as an electronically placed manufacturer order generates a similar record. These data are generated each time a particular system is used,[1] even if there is no interest in data as an outcome.

On the other hand, the capture of other data types requires special effort. For instance, to get a valid and a representative view of buyer satisfaction or to uncover the attitudinal drivers of brand choice, a firm usually has to field a consumer survey. Such data is collected purposefully since its capture requires special provisions that are only put in place for the explicit purpose of capturing that particular information.[2] Thus in contrast to the incidental data capture, the gathering of the purposeful data is usually independent of the firm's operational systems. Under most circumstances, these data are only collected if they are to be used for a specific purpose.

The data acquired from outside data vendors is a little more difficult to classify. The commonly used geodemographics exhibit the characteristics of both the incidental and

purposeful data. Being derived from the U.S. Census information makes it incidental, while being targeted at a relatively specific goal makes it purposeful. In the end, however, it is more reasonable to consider it purposeful because its creation requires an additional effort and if its sales ceased (due to a lack of demand), so would its creation.

The vast majority of the core data is incidental, as is some of the causal data—because of that, it also tends to be organized around events, rather than customers (see Figures 6.3 and 6.4). This is an important point to keep in mind while considering adding some additional, i.e., purposeful, data to the database. Whether the added data comes from an outside data vendor/aggregator or from a direct data capture (such as a survey), its coverage—defined as the % of all database customers for whom the specific purposeful data was added—will almost always be limited to only a subset of all customers in the database.[3] In order to assure the projectability of sample-based database appends, the originally event-centric core data needs to be reorganized into customer-centric matrix to allow for proper coverage and/or bias assessment, or the selection of an appropriate sample.

Analytic Dataset Creation

Some databases come "fully loaded" with a rich assortment of core and causal data, all organized in a customer-centric fashion and ready to be taken to the proverbial "next level." A more common scenario, however, is a database in need of additional, usually causal appends, and proper organization. In such cases it is more cost and time effective to create a database extract-based analytical dataset.

An *analytical dataset* is a subset of the database that has been enriched with the appropriate outside (to the database) data, properly organized and cleansed. In the data sense, it is the starting point for the database analytical process described in this book. As previously mentioned, its creation involves three separate steps:

1. Extracting a subset of the database.
2. Enriching the database extract.
3. Organizing and cleansing.

Extracting a Subset of the Database

The most common database type today is one built around transactional information. The majority of larger retailers keep track of their transactions, as do most manufacturers. The resultant data reservoirs are enormous—the previously mentioned Walmart database has been estimated to contain roughly 435 terabytes of data at the time of this book's writing. And although most other corporate databases are considerably smaller, quite a few of them are multi-terabyte systems, which can easily translate into tens of millions or even billions of records.

Although such large volumes of data can be helpful in assuring reporting accuracy, it can be a hindrance in more advanced statistical modeling. The first reason behind it is obvious and it requires little explanation: It is quite difficult and time-consuming to manipulate such large files, particularly while conducting exploratory data analysis involving a large number of individual metrics and tests. The second reason is somewhat less obvious, but it has more direct impact on the validity of findings: Large sample sizes have a potentially skewing effect on the robustness of statistical parameters and tests. In

view of the potentially weighty implications of that dependence, a more in-depth explanation seems warranted.

Central to the estimation of many statistical parameters is the notion of *standard error*, which is an estimated level of imprecision of a particular statistic. In the computational sense, there are multiple methods of calculating standard errors, depending on the type of statistic used which in turn reflects different possible applications of this concept. In the realm of database analytics, probably the most frequent application of the notion of standard error takes place in the context of sample mean values estimation. In this case, the standard error is computed simply by dividing the sample standard deviation by the square root of the sample size. Interpretation-wise, the larger the sample size, the smaller the standard error. And that is the crux of the problem.

Standard error in and out of itself is of little interest to business analysts—however, it is one of the key inputs into a frequently used *statistical significance testing*,[4] the goal of which is to determine if the observed differences between means (such as the difference in the purchase rate between treated and control groups) are factual or spurious. As the standard error estimate decreases (due to a large sample size), the ever-smaller differences are deemed "statistically significant," which often leads the less experienced analysts to conclude that the otherwise trivially small differences (such as between treated and control groups in an impact experiment) represent meaningful practical findings (such as a positive treatment-attributable incrementality, in the case of the said experiment). In other words, excessively large sample size artificially deflates the size of the standard error estimates which in turn increase the likelihood of "false positive" findings—i.e., ascribing a factual status to trivially small, spurious differences.

The practical consequences of this dependence can be considerable—almost disproportionately large given the somewhat obtuse nature of the concept. For example, an organization might be led to believe that the initial tests of the contemplated strategic initiatives, such as pricing policy or promotional mix allocation changes, are encouraging enough to warrant a full scale commitment, when in fact the analysis (when properly executed) might not support such conclusions. Hence this seemingly trivial notion might be the database analytics' version of the Butterfly Effect,[5] which is why it demands a deeper treatment, presented later.

But for now, let's concentrate on the delineation of effective database extract selection rules, in view of the inappropriateness of using an un-pruned database universe as the basis for analysis.

Extract Selection Rules

The most important considerations in selecting a subset of an entire database are *representativeness* and *sizing*. The former speaks to the degree of compositional similarity between the database and the extract, while the latter spells out the minimum required number of records. Though different in terms of their focus, the two are highly interconnected. An extract-wide sample size is in a large part determined by the composition of the sample, specifically, the number of individual sub-segments and the nesting structure (i.e., how many tiers of sub-segments, or segments within segments, are there). The extract selection rules, therefore, should be framed in the context of the expected representativeness and sizing of the contemplated sample.

The often recommended—and used—*random selection* is actually rarely the best approach to take in selecting a subset of a customer/prospect database. The reason behind

that counterintuitive statement is that this otherwise convenient approach is likely to lead to over-sampling of large customer/prospect groupings and under-sampling of the small ones. This is particularly the case when it comes to the very best customers, who usually comprise a relatively small proportion of the total customer base (see the Pareto's Principle, often referred to as the "80–20" rule). As such, the best customers are likely to be under-represented in a randomly-drawn sample, which would obviously inhibit a deeper understanding of their behaviors (not to mention contributing to a skewed interpretation of the overall findings).

To mitigate the possibility of such undesirable outcomes, the *stratified sampling* scheme can be used instead. This sampling technique offers a higher likelihood of bringing about the typically desired random selection, by building the selection logic around appropriately defined customer segments.[6] The specific stratified customer database extract selection steps are outlined below:

- Step 1: Explicitly describe the content of the customer database by identifying the following:
  - customer clusters (e.g., spending or lifetime value segments, etc.).
  - descriptive variables and their quality (i.e., % of coverage).
  - longitudinal depth of purchase data (i.e., how far back does the data go?).
- Step 2: Identify the most disaggregate customer cluster, i.e., what is the most narrowly-defined group of customers to be used as the focus of analysis?
- Step 3: Flag customer records in accordance with cluster membership and determine the number of records per cluster.
- Step 4: Select a random sample of 500–1,000 customers from each of the clusters.
- Step 5: Contrast the profile of the sample extract with the parent population by comparing the means and distributions of the descriptive variables outlined in Step 1. If differences exceed 1 standard deviation on the key variables' means and/or distributions are significantly different, re-sample.

An important consideration governing the appropriateness of the resultant sample is its analytic adequacy. In other words, will it support the most disaggregate, in terms of sample composition homogeneity, analysis? The general rule of thumb is to use the most narrowly defined group expected to be used in the ensuing analyses as the starting point and work up the level of aggregation chain to arrive at the final sample.

Once selected, the sample extract usually needs to be enriched with additional, typically causal data to enhance its explanatory power.

Extract Data Enrichment

Additional data can be added to the selected sample either from other internal systems or from outside partners or data suppliers. However, one should not lose sight of the fact that since much of the data organizations capture are by-products of their ongoing operations, data tend to be scattered throughout the organization's various systems. Common examples include campaign data (mail list, offers and responses), field sales information (customer visits, inquiries, etc.), outgoing telemarketing (contact lists, call dispositions, etc.), satisfaction surveys, complaints. Although many organizations have been trying to integrate much of that data into a single data reservoir—the so-called

"360° customer view"—it tends to be a slow process and more often than not much of that data remains scattered across various internal systems.

At the same time, it is often beneficial to look outside of the organization for sources of potentially explanatory data. Although there are dozens of data suppliers specializing in data compilation and aggregation, in general, the bulk of the third-party data can be classified as either U.S. Census derived aggregates or special interest survey extrapolations.

The *U.S. Census derived data* represents geography-based aggregation of the detailed resident and business information collected by the U.S. Census Bureau. The resultant descriptors, commonly referred to as "geodemographics," are built around metrics describing age, gender, marital status, children, education, income and employment, race and ethnicity, home values and home ownership. The finest level of granularity at which these data are made available by the Census Bureau is the block level, which is the smallest geographic unit for which the Census Bureau provides (to outside, commercial entities) 100% populated data tables. In urban areas it typically corresponds to city blocks bounded by streets. In total, as of the last U.S. Census (conducted in year 2000), the United States is divided into over 8 million individual blocks. The bulk of the publicly available—and thus commonly appended to commercial, transactional databases—U.S. Census data is at the block level.[7]

Due to privacy considerations, however, some of the Census data is only made available at the census tract level. As defined by the U.S. Census Bureau, census tracts are "*small, relatively permanent subdivisions of a county*. . .[which] *normally follow visible features, but may follow governmental unit boundaries and other non-visible features in some instances; they always nest within counties* [and] *average about 4,000 inhabitants*." As of the 2000 U.S. Census, the United States was broken down into 65,443 census tracts.

The other, commercially available and frequently seen source of explanatory data often used to expand the informational value of transactional data extracts comes in the form of special interest, sample survey-based *behavioral and preference estimates*. One of the two main sources of these data are consumer panels, exemplified by AC Nielsen's scanner-based purchase tracking panel, Yankielovich consumer lifestyle tracking studies or Market Facts' online and catalogue purchase tracking panel. The main advantage of consumer panels, in general, is the breadth of topical coverage and longitudinal continuity, both of which support an assessment of longer-term trends. The most noteworthy drawbacks, on the other hand, are participant self-selection and demand effects, which are another way of saying that any panel-participant-reported facts or opinions may not accurately depict the actual behaviors or feelings of consumers at large. Overall, however, the good outweighs the bad, particularly when trying to understand observed consumer behaviors (i.e., database-contained purchases). The most frequently used panel-derived descriptors are consumer lifestyles and psychographics, online and catalogue purchase propensities and innovation adoption propensities, such as cell phone usage, broadband internet connection purchase for individuals or distributed computing for businesses.

The other main source of the commercially available, sample-based estimates are consumer and business credit reporting bureaus. On the business data side, organizations such as Dunn & Bradstreet make available a range of general industry characteristics and financial performance metrics, while on the consumer data side, the three main consumer credit bureaus (Experian, Equifax and TransUnion) make available consumer credit and financial-asset-related estimates. In addition to the mass data suppliers, a number of

smaller firms offer sample-based estimates of the otherwise hard-to-quantify consumer metrics, such as household net worth, or automobile registration.[8]

The quality—defined in terms of coverage and accuracy—varies across the types of data. Consumer and business credit information sourced from the three major credit bureaus defines, naturally, the upper end of the quality spectrum, though its usage for analysis and other purposes is also highly restricted and tightly regulated. Other data types, such as household net worth, lifestyle or purchase propensity estimates are usually considerably less accurate, since they represent sample-based extrapolations. Overall, these data types hold some explanatory power if used in conjunction with other information and applied to groups of consumers or businesses, but tend to be relatively inaccurate at the individual entity level, such as a household.

Usage: Research vs. Marketing

A major consideration in acquiring database information-enriching outside data is its intended usage. Most commercial data aggregators and resellers adhere to a dual pricing scheme: Data that is only used for research and analysis is priced lower, often considerably, than the same data that is to be used for promotional purposes. The former is usually defined as the analysis of data leading to the creation of knowledge about consumers, their behaviors and preferences. The latter is usually taken to mean the using of the data supplied by the data provider as the basis for targeting specific consumers.

An even more important consideration in selecting the appropriate outside data are the potential legal constraints. As the privacy concerns are becoming more pronounced, there is a growing trend of limiting the usage of consumer data by commercial entities. That said, most of the restrictions tend to focus on unauthorized data transfers/selling and using personally identifiable information as the basis for promotional initiatives. One of the common means of enforcing privacy restrictions is the anonymization of data appends by means of removing all personally identifying information, such as names and addresses. However, even if the appended data were not stripped of all personally identifying information it is important to make sure that the usage of the data does not exceed the limits stipulated by the user license or the applicable laws.

Organizing and Cleansing

As previously discussed and illustrated in Figure 6.7, an analytic data file should be customer-centric (in the data structure sense), with individual events, such as purchases, promotional responses or outside-attributed propensities, relating to customers as attributes. Not only is that type of an organizational schema an expression of the intended focus of most database analytical efforts, but it also makes appending of outside data easier (by supporting file merging at an individual customer/household level). It follows that by the time the outside data has been added to the original database extract, the so-enlarged file is already properly organized. In the event that is not the case, the analytic dataset needs to be rearranged in accordance with the general rationale shown in Figure 6.8.

Once properly organized, the dataset needs to be cleansed before any meaningful analysis can take place. According to a TDWI report, "Data Quality and the Bottom Line," poor data quality costs U.S. businesses approximately $600 billion annually.

Obviously, data-quality-related losses are inherently difficult to quantify, hence it is possible that an actual magnitude of these costs might be somewhat higher or lower. That said, even if the "true" losses are only 50% or so of the above estimate, it is still a significant enough problem to be given a serious consideration. In a more abstract sense, it seems clear that adequate data quality controls should be a necessary component of the database analytics-driven knowledge creation processes, if the results are to be valid and reliable. The two key data quality due diligence steps are data cleansing and normalization. Technically, data normalization is a subset of data cleansing, but in view of both its importance and a certain degree of technical complexity it will be discussed as a stand-alone concept.

Data cleansing can be broadly defined as the process of repairing and normalizing of the contents of an extracted dataset. *Data normalization*, on the other hand, is the process of identification and removal of outlying and potentially skewing and influential data values and correcting for undesirable distributional properties by means of missing and derivative value substitution.

Data Normalization: Outlier Identification

An outlier is a value that falls outside of what is otherwise considered an acceptable range. Some outliers are illustrative of data errors, but others might represent accurate, though abnormally large or small values that are extremely rare. Depending on the size of the dataset (i.e., the number of observations), outliers can be visually identified by means of simple two-dimensional scatterplots, or can be singled out by means of distribution scoring, which is the process of "flagging" individual records whose values on the variable of interest fall outside of the statistically defined norm, such as ±3 standard deviations away from the mean.

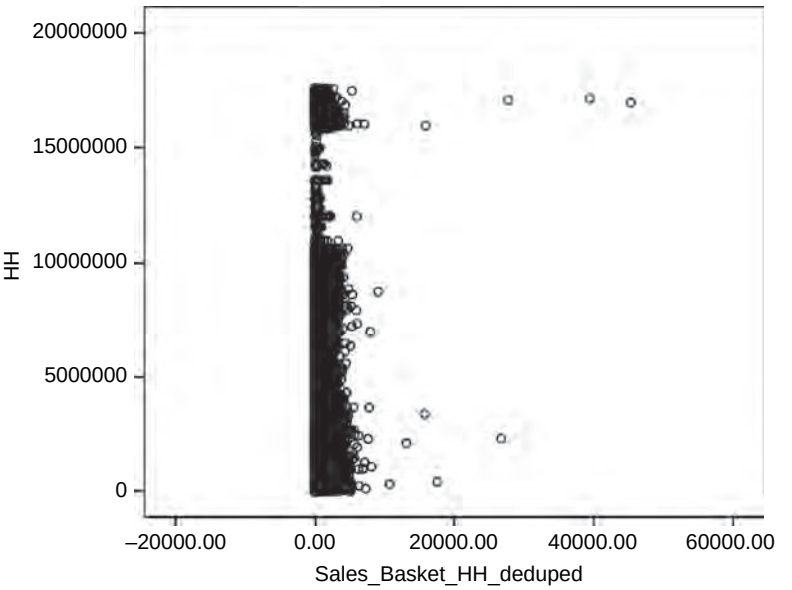

Figure 7.1 Scatterplot Example: Household Weekly Grocery Spend

Visual representations of data, such as the aforementioned scatterplots, offer the simplest method of identifying outlying observations, as illustrated below by a plot of household weekly spending levels at a grocery store chain.

However, the visual detection of outliers becomes practically impossible once the number of dimensions (i.e., defining variables) is greater than two and the data can no longer be easily shown as a simple two-dimensional plot. In addition, even when a two-dimensional representation is sufficient, there is an obvious difficulty associated with the reliance on the visual outlier detection, which is, where to draw a line between an outlier which should be eliminated and a large but retainable value. As illustrated by the above chart representing real-life data, the most extreme—i.e., the furthest to the right in the above illustration—values are easy to classify as outliers, yet the closer one gets to the bulk of the distribution, the more difficult and arbitrary the decision becomes. Outlier treatment needs to be approached with caution, as eliminating too many large values carries with it undesirable statistical and practical consequences. On the one hand, retaining too many excessively large values may cause an inflation of "average" and distributional measures, which in turn may lead to practically problematic consequences, such as unrealistically inflated high-value customer definitions. Still, excessive "large value clipping" will often reduce the variability of the data, which tends to diminish the explanatory power of the subsequent data analysis. In practical terms, it may exclude the very best customers, whose spending levels are considerably above the rest.

The best way to circumvent the subjectivity of visual outlier identification is through the aforementioned dataset flagging, which takes advantage of basic distributional properties of data to objectively identify not only the outlying data points, but those that may exert excessive amounts of *influence* on the analysis of data. In the context of database analytical applications, the influence of a particular database record is a function of that record's deviation from the norm and its leverage, expressed as:

Influence = Deviation from Mean * Leverage

In terms of individual values, influence of a particular value is determined by estimating the standard deviation of actual values (deviation from the mean) from the mean value and the leverage of (potentially) outlying values. Leverage is an expression of "outlying-ness" of a particular value, which is its distance away from the average value—the further away from the mean a particular value, the more leverage it has. Hence influence is simply a measure of the distance, expressed in terms of standard units, away from the center of the distribution.

The challenge associated with the above method—particularly in the sense of measuring the degree of non-conformity (i.e., the outlyingness of individual data points)—is the potential for *masking* of some outliers. In other words, since all data points, including any outliers, are used in computing a mean, it is possible that some of the "less extreme" outliers will be masked by artificially inflated mean values. It is important to point out that the masking problem occurs with both the physical distance and magnitude of difference-based measures.[9]

A relatively easy fix is available. Prior to computing the mean of a particular sample, rank-order all sample records and exclude the lower and upper 5%–10% of the records, prior to calculating the mean. Doing so will prevent any potential outliers from effecting the mean and thus eliminate the masking problem outlined above.

The appropriately computed mean can then be used as the basis for quantifying the extent of deviation from an expected level for each of the records in the analytical dataset. Of course, there are typically a number of candidate variables to be used as the basis for classifying a particular record as an outlier. To that end, it is recommended that only the behavioral metrics, previously described as the "behavioral core" variables should be used as the basis for the outlier determination. This is because these measures present the greatest danger of individual customer record misclassification or group level mischaracterization, which is why it is so important to define outlyingness in accordance with their particular characteristics.

The second broad category of data—the causal data—can be used as the basis for outlier detection, but it is generally not recommended. The primary reason for that is that although these types of metrics can be used for customer or prospect grouping (e.g., demographic or lifestyle segmentation), more often than not their role is to provide explanation and/or description of database records, which suggests an overall weaker skewing effect. In addition, in a "typical" database, a considerable proportion of the causal data represents third-party approximations or group-level estimates (e.g., block level geodemographics), both of which tend to be normalized in the process of their computation. In other words, causal metrics are considerably less likely to have outlying values.

The Process

Regardless of the approach used, the definition of what constitutes an outlier will always carry with it a certain level of ambiguity, or at least subjectivity. Hence outlier identification is as much about the process as it is about thresholds. Putting in place a single and consistent (across time and applications) method for detecting and remedying outlying values will at the very least diminish the possibility of introducing a selection-related bias.

In transactional databases the metric of interest is usually represented by an average or cumulative (customer/prospect) spending level, or an otherwise stated measure of sales/revenue. As pointed out earlier, value outlyingness can be operationalized in terms of the number of standard deviation units that the value of interest is greater or smaller than the appropriately computed mean value. (The appropriately computed mean computation excludes the two tail ends of the value distribution, typically 5% or 10% of the most extreme values. A similar adjustment should also be made when computing the standard deviation metric.).

What remains is the setting of outlier threshold—in other words, at what point an otherwise large value becomes an outlier? In thinking about this issue, consider the goals of the planned analysis and the general inner-workings of statistical methods to be used. Although both vary, a common objective is the identification of high-value current and prospective buyers for targeting purposes, which typically makes use of a variety of regression methodologies. Regression parameters are evaluated in terms of their level of significance (more on that later), which is ultimately tied to distributional properties, including the notions of standard error and standard deviation. The commonly used 95% significance level expresses the validity of an estimated parameter in terms of the likelihood of it falling within ±2 standard deviations away from the mean. Why not calibrate the acceptable value range to the anticipated level of precision? Using such an objective benchmark, only records falling outside the standard-deviation-expressed

range of allowable departures from the mean should be flagged as abnormal. This rationale can be translated into the 4-step process outlined below:

- Step 1: Compute the mean and the standard deviation of the variable of interest.
- Step 2: Select the desired allowable limits; i.e., 3 standard deviations away from the mean = 95% of the values, 4 standard deviations away from the mean = 99% of the values, etc.
- Step 3: Compute the maximum allowable upper values: *mean + upper allowable limit* and the maximum allowable lower values: *mean – lower allowable limit*.
- Step 4: Flag as abnormal the records falling outside the allowable range, both above and below.

Demonstrably outlying records should be eliminated from the dataset. Before they are deleted, however, it is worthwhile to discern if they are more-or-less randomly distributed across the customer groupings, or concentrated in a particular group or groups. If it is the latter, the basic descriptive characteristics of the effected groups should be compared in terms of the before and after the outlier deletion. If significant differences are uncovered, such as mean differences in excess of one standard deviation, the sample-based analysis may not be representative of that group's entire population in the source database. The most obvious remedy is to re-draw that particular part of the sample or to exclude it from the extract if re-sampling is not feasible within the available timeframe. In addition, the reason behind the high outlier concentration should be investigated.

Data Normalization: Distribution Correction

First, let's take a minute to take a closer look at the general notion of a "statistical distribution" and its best-known form, the so-called "standard normal distribution." Statistics is in essence an application of rigorous mathematical techniques to a sample with the goal of making inferences that can be applied to the entire population from which a particular sample was drawn. Doing so necessitates the making of a number of substantive assumptions about the likelihood that the observed, sample-based quantities and relationships are representative of those to be found in the parent population. In order for the sample-to-population extrapolations to be reasonably accurate, values of individual variables must be distributed, i.e., scattered around a center such as the mean, in a predictable fashion. When the observed values of a particular variable follow a generalizable distribution they have a predictable likelihood of occurrence, which in turn forms the basis for sample-to-population generalizations.

It should be noted that density distributions represent empirically based generalizations, which means that they can take a variety of mathematically described shapes. *Normal distribution* is the best known and the most frequently used form of a continuous variable's probability distribution.[10] It can be transformed into a *standard* normal distribution with the mean equal 0 and a standard deviation (which is a scale-neutral unit of measurement) of 1 by re-expressing the original metric in a standard format of the so-called "z-scores" computed as follows:

$$z = \frac{x - \mu}{\sigma}$$

where,

x = an observed value
μ = mean
σ = standard deviation

The "popularity" of the standard normal distribution can to a large degree be attributed to the Central Limit Theorem, which states that a sample mean will follow an approximately normal distribution if the sample is large enough, even if its source population is itself not normally distributed. Standard normal distribution is also very simple, as it is described by only two parameters: a mean equal to 0 and a standard deviation equal to 1. This function is important to statistics because it lies at the root of many other continuous distributions. By extension, it is also of central importance to database analytics because most of the statistical techniques essential to knowledge creation require the data to be normally distributed. Normally distributed data is characterized by the familiar symmetrical bell-shaped curve, which implies a clustering of the majority of values around the center of the distribution; it also implies that the more outlying values are predictable as well as equal (i.e., balanced right and left departures from the mean). Figure 7.2 below depicts the generalized form of the standard normal distribution.

Somewhat complicating the notion of data distribution is the distinction between univariate (i.e., one variable) and multivariate (multiple variables) distributional characteristics. The two-dimensional conceptualization depicted in Figure 7.2 captures a univariate distributional view—i.e., a scattering of observed values is depicted in the context of single variable, such as the level of spending or period sales. Since most database extracts contain multiple continuous variables, the univariate distributional assessment would entail an equal (to the number of the said variables) number of single-variable distributions. However, as discussed in the subsequent chapters, the knowledge creation process described in this book calls for multivariate statistical techniques, which in turn entails the assessment of the underlying data's *multivariate normality*. It is important to note that a multivariate normal distribution is not a mere composite of univariate normal distributions. In other words, even if every variable in the dataset is normally distributed, it is still possible that the combined distribution is not multivariate normal. For that reason, the extract dataset's distributional properties should be

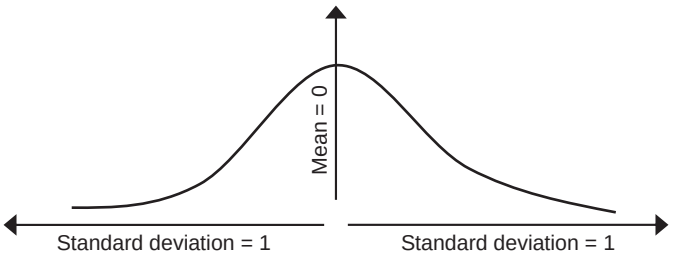

Figure 7.2 Standard Normal Distribution

evaluated in both the univariate and multivariate contexts. As detailed below, the univariate assessment can be easily carried out graphically (see Figure 7.4 below), but the multivariate examination requires a numeric test, such as the one suggested by Mardia,[11] now readily available as a part of standard statistical packages, such as SAS or SPSS.

In practice, many sample database extracts are often not normally distributed in regard to the key customer behavioral characteristics, such as spending or sales levels, while at the same time are approximately normally distributed with regard to many causal characteristics, such as demographics or firmographics. In other words, extract files are rarely multivariate normally distributed. One reason for that is the often-cited "80–20 rule" (Pareto principle), according to which, the top 20% of the firm's customers account for about 80% of its revenue. Naturally then, it will result in an asymmetrical clustering of data points, which is statistically undesirable but nonetheless quite common in transactional databases. Equally problematic is the excessive "flatness" of the distribution, which is typically a result of a monotonic value distribution sometimes associated with geodemographic descriptors.

Generalizing beyond the above examples, the analysis of the basic distributional properties of the normal distribution suggests two basic normalcy evaluation criteria: S*kewness*, which measures the symmetry (or more precisely, the lack of it), and *kurtosis*, which captures the peakedness of data relative to a normal distribution. A normal distribution is expected to exhibit an acceptable amount of symmetry (skewness) and "tightness" around the mean of the distribution (peakedness). As detailed in the next section, there are several graphical and numeric approaches available for assessing both distributional characteristics.

Normality of the distribution is important because many of the "staple" statistical techniques and tests used in database analyses require the underlying data to be at least "approximately normally" distributed. Absent that, the knowledge creation process can become considerably more challenging, as alternative data crunching methodologies may need to be identified. It is important to keep in mind, however, that the notion of distributional normalcy is associated with continuously distributed (also referred to as "metric") data. These are variables measured with either interval or ratio scales, as exemplified by Likert-type attitude measures or product sales, respectively. Discrete variables' (also referred to as "non-metric") properties are expressed with the help of other concepts, discussed later.

Skewness Identification and Correction

A distribution is *skewed* if one of its tails is longer than the other. A *positively skewed* distribution (also referred to as being "skewed to the right") is one that has a long tail extending to the left of the center, which occurs when the mean of the sample is greater than its mode.[12] It follows that a *negatively skewed* distribution (also referred to as "skewed to the left") has a long tail extending to the right of the center, as a result of the mode being greater than the mean. The general shapes of positively and negatively skewed distributions are illustrated in Figure 7.3.

In a statistical sense, a skewed variable is one whose mean is not in the center of the distribution. Both positive and negative skews can be operationalized by comparing the actual values against a null hypothesis of zero and testing the difference for statistical significance. A statistically significant actual-minus-null deviation is then taken as an indication of an action-demanding departure from normality. In practice, however, many

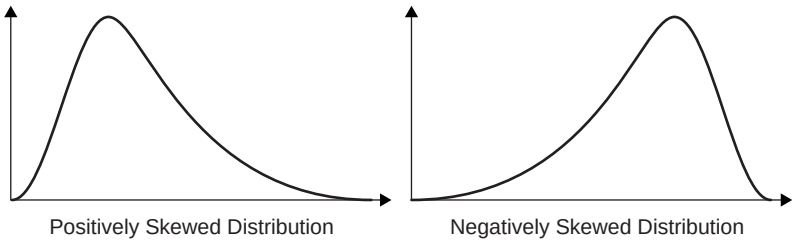

Figure 7.3 Negatively and Positively Skewed Distributions

commonly used statistical techniques, such as regression, have been shown to be relatively robust with regard to modest deviations from normality of the input data, while large-sample-based significance tests have been shown to be highly prone to yielding false positive results.[13] Thus in assessing skewness of the distribution of a particular continuous variable, the question should be: Not whether or not it is skewed in an absolute sense, but whether it is skewed in the sense of diminishing the validity and/or reliability of any subsequent analysis.

Put that way, there are no absolute skewness benchmarks or thresholds. Also, as noted above, most of the mainstream statistical techniques, such as the various regression formulations, are robust with regard to some departures from the expected normal data distribution. This eases the assessment of the quality of continuously distributed data, simply because more tolerance in distributional deviation places the focus on the identification of gross deviations from expected norms. At the same time, not all statistical techniques—as discussed in the ensuing chapters—require normally distributed data. However, the assessment of the extract's distributional properties should be undertaken without regard to ensuing statistical analysis, as it is a key step in ascertaining representativeness of the sample.

The notion of skewness takes on a somewhat different meaning when it comes to non-continuously distributed categorical variables. Often referred to as non-metric or qualitative data (measured with either a nominal or an ordinal scale), these variables are typically evaluated in terms of the relative proportions. In other words, for a dichotomously coded variable, such as gender (male vs. female) or a response indicator (yes vs. no), the ratio of the two response categories can be compared, with the same approach being applied to a larger number of response categories.

The simplest and perhaps the quickest method of assessing univariate distributional properties of the extracted dataset is a histogram, which can be used with both continuous as well as discrete metrics. A simple visual representation of the individual variable's response categories' frequency distribution, this method efforts a quick visual means of detecting skewed distributions, as exemplified below.

A histogram, however, can be inconclusive in many non-extreme situations, i.e., where the data is somewhat, but not strongly skewed, simply because it does not have built-in objective determination thresholds. In other words, relying only on visual depictions, it is often difficult to tell the difference between "somewhat skewed" and "approximately normal" distributions. Fortunately, there are several computational tests that can be used to measure the degree of skewness which circumvent the limitations of histograms and other visual methods, with the three best known ones shown below.

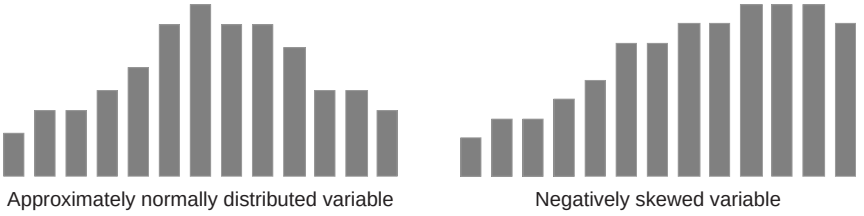

Figure 7.4 Sample Histogram

1. Pearson's coefficient of skewness =

$$\frac{(mean) - (md)}{stdev} = \frac{3(mean) - (medium)}{stdev}$$

where,

md = mode
stdev = standard deviation

2. Quartile measure of skewness =

$$\frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$$

where,

Q_1, Q_2 & Q_3 = quartile 1, 2 and 3, respectively

3. Coefficient of skewness =

$$\frac{\Sigma(x - \mu)^3}{N\sigma^3}$$

where,

μ = mean
σ = standard deviation
N = sample size

Although the above tests yield an objective quantification of the degree of skewness, they lack a normative threshold that analysts can use to identify unacceptably or excessively skewed distributions. Considering that skewness can be expressed on a continuum ranging from '0' for perfectly symmetrical, normally distributed sample to an infinitely large (at least in theory) positive or negative values, and also keeping in mind that some skewness can be tolerated even by methods requiring normally distributed data, an objective threshold is needed to guide an analyst in delineating the point at which skewness becomes a problem. The following decision rule is the recommended solution[14]:

if $|skewness\ statistic| > \sqrt{\frac{6}{N}}$ then dataset is significantly skewed

where,

N = sample size

In other words, if the absolute value of skewness statistic exceeds two standard errors of skewness, approximated by

$$\sqrt{\frac{6}{N}},$$

the dataset should be considered significantly skewed. If that is the case, two separate courses of action are available:

1. The previously detailed outlier detection and elimination.
2. Mathematical dataset transformations.

In general, *data transformation* can be defined as the re-expressing of the data in different units with the goal of bringing about desired data qualities, which usually means correcting for departures from normality. There are multiple types of mathematical transformations available and the selection of a method is usually driven by methodological considerations (for instance, one of the key requirements of linear regression is that data follow normal distribution, while other statistical techniques impose no distributional requirements), empirical test results (such as the above discussed skewness tests) and proven empirical "rules of thumb" (for instance, it has been shown that frequency-count-based data can be made more normal by taking their square roots). A transformed value then is as a function of an original value and a type of a transformation used. In the data analytical sense, transforming data amounts to replacing an original value of a particular variable with a result of a specific transformation correcting for specific distributional deficiency. A word of caution: Although useful in correcting for undesirable characteristics of data and thus ascertaining the robustness of findings, the use of transformations can nonetheless lead to biased or outright improbable results. Furthermore, using transformed data as input into analyses results in an additional step of translating estimated coefficients into directly interpretable values (for instance, if log transformation, discussed below, was used in estimating a linear regression model, the resultant regression coefficients would need to be transformed back into a standard format by means of computing antilogs of their values before the results could be used in the manner outlined in the *Behavioral Predictions* chapter).

In general, transformations can be either linear or non-linear. A *linear transformation* changes the scale, but not the shape of the distribution, while a non-linear transformation changes the shape of the distribution. In terms of purpose, linear transformations are typically used to *standardize* (or z-standardize) variables, which entails converting values expressed in original units of measurement to standard deviations from the mean. It is important to note that contrary to what is often believed, standardization does not normalize a distribution, hence a skewed distribution will remain skewed following variable standardization. This means that a linear transformation of data should be considered if the goal is to simplify cross-variable comparisons. On the other hand, a *non-linear transformation* should be considered if the shape of an underlying distribution needs to be changed, or more specifically, if non-normally distributed data is to be forced to assume the previously mentioned standard normal distribution.

As noted earlier, the primary benefit of linear transformation, such as the aforementioned standardization is enabling direct, side-by-side comparisons of otherwise not directly comparable measures. By re-expressing differently scaled (i.e., measured with magnitudinally dissimilar unit, such as miles vs. years) variables as standard-deviation-expressed z-scores, the relative influence of metrics can be assessed, ultimately enabling importance-based rank ordering of measures of interest. Other than variable standardization, linear transformation methods include adding a translation constant (adding a constant value to each raw data point with the goal of shifting of the origin of the X-axis) or a multiplicative constant (multiplying each raw value by a constant with the goal of scaling, which is expanding or contracting of the underlying distribution). Regardless of type, linear transformations are focused on re-scaling of individual variables without impacting the aggregate distribution.

Non-linear transformations, on the other hand, are focused on changing the shape of the underlying distribution. This is a somewhat more esoteric goal, driven less by practical demands of result interpretation and more by methodological requirements of statistical techniques. Depending on the type (i.e., skeweness vs. kurtosis) or the strength of non-normality, a number of different transformation options are available, all of which fall under either of the two broadly defined categories: *logarithmic transformations* (or log, for short) and *power transformations*. In the majority of cases, logarithmic (with natural, ln, and base-10, log10, being the most common ones) transformation can be an effective normalization-infusing treatment for positively skewed data. That said, if the degree of skeweness is relatively mild, a logarithmic transformation may over-correct, in effect replacing positive with negative skewness (at the same time, severely positively skewed data may not be correctable with log transformation). Power transformations are usually a viable option in the instances where a logarithm fails to deliver the adequate degree of normalization. The following are the most commonly used power transformations, along with their targeted corrections:

- x^2 (raising the original value to the second power)—to reduce negative skewness
- x^3 (raising the original value to the third power)—reduces extreme negative skeweness
- $\sqrt{2}$ (the square root of the original value)—reduces mild positive skeweness
- $-1/x$ (negative reciprocal of the original value)—reduces more extreme positive skeweness

However, unlike the outlier elimination which should be carried out as a part of any a priori data cleansing, data transformations are usually executed in conjunction with specific statistical techniques, because as noted earlier, not all methodologies make specific distributional assumptions. Hence the topic of data transformations will be revisited in subsequent chapters.

Kurtosis Identification and Correction

The second of the potential distributional data challenges is peakedness, or kurtosis of data, relative to normal standard distribution. Positive kurtosis indicates relatively peaked distribution, while negative kurtosis indicates a relatively flat one, as depicted in Figure 7.5 below.

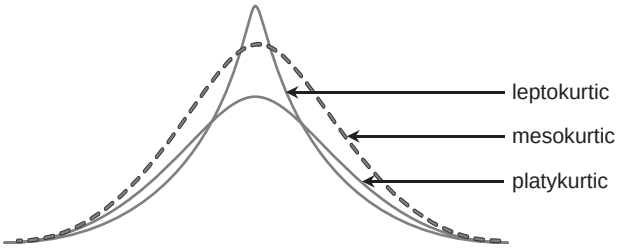

Figure 7.5 Mesokurtic, Leptokurtic and Platykurtic Distributions

Similar to skewness, normal distributions have approximately 0 kurtosis, and the departure from normality represents a movement along the 0-to-infinity continuum. A normally "high," in the physical sense, distribution (i.e., a shape that is neither excessive peaked nor flat) is usually called mesokurtic, in contrast to an abnormally high distribution which is called leptokurtic and an abnormally flat one which is called platykurtic.

The easiest method of detecting kurtosis is by visually examining a histogram of the distribution, but as it is the case with skewness, it is not always possible to visually differentiate between what might be an acceptably small deviation from the norm (i.e., not undesirably impacting subsequent analysis) and a significant departure. This difficulty can be to a large degree circumvented by employing a numerical kurtosis test, shown below:

$$\text{Kurtosis coefficient} = \frac{\Sigma(x-\mu)^4}{N\sigma^4}$$

where,

μ = mean
σ = standard deviation
N = sample size

Again, the question of "where to draw the line" arises: Since few distributions are perfectly mesokurtic, i.e., have 0 kurtosis, how large a departure from the idealized norm can be accepted before the validity and reliability of subsequent data analysis are negatively impacted? The following decision rule can be used[15]:

$$\text{if } |\text{kurtosis } \textit{statistic}| > \sqrt{\frac{24}{N}} \text{ then dataset excessively non-mesokurtic}$$

where,

N = sample size

In other words, if the absolute value of the kurtosis statistic exceeds two standard errors, approximated by

$$\sqrt{\frac{24}{N}},$$

the dataset should be considered significantly leptokurtic or platykurtic, as applicable.

Data Repairing: Missing Value Substitution

One of the key criteria for evaluating the initial analytic quality of database extracts is the degree of completeness, which is usually expressed at an individual variable level. For instance, if 1,000 individual records were extracted, with each containing 100 individual variables, what is the proportion of missing to non-missing values for each of the 100 variables across all 1,000 records?

Frankly, it is rare for an extract data file to be 100% complete. First, organizations vary in terms of their data capture and maintenance proficiency. But even those that excel in that area still need to contend with the inevitable "missing value" challenges. In other words, due to factors including human error, occasional technical glitches or imperfect data capture methods, virtually all data types will exhibit some degree of incompleteness. In general, the behavioral core data typically yields a smaller proportion of missing values than the causal enhancement data. At a more micro level, any electronically captured point-of-sales data, such as the UPC-scanner-based or online transactions will usually exhibit the highest level of completeness, with the typical missing value proportion of less (at time considerably) than 10%. On the other hand, the (behavior) augmenting causal data, such as demographics, firmographics, lifestyle or ascribed purchase propensities tend to yield considerably lower completeness rate, in some cases as low as 5%–10%, particularly for the third-party geodemographic overlays.

Most data analyses cannot proceed unless these "data holes" are filled (by default, some of the commonly used statistical software packages such as SAS or SPSS will eliminate all records containing missing values on variables used in the analysis, a process often referred to as "pairwise deletion"). These unintended deletions of missing data-containing records are obviously troublesome from the standpoint of maintaining a sufficiently large analysis sample size. Even if missing values are randomly scattered throughout the extract dataset, a (favorable) condition termed *missing completely at random* (MCAR), the analysis sample may become prohibitively small as a result of missing case deletion.

Another, even potentially more handicapping consequence of such unchecked elimination of missing value records is a systematic bias creation potential, which is a result of underlying—though usually not self-evident—commonalities shared by missing value cases. This might be particularly evident in the context of the previously outlined stratified sample, where the overall universe of database records is comprised of several, clearly discernible sub-categories. In this case, it is possible that any missing value driven record deletion would impact some segments noticeably more than some others.

And finally, even if the effective (i.e., post-deletion) sample remains robust in terms of its size and unbiased in terms of its composition, the amount of variability in the data will certainly have been reduced, which may potentially adversely affect the robustness of findings. In other words, since the amount of data variability is directly related to the explanatory and/or predictive power of data analyses (i.e., low variability attests to very few or very weak cross-record differences), reducing it runs the danger of diminishing the informational content of the data.

In view of these potentially significant missing value deletion consequences, the safest approach to dealing with missing data often turns out to be a reasoned *a priori replacement* strategy. However, for reasons detailed below it is not always possible to take this course of action and even when it is possible, it entails its own due diligence process.

First and foremost, each variable needs to be assessed in terms of its usability—i.e., does its coverage warrant inclusion in the ensuing analysis, or should it be outright

eliminated. For instance, a metric which is 90% populated will almost always warrant inclusion in future analysis; on the other hand, a metric which is 90% missing will almost warrant exclusion from any analysis. In practice, however, most variables will fall into that grey area of indecision that tends to span the middle ground between the two extremes. What then?

To some degree, the answer depends on the type of variable. The behavioral core metrics should be held to a higher standard of completeness simply because they are manifestations of factual actions and in a statistical sense, tend to serve as predictive targets, or dependent variables that do not have immediate substitutes or proxies. A database record lacking transactional details lacks the most fundamental classificatory dimension enabling it to be correctly categorized, which renders its informational value null. Assigning any value, be it the mean, median or a regressed value would amount to creating data. In other words, database records with missing behavioral core metric values should be eliminated from further analyses.

Causal data, on the other hand, can be held to a less stringent standard. As enhancements to customer behaviors, these variables are descriptive (rather than classificatory) in nature and usually have multiple proxies or substitutes—i.e., they are a part of the multivariate mix, rather than being a univariate target. In a statistical sense they tend to be deployed as predictor (also called independent) variables, usually as a component of a large multivariable mix, as shown in later chapters. In other words, a record with missing values on some of the descriptive causal variables still makes a positive informational contribution with other, populated causal variables.

The question that arises, however, is what should be the upper limit of missing values (i.e., the proportion of missing) that should be deemed acceptable? There is no agreement among analysts (or among theoreticians, for that matter) as to what such a threshold should be and as a result, treatments vary widely across situations. That said, it is reasonable to assume that a missing value metric, which could be called "proportion of missing values" is a continuous, randomly distributed stochastic variable that could be examined within the notion of standard normal distribution. It also seems reasonable to conclude that basic distributional properties of the normal distribution can be used as bases for identifying an objective missing value evaluation threshold. In particular, the proportion of all observations accounted for by the set number of standard deviations away from the mean seems particularly appropriate, as it expresses the probability of the actual value falling within a certain range. Of course, the choice of how many standard deviations away from the mean may constitute an outlier is usually somewhat arbitrary.

As shown in Figure 7.6, roughly two-thirds of all observations can be accounted for within ±1 standard deviation away from the mean, which increases to about 95% of all observations within ±2 standard deviations and more than 99% within ±3 deviations away from the mean. Although 95% or even 99% would be the ideal standards, in practice, setting the threshold at such a high level would lead to the elimination of the vast majority of individual metrics. At the same time, recent analyses[16] indicate that any variable which is populated no less than 68%—or in reverse—which has less than 32% of its values missing can still be "repaired" without significantly affecting its basic distributional properties. This means that causal variables which have more than roughly one-third of their values missing should be excluded from further analysis.

The basic distributional properties of the standard normal distribution also offer value replacement hints. First and foremost, the practice of replacing all missing values with a single value (usually one of the measures of central tendency, typically either the mean

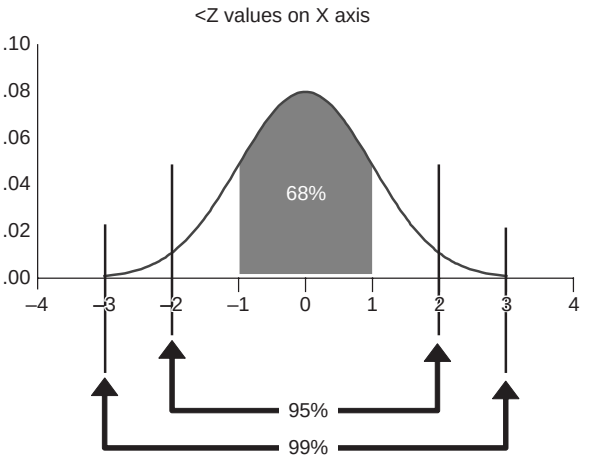

Figure 7.6 Properties of Standard Normal Distribution

or the median) should be avoided as it tends to diminish the variability in the data, which can in turn lead to bias, unreliable effect estimates, ultimately undermining the robustness of the resultant information. A somewhat better value replacement strategy used by database analysts involves mimicking the generalized normal distribution proportions, which means randomly assigning either mean or median values to some, while mean/median ±1 or 2 standard deviations to others, in a proportion depicted in Figure 7.6 above. In a number of instances, however, particularly when missing values are not randomly distributed deletion of incomplete records might be a safer road to take.

Missing Data Imputation

In instances where the deletion of missing values containing records is not a viable option, the most effective method of dealing with missing values is *value imputation*, which is the process of estimating the most likely missing value by using a non-missing value in the sample. The imputation process can either result in physically replacing the missing value—i.e., substituting an actual value in place of the missing one—or just imputing the distributional characteristics, such as means and standard deviations, or relationships from the available data without actually physically replacing the missing values. Although this discussion is concerned primarily with the former, i.e., the physical value replacement, both approaches will be discussed to paint a complete picture of the available replacement options.

PHYSICAL REPLACEMENT OPTIONS: COLD DECK IMPUTATION

Perhaps the simplest approach to missing value imputation is to replace them with an externally derived constant. However, because a single value is imputed into numerous cases, this approach will lead to an artificial reduction in the variability of data, which is likely to bias coefficient estimates, ultimately diminishing the explanatory or predictive validity of findings.

Mean Imputation. Missing values are replaced with the commonly used measures of central tendency, such as the mean, estimated with non-missing values. Its main advantage is the conceptual and operational ease (SAS and SPSS have built-in functions to carry out this operation). The disadvantages, however, are numerous. First of all, as stated earlier, the true variance in the data will be understated, which will reduce the reliability of subsequent analyses. Secondly, the actual distribution of values is likely to be distorted. Thirdly, cross-variable relationships will be depressed because of a constant value being imputed into numerous cases. Although commonly used, this method produces results inferior to regression or model-based methods discussed below.

Regression-Based Imputation. Regression analysis, a multivariate statistical technique described in Chapter 6, finds the best-fitting substitute for the particular missing value based on the relationship of the missing value variable with other variables in the dataset. Obviously, a multivariate-regression-based approach makes a better use of the available data, hence intuitively it should be more efficient (efficiency is a function of the algorithm's ability to yield unbiased coefficient estimates while also being easy to implement), which is indeed the conclusion of the investigative research.[17] This particular method, however, also has several distinct disadvantages. First of all, it is considerably more complex as it requires the calibration of a multivariate statistical model. Secondly, it reinforces the relationships already in the data, potentially diminishing the validity of future findings. Thirdly, the variance of the distribution will likely be diminished, unless stochastic values are added to the estimated values, which will further increase the level of complexity. Lastly, it makes an assumption that the missing value variable is highly correlated with other variables. Ultimately, empirical comparisons of the efficacy of this method found it to be less efficient than the two probabilistic approaches described next.

Probabilistic (Also Called Model-Based) Approaches. Sometimes the simplest solutions are the most effective ones. That truism does not seem to hold in the context of missing data imputation, as the most methodologically complex methods have also been found to yield the most robust replacement values in general, while at the same time also being most efficient. There are two distinct model-based missing value replacement methods. The first is the *maximum likelihood estimation* which uses all available data to generate the correct likelihood for the unknown parameters. Although there are numerous maximum likelihood computational methods, in general, they are all based on the assumption that the marginal distribution of the available data provides the closest approximation of the unknown parameters. The good news is that the main statistical analysis systems already implemented the maximum likelihood methods for missing data (e.g., SPSS Missing Value Analysis). The bad news is these methods are quite computationally intensive, which translates into higher processing power requirements.

The second of the probabilistic approaches is *Bayesian imputation*, which represents a probability-based way to estimate the conditional and marginal distribution for missing data. Computationally, it is based on a joint posterior distribution of parameters and missing data, conditioned on modeling assumptions and the available data.

NON-REPLACEMENT OPTIONS: ALL-AVAILABLE INFORMATION APPROACH

By expressly taking into account all non-missing data, this approach estimates the cross-variable relationships, i.e., correlations discussed in the next chapter, and maximizes the pairwise information available in the sample. Each correlation is based on a unique set of

observations and each correlation is computed with a potentially different number of observations. The resultant correlations are representative of the entire sample and are used (in subsequent analyses) in lieu of the raw sample. Naturally, unless the missing values are truly randomly distributed, the correlations will be biased. However, even if the missing values are random, any of the correlations between X and Y can be inconsistent with other so-computed correlations due to dataset-wide interrelationships among all variables.[18] In other words, the range of values for any X–Y correlation is constrained by the correlation of X and/or Y to a third variable, Z, as shown below (based on Pearson's product-moment correlation coefficient *r*—see Chapter 5 for more details):

$$Range\ of\ r_{xy} = r_{xz}\, r_{yz} \pm \sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}$$

Missing value imputation is an important consideration, likely to have a considerable impact on the robustness of analytic results; hence it demands careful deliberation, particularly when choosing an approach. The first step in the data engineering process should always entail deciding between the outright elimination of the incomplete metrics and imputing the missing values. In many situations, particularly those involving large transactional databases that can easily absorb sample size reductions, eliminating poorly populated data might be the most appropriate corrective step to take, provided that the missing value case deletion would not be biased (i.e., systematically eliminate certain types of records, while keeping others). Empirical research almost universally found pairwise deletion (another term for missing value case elimination) to be the most efficient approach to missing values, and certainly the safest one from the standpoint of result validity.

If, however, throwing out missing value cases is not appropriate or feasible, the probabilistic methods—namely maximum likelihood estimation or Bayesian imputation—should be employed as, again, empirical analyses found those methods to be the most efficient. The conceptually and operationally easier to tackle—and more frequently used—mean or median substitution should be avoided as much as possible. By artificially deflating the amount of variability contained in the dataset and distorting the distribution of values, these methods can introduce a considerable bias into the analysis, ultimately diminishing the validity of the results.

Lastly, to further enhance the robustness of missing value imputations, an iterative approach should be used, paralleling the process developed by Rubin[19]:

1. Impute missing values using an appropriate model.
2. Repeat the above process, n-number of times (usually, 3–5) to produce n-complete datasets.
3. Perform the desired analysis on each dataset using complete-data methods.
4. Average the resultant parameter estimates across the individual datasets to arrive at a single point estimate.
5. Compute the standard errors of the estimates as follows:

$$SE_m = \frac{s}{\sqrt{n}}$$

where,

s is the standard deviation
n is the number of observations

Although obviously more laborious, the multiple imputation process has distinct advantages, especially:

- Repeated re-estimation is the only valid method of generating standard error estimates, which in turn makes it possible to get approximately unbiased estimates of parameters.
- Multiple imputation is relatively straightforward to implement, thus yields an attractive cost–benefit trade-off.

Data Repairing: Derivative Value Substitution

A *derivative value* is a result of data transformation aiming to correct for undesirable distributional characteristics of the original, or raw, value. Examples include a log of weekly sales or a square root of cumulative purchases. Derivative values are usually computed with a specific application in mind, such as a regression model, and are used in place of the original values. Results based on these values should not be interpreted prior to coefficients being "translated" into their original forms. For instance, before they can be interpreted as elasticities, regression coefficients based on a logarithmic transformation must be re-expressed in the original, non-logarithmic form, which requires computing anti-logs for all coefficients stemming from previously transformed variables.

Transforming data does not guarantee that the desired distributional properties will be indeed attained, hence it is critical to assess the results of it in the context of the sought-after end objectives. It is always possible that some of the metrics will fail to take on the methodology-mandated characteristics; however, empirical evidence suggests that certain types of continuous data transformations can be highly effective at reaching the stated variable distributional goals. The vast majority of data transformations are geared toward bringing about the normality of the distribution, in effect correcting for varying degrees of the previously discussed skewness, kurtosis or both. Figure 7.7 shows some of the more commonly seen data normality deviations, along with the recommended fixes.

It is important to keep in mind that the above-recommended transformations, as well as all data transformations in general are univariate distributional fixes—i.e., their goal is to correct for a single variable's lack of normality. As previously pointed out, even if all individual metrics are normally distributed, that does not guarantee multivariate normal distribution, which takes on special importance in the context of segmentation (Chapter 9) as well as targeting and behavioral-predictions (Chapter 10) focused analytical initiatives. However, unlike the univariate variable distributions which are manifest data qualities, multivariate normality is a situational characteristic, which is a function of the variate[20] and a statistical technique selection, which are obviously highly situational. These considerations will be discussed in more detail in the ensuing chapters.

Metadata

Somewhat tautologically defined, *metadata* is data about data. Operationally, it is a summary view of the individual variables contained in the dataset expressed in terms of the key statistical descriptors, such as value ranges and the corresponding central tendencies, the average amount of variability, as well as the assessment of coverage and accuracy. Although historically more familiar to academicians than to practitioners, the concept of

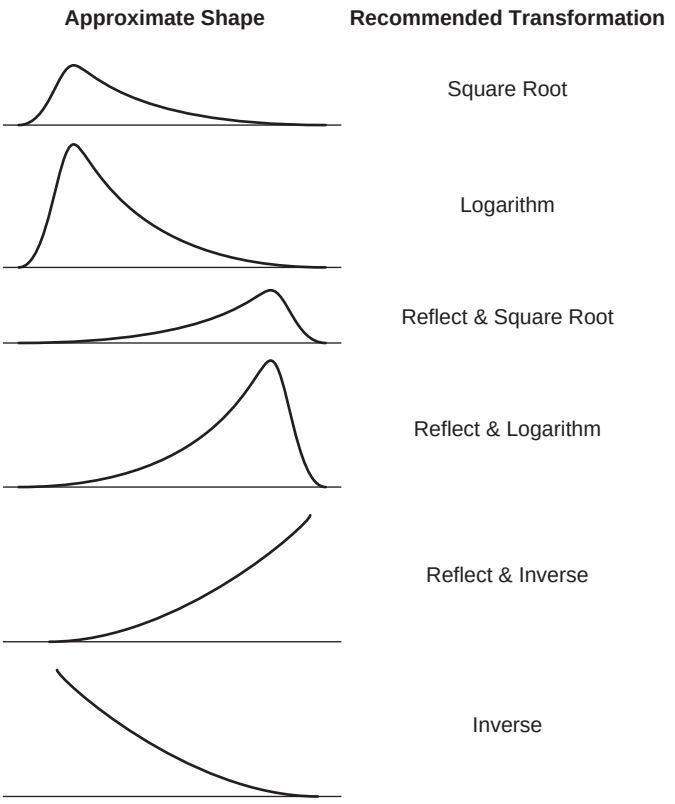

Figure 7.7 Common Normality Departures and Recommended Fixes

metadata is gaining popularity among the latter as the amount and diversity of data contained in corporate repositories continues to grow to the point of becoming overwhelming. In the world of corporate databases, even the well-annotated ones, i.e., those accompanied by clear data model descriptions and comprehensive data dictionaries, are often just a hodgepodge of some well and some sparsely populated variables or discontinued or definitionally amended metrics, with little indication as to what is good and what is not.

The primary value of metadata is the time savings it can deliver. For an analyst to have an a priori knowledge of the dataset means being able to take the right data engineering steps, which in turn enables one to focus on specific statistical techniques that can be supported by the available data. When the metadata information is not available, database analyses can become riddled with time-consuming and confidence-shaking corrective re-work, largely due to the underlying data challenges not having been discovered, and corrected, in a timely fashion.

Metadata Template

Although the type and the number of specific variables can differ considerably across datasets, the informational foundation of the metadata associated with each dataset is relatively constant. A general outline of a metadata template is shown below:

		Metrics								
		Mean	Median	St. Deviation	Skewness	Kurtosis	Min	Max	% Missing	Coding*
Behaviors	Transactions									
	Promotional responses									
	Customer-initiated actions									
Causes	Demographic descriptors									
	Lifestyle indicators									
	Purchase propensities									
	Transactional channels									
	Financial indicators									

* quantitative vs. qualitative

Figure 7.8 Sample Metadata Template

The sample template is shown at a relatively aggregate level, i.e., it illustrates variable types, rather than listing specific variables which are likely to differ across datasets. The focus of the metadata is on a comprehensive overview of the key characteristics of datasets as it relates to using the data contained therein as a foundation for the development of unique, competitively advantageous knowledge. The data evaluation embedded inside of a metadata template needs to differentiate between the two key data types discussed earlier: the behavioral core (i.e., "behaviors") and the augmenting causal drivers (i.e., "causes") to aid in specific statistical technique selection.[21] In practice, behavioral metrics, especially transactions, have the highest likelihood of outlying observations. For example, the UPC-scanner data captured by virtually all retailers utilizing electronic point-of-sales systems and barcoding usually contains outlying negative values, commonly due to refunds and/or exchanges. Unless removed, these values will not only skew the basic transactional characteristics, but may also confuse subsequent sales explanatory analytics.

As shown in the outline in Figure 7.8, the main focus of metadata is on the description of the basic distributional properties of the dataset with the help of measures of central tendency (mean and median), variability (standard deviation), distributional shape (skewness and kurtosis) and extremity (minimum and maximum). Combined, these evaluative statistics clearly describe the usefulness of individual variables as inputs into specific types of statistical analyses.

Further adding to that assessment are the remaining two variables: coverage (% missing) and data type (coding). The former captures the degree of missing values, while the latter qualifies variables' informational content as either qualitative (i.e., nominal or ordinal) or quantitative (i.e., interval or ratio). High incidence of missing values can make an individual variable—particularly, a behavioral one—not analytically suitable. On the other hand, qualitatively coded values, including 0–1 indicators (often called "dummy" codes) or category labels (such as "male–female" gender classification, "under-18, 18–44, 45–65, 66+" age group categorization or "high–medium–low" spending groupings) will limit the analytic usability of such-coded variables.[22] For instance, a qualitatively coded lifetime value (e.g., high–medium–low) will disqualify such variables from being employed as a dependent variable in linear regression.

Using Metadata

Metadata is compiled—what next? Evaluate the usefulness of the individual dataset variables in the context of the analytic plan, discussed in Chapter 3. Is the available data of sufficient quality to support the knowledge creation goals laid out in the analytic

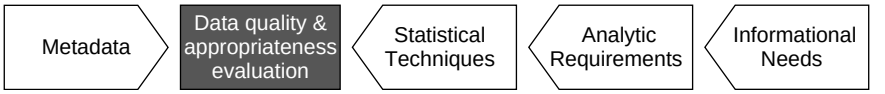

Figure 7.9 Using Metadata to Assess Analytic Preparedness

roadmap? In short, the main benefit of clear and concise metadata information is to evaluate the availability and quality of raw materials (i.e., data) prior to launching data analysis/model building efforts. Doing so entails working backwards from the desired informational outcomes, as shown in Figure 7.9.

In order to bring about the desired informational end state—i.e., the sought after knowledge—data needs to exhibit the desired quality and statistical characteristics in order to be analyzable in a valid and reliable fashion, which underscores the importance of an a priori preparation of an explicit analytic roadmap. The *data quality and appropriateness evaluation* is a process of contrasting the informational needs-dictated analytic requirements and the realities of the available data. Hence, the distribution, quality and coding constraints characterizing the available data (Metadata) are compared to Statistical Techniques' requirements, where the latter stem from Informational Needs-driven Analytic Requirements. In a sense, this is the assessment of the feasibility of the stated informational goals and metadata is the key enabler of this process.

Mini-Case 7.1: My Know-How and Your Data

Predictive analytics is a hot topic in database marketing. Broadly defined, it is the use of statistical modeling techniques for estimating the probability of future occurrences of outcomes of interest and/or magnitude of those outcomes. Its primary appeal is that it allows marketers to "cherry pick" consumers to be targeted with specific offers, by systematically assessing consumers' propensity to respond and then selecting those exhibiting the highest probability. Predictive analytics is particularly attractive to direct marketers, because of significant (on average, $1 to $10 per piece) variable costs associated with that marketing channel—it has been shown that the ability to focus direct mail campaigns on only the most prone-to-respond consumers can increase the response rate by 300% or more, which would have a multiplicative impact on the campaign ROI.

However, predictive analytics is a technically involved undertaking and given the often considerable amount of highly specialized expertise required to build and validate statistical models, many organizations opt to use outside suppliers, rather than relying on in-house capabilities. The outside suppliers, which can range from consulting firms to large global service organizations, contribute the analytic know-how, while the clients provide raw data. Given that, even those (suppliers) with a significant amount of experience in a particular industry need to invest at times considerable amount of time and effort in data due diligence to understand and fully account for any client-specific data peculiarities. However, even before any unique characteristics of data can be considered, the data needs to be properly

structured as an analytical dataset—let's take a closer look at a typical transaction data file sourced from grocery store infrared bar code scanners. Here are a few data records:

6|2011–10–15|1780012631|BENEFUL PLYFL LFE15.5LB|3000054|DOG FOOD DRY MOIST|4054007|PREMIUM|1|15.99|015636070|2614|21110161 2513725|2011–10–15|1113252147|ALPO PRM SL BEEF 13.2Z|3000055|DOG FOOD WET|4055002|PREMIUM NUTRITION|6|4.50|-.30

115766510|2604|21110161859092|2011–10–15|1780013462|BENEFUL HEALTHY WT 7LB|3000054|DOG FOOD DRY MOIST|4054007|PREMIUM|1|10.29|0555200|726|211102116861808|2011–10–16|3810013871|MST MTY RISE SHINE 72OZ|3000054|DOG FOOD DRY MOIST|4054001|NON PREMIUM|1|5.99|016025700|31|211032014379373|2011–03–19|1780040523|PUR PUPPY CHOW 17.6LB|3000054|DOG FOOD DRY MOIST|4054007|PREMIUM|1|13.99|018742990|38|211032014315270|2011–03–19|1780013468|BENEFUL HLTH RAD 15.5LB|3000054|DOG FOOD DRY MOIST|4054007|PREMIUM|1|15.99|0

The data shown above captures sales of dry and wet pet food. Although it is sourced from individual stores (ultimately, single check-out terminals), it is typically aggregated to a pre-determined level of geography, such as a region, and covers an agreed upon stretch of time, such as a year (the resultant file is almost always very voluminous, typically containing several hundred thousand or several million records). The raw data file exemplified above is formatted as a pipe (|) delimited text file with no column breaks and contains no column headers—those are usually made available separately and have to be matched with appropriate fields. Content-wise, there are 13 distinct variables (e.g., store ID, transaction ID, transaction date, UPC ID, product category ID, item price, markdowns, etc.), some of which are numeric (such as item price), others are alphanumeric (i.e., comprised of both letters and digits, such as store ID) and still others are formatted as dates. When the above exemplified data is read into an appropriate data analytical system, typically SAS, SPSS or R, the individual variable names need to be associated with the appropriate columns of data, so that the original pipe-delimited, heading-less file (where rows demark individual records but columns do not delimit individual variables) is transformed into a data matrix where row = records and columns = variables.