2
What Is Data Quality and Why Should We Care?

Caring about data quality is key to safeguarding and improving it. As stated, this sounds like a very obvious proposition. But can we, as the expression goes, “recognize it when we see it”? Considerable analysis and much experience make it clear that the answer is “no.” Discovering whether data are of acceptable quality is a measurement task, and not a very easy one. This observation becomes all the more important in this information age, when explicit and meticulous attention to data is of growing importance if information is not to become misinformation.

This chapter provides foundational material for the specifics that follow in later chapters about ways to safeguard and improve data quality.[1] After identifying when data are of high quality, we give reasons why we should care about data quality and discuss how one can obtain high-quality data.

Experts on quality (such as Redman [1996], English [1999], and Loshin [2001]) have been able to show companies how to improve their processes by first understanding the basic procedures the companies use and then showing new ways to collect and analyze quantitative data about those procedures in order to improve them. Here, we take as our primary starting point primarily the work of Deming, Juran, and Ishakawa.

2.1. When Are Data of High Quality?

Data are of high quality if they are “Fit for Use” in their intended operational, decision-making and other roles.[2] In many settings, especially for intermediate products, it is also convenient to define quality as “Conformance to Standards” that have been set, so that fitness for use is achieved. These two criteria link the

[1] It is well recognized that quality must have undoubted top priority in every organization. As Juran and Godfrey [1999; pages 4–20, 4–21, and 34–9] makes clear, quality has several dimensions, including meeting customer needs, protecting human safety, and protecting the environment. We restrict our attention to the quality of data, which can affect efforts to achieve quality in all three of these overall quality dimensions.

[2] Juran and Godfrey [1999].

role of the employee doing the work (conformance to standards) to the client receiving the product (fitness for use). When used together, these two can yield efficient systems that achieve the desired accuracy level or other specified quality attributes.

Unfortunately, the data of many organizations do not meet either of these criteria. As the cost of computers and computer storage has plunged over the last 50 or 60 years, the number of databases has skyrocketed. With the wide availability of sophisticated statistical software and many well-trained data analysts, there is a keen desire to analyze such databases in-depth. Unfortunately, after they begin their efforts, many data analysts realize that their data are too messy to analyze without major data cleansing.

Currently, the only widely recognized properties of quality are quite general and cannot typically be used without further elaboration to describe specific properties of databases that might affect analyses and modeling. The seven most commonly cited properties are (1) relevance, (2) accuracy, (3) timeliness, (4) accessibility and clarity of results, (5) comparability, (6) coherence, and (7) completeness.[3] For this book, we are primarily concerned with five of these properties: relevance, accuracy, timeliness, comparability, and completeness.

2.1.1. Relevance

Several facets are important to the relevance of the data analysts' use of data.

- Do the data meet the basic needs for which they were collected, placed in a database, and used?
- Can the data be used for additional purposes (e.g., a market analysis)? If the data cannot presently be used for such purposes, how much time and expense would be needed to add the additional features?
- Is it possible to use a database for several different purposes? A secondary (or possibly primary) use of a database may be better for determining what subsets of customers are more likely to purchase certain products and what types of advertisements or e-mails may be more successful with different groups of customers.

2.1.2. Accuracy

We cannot afford to protect against all errors in every field of our database. What are likely to be the main variables of interest in our database? How accurate do our data need to be?

[3] Haworth and Martin [2001], Brackstone [2001], Kalton [2001], and Scheuren [2001]. Other sources (Redman [1996], Wang [1998], Pipino, Lee, and Wang [2002]) provide alternative lists of properties that are somewhat similar to these.

For example, how accurate do our data need to be to predict:

- Which customers will buy certain products in a grocery store? Which customers bought products (1) this week, (2) 12 months ago, and (3) 24 months ago? Should certain products be eliminated or added based on sales trends? Which products are the most profitable?
- How will people vote in a Congressional election? We might be interested in demographic variables on individual voters – for example, age, education level, and income level. Is it acceptable here if the value of the income variable is within 20% of its true value? How accurate must the level of education variable be?
- How likely are individuals to die from a certain disease? Here the context might be a clinical trial in which we are testing the efficacy of a new drug. The data fields of interest might include the dosage level, the patient's age, a measure of the patient's general health, and the location of the patient's residence. How accurate does the measurement of the dosage level need to be? What other factors need to be measured (such as other drug use or general health level) because they might mitigate the efficacy of the new drug? Are all data fields being measured with sufficient accuracy to build a model to reliably predict the efficacy of various dosage levels of the new drug?

Are more stringent quality criteria needed for financial data than are needed for administrative or survey data?

2.1.3. *Timeliness*

How current does the information need to be to predict which subsets of customers are more likely to purchase certain products? How current do public opinion polls need to be to accurately predict election results? If data editing delays the publication/release of survey results to the public, how do the delays affect the use of the data in (1) general-circulation publications and (2) research studies of the resulting micro-data files?

2.1.4. *Comparability*

Is it appropriate to combine several databases into a data warehouse to facilitate the data's use in (1) exploratory analyses, (2) modeling, or (3) statistical estimation? Are data fields (e.g., Social Security Numbers) present within these databases that allow us to easily link individuals across the databases? How accurate are these identifying fields? If each of two distinct linkable databases[4] has an income variable, then which income variable is better to use, or is there a way to incorporate both into a model?

[4] This is illustrated in the case studies of the 1973 SSA-IRS-CPS exact match files discussed in Section 17.3 of this work.

2.1.5. *Completeness*

Here, by *completeness* we mean that no records are missing and that no records have missing data elements. In the survey sampling literature, entire missing records are known as *unit non-response* and missing items are referred to as *item non-response*. Both *unit* non-response and *item* non-response can indicate lack of quality. In many databases such as financial databases, missing entire records can have disastrous consequences. In survey and administrative databases, missing records can have serious consequences if they are associated with large companies or with a large proportion of employees in one subsection of a company. When such problems arise, the processes that create the database must be examined to determine whether (1) certain individuals need additional training in use of the software, (2) the software is not sufficiently user-friendly and responsive, or (3) certain procedures for updating the database are insufficient or in error.

2.2. Why Care About Data Quality?

Data quality is important to business and government for a number of obvious reasons. First, a reputation for world-class quality is profitable, a "business maker." As the examples of Section 3.1 show, high-quality data can be a major business asset, a unique source of competitive advantage.

By the same token, poor-quality data can reduce customer satisfaction. Poor-quality data can lower employee job satisfaction too, leading to excessive turnover and the resulting loss of key process knowledge. Poor-quality data can also breed organizational mistrust and make it hard to mount efforts that lead to needed improvements.

Further, poor-quality data can distort key corporate financial data; in the extreme, this can make it impossible to determine the financial condition of a business. The prominence of data quality issues in corporate governance has become even greater with enactment of the Sarbanes–Oxley legislation that holds senior corporate management responsible for the quality of its company's data.

High-quality data are also important to all levels of government. Certainly the military needs high-quality data for all of its operations, especially its counter-terrorism efforts. At the local level, high-quality data are needed so that individuals' residences are assessed accurately for real estate tax purposes.

The August 2003 issue of *The Newsmonthly of the American Academy of Actuaries* reports that the National Association of Insurance Commissioners (NAIC) suggests that actuaries audit "controls related to the completeness, accuracy, and classification of loss data". This is because poor data quality can make it impossible for an insurance company to obtain an accurate estimate of its insurance-in-force. As a consequence, it may miscalculate both its premium income and the amount of its loss reserve required for future insurance claims.

2.3. How Do You Obtain High-Quality Data?

In this section, we discuss three ways to obtain high-quality data.

2.3.1. Prevention: Keep Bad Data Out of the Database/List

The first, and preferable, way is to ensure that all data entering the database/list are of high quality. One thing that helps in this regard is a system that edits data before they are permitted to enter the database/list. Chapter 5 describes a number of general techniques that may be of use in this regard. Moreover, as Granquist and Kovar [1977] suggest, "The role of editing needs to be re-examined, and more emphasis placed on using editing to learn about the data collection process, in order to concentrate on preventing errors rather than fixing them."

Of course, there are other ways besides editing to improve the quality of data. Here organizations should encourage their staffs to examine a wide variety of methods for improving the entire process. Although this topic is outside the scope of our work, we mention two methods in passing. One way in a survey-sampling environment is to improve the data collection instrument, for example, the survey questionnaire. Another is to improve the methods of data acquisition, for example, to devise better ways to collect data from those who initially refuse to supply data in a sample survey.

2.3.2. Detection: Proactively Look for Bad Data Already Entered

The second scheme is for the data analyst to proactively look for data quality problems and then correct the problems. Under this approach, the data analyst needs at least a basic understanding of (1) the subject matter, (2) the structure of the database/list, and (3) methodologies that she might use to analyze the data. Of course, even a proactive approach is tantamount to admitting that we are too busy mopping up the floor to turn off the water.

If we have quantitative or count data, there are a variety of elementary methods, such as univariate frequency counts or two-way tabulations, that we can use. More sophisticated methods involve Exploratory Data Analysis (EDA) techniques. These methods, as described in Tukey [1977], Mosteller and Tukey [1977], Velleman and Hoaglin [1981], and Cleveland [1994], are often useful in examining (1) relationships among two or more variables or (2) aggregates. They can be used to identify anomalous data that may be erroneous.

Record linkage techniques can also be used to identify erroneous data. An extended example of such an application involving a database of mortgages is presented in Chapter 14. Record linkage can also be used to improve the quality of a database by linking two or more databases, as illustrated in the following example.

Example 2.1: Improving Data Quality through Record Linkage

Suppose two databases had information on the employees of a company. Suppose one of the databases had highly reliable data on the home addresses of the employees but only sketchy data on the salary history on these employees while the second database had essentially complete and accurate data on the salary history of the employees. Records in the two databases could be linked and the salary history from the second database could be used to replace the salary history on the first database, thereby improving the data quality of the first database.

2.3.3. Repair: Let the Bad Data Find You and Then Fix Things

By far, the worst approach is to wait for data quality problems to surface on their own. Does a chain of grocery stores really want its retail customers doing its data quality work by telling store managers that the scanned price of their can of soup is higher than the price posted on the shelf? Will a potential customer be upset if a price higher than the one advertised appears in the price field during checkout at a website? Will an insured whose chiropractic charges are fully covered be happy if his health insurance company denies a claim because the insurer classified his health provider as a physical therapist instead of a chiropractor? Data quality problems can also produce unrealistic or noticeably strange answers in statistical analysis and estimation. This can cause the analyst to spend lots of time trying to identify the underlying problem.

2.3.4. Allocating Resources – How Much for Prevention, Detection, and Repair

The question arises as to how best to allocate the limited resources available for a sample survey, an analytical study, or an administrative database/list. The typical mix of resources devoted to these three activities in the United States tends to be on the order of:

Prevent: 10%
Detect: 30%
Repair: 60%.

Our experience strongly suggests that a more cost-effective strategy is to devote a larger proportion of the available resources to preventing bad data from getting into the system and less to detecting and repairing (i.e., correcting) erroneous data. It is usually less expensive to find and correct errors early in the process than it is in the later stages. So, in our judgment, a much better mix of resources would be:

Prevent: 45%
Detect: 30%
Repair: 25%.

2.4. Practical Tips

2.4.1. Process Improvement

One process improvement would be for each company to have a few individuals who have learned additional ways of looking at available procedures and data that might be promising in the quest for process improvement. In all situations, of course, any such procedures should be at least crudely quantified – before adoption – as to their potential effectiveness in reducing costs, improving customer service, and allowing new marketing opportunities.

2.4.2. Training Staff

Many companies and organizations may have created their procedures to meet a few day-to-day processing needs, leaving them unaware of other procedures for improving their data. Sometimes, suitable training in software development and basic clerical tasks associated with customer relations may be helpful in this regard. Under other conditions, the staff members creating the databases may need to be taught basic schemes for ensuring minimally acceptable data quality.

In all situations, the company should record the completion of employee training in appropriate databases and, if resources permit, track the effect of the training on job performance. A more drastic approach is to obtain external hires with experience/expertise in (1) designing databases, (2) analyzing the data as they come in, and (3) ensuring that the quality of the data produced in similar types of databases is “fit for use.”

2.5. Where Are We Now?

We are still at an early stage in our discussion of data quality concepts. So, an example of what is needed to make data “fit for use” might be helpful before continuing.

Example 2.2: Making a database fit for use

Goal: A department store plans to construct a database that has a software interface that allows customer name, address, telephone number and order information to be collected accurately.

Developing System Requirements: All of the organizational units within the department store need to be involved in this process so that their operational needs can be met. For instance, the marketing department should inform the database designer that it needs both (1) a field indicating the amount of money each customer spent at the store during the previous 12 months and (2) a field indicating the date of each customer’s most recent purchase at the store.

Data Handling Procedures: Whatever procedures are agreed upon, clear instructions must be communicated to all of the affected parties within the department store. For example, clear instructions need to be provided on how to handle missing data items. Often, this will enable those maintaining the database to use their limited resources most effectively and thereby lead to a higher quality database.

Developing User Requirements – How will the data be used and by whom? All of the organizational units within the department store who expect to use the data should be involved in this process so that their operational needs can be met. For example, each unit should be asked what information they will need. Answers could include the name and home address for catalog mailings and billing, an e-mail address for sale alerts, and telephone number(s) for customer service. How many phone numbers will be stored for each customer? Three? One each for home, office, and mobile? How will data be captured? Are there legacy data to import from a predecessor database? Who will enter new data? Who will need data and in what format? Who will be responsible for the database? Who will be allowed to modify data, and when? The answers to all these questions impact the five aspects of quality that are of concern to us.

Relevance: There may be many uses for this database. The assurance that all units who could benefit from using the data do so is one aspect of the relevance of the data. One thing that helps in this regard is to make it easy for the store's employees to access the data. In addition, addresses could be standardized (see Chapter 10) to facilitate the generation of mailing labels.

Accuracy: Incorrect telephone numbers, addresses, or misspelled names can make it difficult for the store to contact its customers, making entries in the database of little use. *Data editing* is an important tool for finding errors, and more importantly for ensuring that only correct data enter the system at the time of data capture. For example, when data in place name, state, and Zip Code fields are entered or changed, such data could be subjected to an edit that ensures that the place name and Zip Code are consistent. More ambitiously, the street address could be parsed (see Chapter 10) and the street name checked for validity against a list of the streets in the city or town. If legacy data are to be imported, then they should be checked for accuracy, timeliness, and duplication before being entered into the database.

Timeliness: Current data are critical in this application. Here again, *record linkage* might be used together with external mailing lists, to confirm the customers' addresses and telephone numbers. Inconsistencies could be resolved in order to keep contact information current. Further, procedures such as real-time data capture (with editing at the time of capture) at the first point of contact with the customer would allow the database to be updated exactly when the customer is acquired.

Comparability: The database should capture information that allows the department store to associate its data with data in its other databases (e.g., a transactions database). Specifically, the store wants to capture the names, addresses, and telephone numbers of its customers in a manner that enables it to link its customers across its various databases.

Completeness: The department store wants its database to be complete, but customers may not be willing to provide all of the information requested. For example, a customer may not wish to provide her telephone number. Can these missing data be obtained, or *imputed*, from public sources? Can a nine-digit Zip Code be imputed from a five-digit Zip Code and a street address? (Anyone who receives mail at home knows that this is done all the time.) Can a home telephone number be obtained from the Internet based on the name and/or home address? What standard operating procedures can be established to ensure that contact data are obtained from every customer? Finally, *record linkage* can be used to eliminate duplicate records that might result in a failure to contact a customer, or a customer being burdened by multiple contacts on the same subject.

This simple example shows how the tools discussed in this book – data editing, imputation, and record linkage – can be used to improve the quality of data. As the reader will see, these tools grow in importance as applications increase in complexity.

Springer

http://www.springer.com/978-0-387-69502-0

Data Quality and Record Linkage Techniques
Herzog, Th.N.; Scheuren, F.J.; Winkler, W.E.
2007, XIV, 234 p., Softcover
ISBN: 978-0-387-69502-0