



Chapter 5

Visual Analytics for Investigating and Processing Data

Abstract In this chapter, we discuss how visual analytics techniques can support you in investigating and understanding the properties of your data and in conducting common data processing tasks. We consider several examples of possible problems in data and how they may manifest in visual representations, discuss where and why data quality issues can appear, and introduce a number of elementary data processing operations and computational techniques that help, in combination with visualisation, understand data characteristics and detect abnormalities.

5.1 Examples of data properties that may affect data analysis

In a perfectly organised world, all data that land on the desk of a data analyst, are collected and verified carefully, documented thoroughly, and cleaned from any occasional problems. The reality often differs from this description, unfortunately. We have seen many data sets that were collected by different people and organisations using different equipment, methods, and protocols. The data are often represented in different formats using different notations, making their fusion a challenging task. Erroneous and missing values are very usual in any data.

In this chapter, we shall write about data sets in general, irrespective of their specifics and representation. The following chapters will address different types of data in detail. So, we shall talk here about a general *dataset* consisting of multiple *data items*. Data items are composed of *fields*, which may contain values of attributes, references to entities, places, or times, or to items in another dataset. All data items in a dataset have homogeneous structure, i.e., consist of the same number of fields having the same meaning and containing the same kind of information. The fields usually have names. The contents of the fields are called the *values* of these fields. Some fields in a dataset may be empty. The absence of a value in a field may

have different meanings: either no value exists, or some value exists in principle but could not be determined. Knowing the meaning of the field can help understand what the absence of a value means. When this is not clear and not described in metadata, it is necessary to obtain additional information about the dataset, e.g., by contacting the data collector.

In many data sets, dummy values, like 999 or -1 , have special meanings, such as “missing value”, “anything else” (e.g., when values in a field are categories or classes), or “error”. Sometimes, different dummy values have the same meaning in a single data set, if it was prepared by different people or at different times. For example, both “n/a” and “-” may mean “not applicable”. It may also happen that the same dummy value has different meanings. A series of data records may lack consistency in measurement units (e.g. metres or kilometres), formatting (decimal dot or decimal comma), and representation of dates and times. If a data set has been collected over a long period of time, consistency may be lacking due to changes in equipment, policies, daily routines, or personal habits and preferences. Data items representing times may be inconsistent due to wrong time zones (e.g. if a tourist did not set a correct time zone in the photo camera after moving to a different continent) or ignoring switches to/from daylight saving time. Textual data components may be misspelled, contain abbreviations and jargon, texts in different languages, etc. The same meanings may be expressed using synonyms, thus complicating processing and analysis.

During processing, field values may lose their *precision*. It may be a result of insufficiently careful transformation of the data format (e.g., from a spreadsheet to a text file) or an attempt to decrease the size of a file with data. Figure 5.1 shows the impact of rounding up geographic coordinates from 5 to 2 decimals. The dots are rendered with 70% transparency for enabling the assessment of the densities. By comparing two maps, it is visible that, as a consequence of the displacement of the points, some real patterns disappear while artefacts, or fake patterns, emerge.



Fig. 5.1: The same set of 3,083 points (Twitter messages posted in London) is displayed with precision of 5 and 2 decimals (left and right).