Chapter 4
Types of Data Manipulations

Study Exercise – data cleaning issues

Use any reference that you might find useful, including our good friend Google.com, to study and draft an answer for the following question.

1. What are the differences and similarities between Error and Noise?
2. What are the difference and similarities between Errors and Missing values?
3. What are the difference and similarities between Errors and Outliers?
4. What are the common strategies to address Missing Values?
5. Do outliers always pose an issue for our analytics? When do they do so, and why?
6. Assume that we have performed some data manipulations to address one or all of the level 3 data cleaning issues. Should we call the data manipulations *Data Cleaning* or *Data Transformation*? Explain.

Answers to questions

Q1: What are the differences and similarities between Error and Noise?

Similarity: They are both issues with the data that we need to resolve.

Difference: Error is avoidable, but noise is not as noise happens due to random irregularities in the phenomena that we measure, but the error happens due to systematic mistakes in data collection.

Q2: What are the differences and similarities between Errors and Missing values?

Similarity: They are both issues with the data that we need to resolve.

Similarity: They both happen due to systematic problems in the data collection

Difference: For errors, we have a wrong measurement, but for missing values, we lost the window of opportunity to be able to measure.

Q3: What are the difference and similarities between Errors and Outliers?

In principle, they bear no similarities. The only similarity they have is that we use the same analytic tools to detect outliers and errors. Essentially, when we detect data objects too different from the rest of the population, we check if the values may have been errors, if not, they are outliers.

Q4: What are the common strategies to address Missing Values?

- Leave as as
- Drop the row
- Drop the column
- Impute a new value

Q5: Do outliers always pose an issue for our analytics? When do they do so, and why?

Most often they do not. In situations where the large numerical differences between the outlier and the rest of the population consume the analytics capabilities, then they need to be dealt with. For instance, see *Data Transformation in Action – An example with data* in *Chapter 3*.

Q6: Assume that we have performed some data manipulations to address one or all of the level 3 data cleaning issues. Should we call the data manipulations Data Cleaning or Data Transformation? Explain.

It doesn’t matter, we use these distinctive terms to learn about various data preparation that needs to be taken place. When it comes to action, one action may be doing more than one thing and we may not be able to find one distinctive name for that manipulation action.