## DEVELOPING A DATA DICTIONARY

The stakeholder approaches you for help in designing a survey that collects the following data:

1. Demographics, including age, level of education, income, state, postcode, suburb, and whether resident in Melbourne.
2. Number of stores respondents have visited.
3. Overall satisfaction with the products and services.
4. Customer perceptions of the value they receive for the price they pay for the products.
5. Customers' historic purchase frequency.
6. The likelihood customers will shop again at the store
7. Opportunities for improvement.

A data dictionary provides a centralised location to describe the data stored in a database. It supports the collection, analysis, and presentation of data and is integral to good database management because it supports a consistent understanding of the data elements and associated terminology. It provides guidelines for when users change or delete data, or populate new data, so improves reliability, dependability, and trustworthiness of data use.

A well-conceived data dictionary would help the analyst to define structures for the data elements that are valid and fit for further analysis, which is beneficial, as different analysis techniques require specific data formats, and having a well-conceived data dictionary makes it easier to access the required formats and structures.

A data dictionary would help the analyst or decision makers minimise data quality issues in advance and reduce the amount of data transformation effort required, because the codification and level of measurement can be predefined to fit the desired analysis. Having descriptions of variables helps avoid any confusion about the meaning of the variables, and also saves any additional investment of time in trying to understand what the variables and associated data elements mean.

For example, take the variable 'Perceived price value'. Perceived price value may be interpreted and measured differently by a member of the marketing team (in this instance, price value is likely to be measured via a survey) compared to a colleague in finance (who is likely to measure price value using financial information such as that relating to past purchases). With a clear definition of the meaning and measurement of Perceived price value, any confusion can be avoided. The data can also be easily used by anyone who was not involved in the data collection process and therefore ensures consistency in data understanding and use.

A data dictionary helps enhance data quality because it enables the enforcement of data entry standards, which ultimately leads to improved decision-making. A data dictionary can also make it easier for users to identify data errors and inconsistencies (e.g. by using distribution plots and checking the results against the data dictionary).

Examples of data dictionary:

| Variable name | Description | Question | Codification | Level of measurement | Explanation |
|---|---|---|---|---|---|
| Age | Respondent's age (in years) | What is your age? | Open-ended (e.g. 28, 54, 20) | Numerical – ratio | Age is usually classified as ratio data or ordinal data. It can be useful to bracket the data into age ranges (e.g. 'younger than 18 years old', '18–24 years old') |
| Education | Respondent's education level | What is your level of education? | • Doctoral D=degree (PhD)<br>• Master's degree<br>• Graduate diploma/certificate<br>• Bachelor's degree<br>• Diploma/Advanced diploma<br>• Certificate I–IV<br>• Senior secondary education (Year 11 and 12)<br>• Junior secondary education (Years 7–10)<br>• Never attended school<br>• Other<br>• I prefer not to say | Categorical - ordinal | Data that is ordinal in nature features a meaningful ranking among the data, but no clear measurement of the distances between values. |
| Income | Respondent's income level | What is your gross income (combined annual income from all | • Less than $24,999<br>• $25,000 – $49,999<br>• $50,000 – $99,999<br>• More than $100,000<br>• I prefer not to answer | Categorical - ordinal | Income is usually classified as ratio or ordinal data. |

| | | | | | |
|---|---|---|---|---|---|
| | | sources before tax)? | | | |
| State | State or territory of the respondent's residential address | In which state or territory do you live? | Open-ended | Categorical – nominal | The states represent different categories. They are measured on a nominal scale because no meaningful order can be derived. |
| Postcode | Postcode of the respondent's residential address | What is your postcode? | Open-ended, four digits | Categorical – nominal | Despite consisting of numbers, a postcode represents a specific category (region). The level of measurement of postcodes is usually nominal. Nominal data has no meaningful order and any numbers attributed to data values are simply for coding purposes. |
| Suburb | Name of the suburb where the respondent resides | What is the name of your suburb? | Open-ended | Categorical – nominal | Suburb names represent different categories. They are measured on a nominal level because no meaningful order can be derived |
| Satisfaction | Overall satisfaction with the store products and services | How satisfied are you with our products and services? | • Very dissatisfied<br>• Somewhat dissatisfied<br>• Neither satisfied nor dissatisfied<br>• Somewhat satisfied<br>• Very satisfied | Categorical – ordinal | Likert scales usually measure ordinal data. It is not an interval or a ratio level of measurement because we cannot interpret or compare the distance between values |
| PerceivedPriceValue | Respondent's perceived value for money of the store's products | How would you rate the value for money of the store's products? | • Very poor<br>• Poor<br>• Neutral<br>• Good<br>• Excellent | Categorical – ordinal | The level of measurement for this example is ordinal as a meaningful ranking of the data is possible |
| PurchaseFrequency | Number of purchases made by the | In the last 12 months, how often | • None 1–10 times<br>• 11–30 times | Categorical – ordinal | The level of measurement for this example is ordinal as a meaningful ranking of the data is possible |

| | | | | | |
|---|---|---|---|---|---|
| | respondent in the given period of 12 months | have you purchased from our store? | • 31–50 times<br>• Over 50 times | | |
| Improvement | Suggestions for products or services in need of improvement | What are your suggestions on how we can improve? | Open-ended (eg qualitative text comments) | Categorical – nominal | Suggestions represent different categories. They are measured on a nominal level because no meaningful order can be derived. |