Chapter 15

Data Preparation and Analysis Strategy
### Focus of this Chapter
- Preparing Data for Analysis

### Relationship to Previous Chapters
- Marketing Research Process (Chapter 1)
- Research Design Components (Chapter 3)

### Relationship to Marketing Research Process

1. Problem Definition
2. Approach to Problem
3. Research Design
4. Field Work
5. Data Preparation and Analysis
6. Report Preparation and Presentation

---

*Figure 15.1  Relationship of Data Preparation to the Previous Chapters and the Marketing Research Process*
Figure 15.2 Data Preparation: An Overview
Figure 15.3  Data Preparation Process

1. Preliminary Plan of Data Analysis
2. Questionnaire Checking
3. Editing
4. Coding
5. Transcribing
6. Data Cleaning
7. Selecting a Data Analysis Strategy
Questionnaire Checking

A questionnaire returned from the field may be unacceptable for several reasons.

- Parts of the questionnaire may be incomplete.
- The pattern of responses may indicate that the respondent did not understand or follow the instructions.
- The responses show little variance.
- One or more pages are missing.
- The questionnaire is received after the pre-established cut-off date.
- The questionnaire is answered by someone who does not qualify for participation.
Figure 15.4 Treatment of Unsatisfactory Responses

- Treatment of Unsatisfactory Responses
  - Return to the Field
  - Assign Missing Values
    - Substitute a Neutral Value
    - Casewise Deletion
    - Pairwise Deletion
  - Discard Unsatisfactory Respondents

Copyright © 2011 Pearson Education, Inc.  Chapter 15 - 7
Editing

Treatment of Unsatisfactory Results

- **Returning to the Field** – The questionnaires with unsatisfactory responses may be returned to the field, where the interviewers recontact the respondents.

- **Assigning Missing Values** – If returning the questionnaires to the field is not feasible, the editor may assign missing values to unsatisfactory responses.

- **Discarding Unsatisfactory Respondents** – In this approach, the respondents with unsatisfactory responses are simply discarded.
Coding

**Coding** means assigning a code, usually a number, to each possible response to each question. The code includes an indication of the column position (field) and data record it will occupy.

**Coding Questions**

Fixed field codes, which mean that the number of records for each respondent is the same and the same data appear in the same column(s) for all respondents, are highly desirable.

- If possible, standard codes should be used for missing data. Coding of structured questions is relatively simple, since the response options are predetermined.

- In questions that permit a large number of responses, each possible response option should be assigned a separate column.
Guidelines for coding unstructured questions:

- Category codes should be mutually exclusive and collectively exhaustive.
- Only a few (10% or less) of the responses should fall into the “other” category.
- Category codes should be assigned for critical issues even if no one has mentioned them.
- Data should be coded to retain as much detail as possible.
A **codebook** contains coding instructions and the necessary information about variables in the data set. A codebook generally contains the following information:

- column number
- record number
- variable number
- variable name
- question number
- instructions for coding
Coding Questionnaires

- The respondent code and the record number appear on each record in the data.
- The first record contains the additional codes: project code, interviewer code, date and time codes, and validation code.
- It is a good practice to insert blanks between parts.
Table 15.1 Restaurant Preference

<table>
<thead>
<tr>
<th>ID</th>
<th>PREFER.</th>
<th>QUALITY</th>
<th>QUANTITY</th>
<th>VALUE</th>
<th>SERVICE</th>
<th>INCOME</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>19</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>7</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
Table 15.2
SPSS Variable View of the Data of Table 15.1

<table>
<thead>
<tr>
<th></th>
<th>Name</th>
<th>Type</th>
<th>Width</th>
<th>Decimals</th>
<th>Label</th>
<th>Values</th>
<th>Missing</th>
<th>Columns</th>
<th>Align</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ID</td>
<td>Numeric</td>
<td>8</td>
<td>0</td>
<td>Respondent Number</td>
<td>None</td>
<td>None</td>
<td>8</td>
<td>Right</td>
<td>Scale</td>
</tr>
<tr>
<td>2</td>
<td>PREFERE</td>
<td>Numeric</td>
<td>8</td>
<td>0</td>
<td>Restaurant Preference</td>
<td>{1, Weak Preference}</td>
<td>None</td>
<td>11</td>
<td>Right</td>
<td>Scale</td>
</tr>
<tr>
<td>3</td>
<td>QUALITY</td>
<td>Numeric</td>
<td>8</td>
<td>0</td>
<td>Quality of Food</td>
<td>{1, Poor}...</td>
<td>None</td>
<td>10</td>
<td>Right</td>
<td>Scale</td>
</tr>
<tr>
<td>4</td>
<td>QUANTITY</td>
<td>Numeric</td>
<td>8</td>
<td>0</td>
<td>Quantity of Portions</td>
<td>{1, Poor}...</td>
<td>None</td>
<td>10</td>
<td>Right</td>
<td>Scale</td>
</tr>
<tr>
<td>5</td>
<td>VALUE</td>
<td>Numeric</td>
<td>8</td>
<td>0</td>
<td>Overall Value</td>
<td>{1, Poor}...</td>
<td>None</td>
<td>10</td>
<td>Right</td>
<td>Scale</td>
</tr>
<tr>
<td>6</td>
<td>SERVICE</td>
<td>Numeric</td>
<td>8</td>
<td>0</td>
<td>Restaurant Service</td>
<td>{1, Poor}...</td>
<td>None</td>
<td>10</td>
<td>Right</td>
<td>Scale</td>
</tr>
<tr>
<td>7</td>
<td>INCOME</td>
<td>Numeric</td>
<td>8</td>
<td>0</td>
<td>Annual Household Income</td>
<td>{1, Less than $20,000}</td>
<td>None</td>
<td>10</td>
<td>Right</td>
<td>Scale</td>
</tr>
<tr>
<td>Column Number</td>
<td>Variable Number</td>
<td>Variable Name</td>
<td>Question Number</td>
<td>Coding Instructions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>-----------------</td>
<td>---------------</td>
<td>-----------------</td>
<td>---------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>ID</td>
<td></td>
<td>1 to 20 as coded</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| 2             | 2               | Preference    | 1               | Input the number circled.  
|               |                 |               |                 | 1=Weak Preference  
|               |                 |               |                 | 7=Strong Preference |
| 3             | 3               | Quality       | 2               | Input the number circled.  
|               |                 |               |                 | 1=Poor  
|               |                 |               |                 | 7=Excellent |
| 4             | 4               | Quantity      | 3               | Input the number circled.  
|               |                 |               |                 | 1=Poor  
|               |                 |               |                 | 7=Excellent |
| 5             | 5               | Value         | 4               | Input the number circled.  
|               |                 |               |                 | 1=Poor  
|               |                 |               |                 | 7=Excellent |
| 6             | 6               | Service       | 5               | Input the number circled.  
|               |                 |               |                 | 1=Poor  
<p>|               |                 |               |                 | 7=Excellent |</p>
<table>
<thead>
<tr>
<th>Column Number</th>
<th>Variable Number</th>
<th>Variable Name</th>
<th>Question Number</th>
<th>Coding Instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7</td>
<td>Income</td>
<td>6</td>
<td>Input the number circled. 1 = Less than $20,000 2 = $20,000 to 34,999 3 = $35,000 to 49,999 4 = $50,000 to 74,999 5 = $75,000 to 99,999 6 = $100,00 or more</td>
</tr>
</tbody>
</table>
Figure 15.6 Data Transcription

Raw Data

- CATI/CAPI/Internet
- Key Punching via CRT Terminal
- Mark Sense Forms
- Optical Scanning
- Computerized Sensory Analysis

Verification: Correct Key Punching Errors

- Computer Memory
- Disks
- Magnetic Tapes

Transcribed Data
Figure 15.7 Selecting A Data Analysis Strategy

- Earlier Steps (1, 2, 3) of the Marketing Research Process
- Known Characteristics of Data
- Properties of Statistical Techniques
- Background & Philosophy of the Researcher
- Data Analysis Strategy
International Marketing Research

- The researcher should ensure that the data have been prepared in a comparable manner across countries or cultural units. This means that comparable procedures must be followed for checking questionnaires, editing, and coding.
- Certain adjustments might be necessary to make the data comparable across countries. For example, the data might have to be adjusted to establish currency equivalents or metric equivalents.
- Transformation of the data might be necessary to make meaningful comparisons and achieve consistent results.
Data generated by a large networked panel could be made accessible to members enabling the discussions to dynamically organize and reorganize within the panel.

Social media respondents co-create, and thus ‘respondents’ become ‘participants’ in a shared enterprise, retaining the rights to set the agenda rather than simply responding to it.
Marketing Research & Social Media (Cont.)

- Data collection involves the Web scraping process that first “crawls” the website to locate and identify the discussion topics, topic ID, topic starter and topic start date. It then uses topic ID to download posts and messages by the topic. It is important that when storing the messages in a database, quotes in posts from others’ text be removed to prevent double counting.
Marketing Research & Social Media (Cont.)

- Text coding and categorization involves human inspection of a random sample of text messages to understand the type of acronyms, shorthand and terminologies used and get a feel for the data. The human coder then develops a coding and categorization scheme and assists in the computer categorization of text until coding results are satisfactory.

- In Text mining and visualization, the coded text data are interpreted by matching with positive or negative comments with overlapping terms that link the positive and negative comments to mentions of the product/brand.
Discarding respondents after analyzing the data raises ethical concerns, particularly if this information is not fully disclosed in the written report.

The procedure used to identify unsatisfactory respondents and the number of respondents discarded should be clearly disclosed.

Although interpretations, conclusions, and recommendations necessarily involve the subjective judgment of the researcher, this judgment must be exercised honestly, free from any personal biases or agendas of the researcher or the client.
Using the Base module, out-of-range values can be selected using the SELECT IF command. These cases, with the identifying information (subject ID, record number, variable name, and variable value) can then be printed using the LIST or PRINT commands. If a formatted list is required, the SUMMARIZE command can be used.

IBM® SPSS® DATA COLLECTION is a complete suite of products for survey and marketing research. It enables you to quickly and efficiently acquire and clean data from a range of sources using an array of methods.

IBM® SPSS® Modeler is a versatile data and text analytics workbench that helps you build predictive models. It is rather intuitive and does not require programming knowledge.

We illustrate specific procedures using the data of Table 15.1.
SPSS Detailed Steps: Variable Respecification

1. Select TRANSFORM.
2. Click COMPUTE VARIABLE.
3. Type "overall" into the TARGET VARIABLE box.
4. Click "quality" and move it to the NUMERIC EXPRESSIONS box.
5. Click the "+" sign.
6. Click "quantity" and move it to the NUMERIC EXPRESSIONS box.
7. Click the "+" sign.
8. Click "value" and move it to the NUMERIC EXPRESSIONS box.
9. Click the "+" sign.
10. Click "service" and move it to the NUMERIC EXPRESSIONS box.
11. Click TYPE & LABEL under the TARGET VARIABLE box and type "Overall Evaluation." Click CONTINUE.
12. Click OK.
SPSS Detailed Steps: Variable Recoding

1. Select TRANSFORM.
2. Select RECODE INTO DIFFERENT VARIABLES.
3. Click income and move it to the INPUT VARIABLE → OUTPUT VARIABLE box.
4. Type "rincome" into the OUTPUT VARIABLE NAME box.
5. Type "Recoded Income" into the OUTPUT VARIABLE LABEL box.
6. Click the OLD AND NEW VALUES box.
7. Under OLD VALUES, on the left click RANGE. Type 1 and 2 in the range boxes. Under NEW VALUES, on the right click VALUE and type 1 into the value box. Click ADD.
8. Under OLD VALUES, on the left click VALUE. Type 3 in the value box. Under NEW VALUES, on the right click VALUE and type 2 into the value box. Click ADD.

9. Under OLD VALUES, on the left click VALUE. Type 4 in the value box. Under NEW VALUES, on the right click VALUE and type 3 in the value box. Click ADD.

10. Under OLD VALUES, on the left click RANGE. Type 5 and 6 in the range boxes. Under NEW VALUES, on the right click VALUE and type 4 in the value box. Click ADD.

11. Click CONTINUE.

12. Click CHANGE.

13. Click OK.
Excel

In Excel, the IF statement can be used to make logical checks and check out-of-range values. The IF statement can be accessed under the FORMULAS > INSERT FUNCTION > IF.

We illustrate specific procedures using the data of Table 15.1.
Excel Detailed Steps: Variable Respecification

1. Click on the cell H1.
2. Type "New Variable" in cell H1.
3. Type "+B2+C2+D2+E2" in cell H2.
4. Click on the "Accept formula value" symbol.
5. Next, right click on cell H2. An excel pop-up menu gets displayed.
6. Select the COPY menu item.
7. Next, select (highlight) cells H3 through H21.
8. Right click in any one of these highlighted cells. An excel pop-up menu gets displayed.
9. Select the PASTE menu item.
10. The values for the new variable are now displayed in the cells.
1. Click on the cell H1.
2. Type "RINCOME" in cell H1.
3. Carefully type the formula 
   "=IF(F2=6,4,IF(F2=5,4,IF(F2=4,3,IF(F2=3,2,IF(F2=2,1,IF(F2=1,1,1))))))" correctly in cell H2.
4. Click on the "Accept formula value" symbol.
5. Next, right click on cell H2. An excel pop-up menu gets displayed.
6. Select the COPY menu item.
7. Next, select (highlight) cells H3 through H21.
8. Right click in any one of these highlighted cells. An excel pop-up menu gets displayed.
9. Select the PASTE menu item.
10. The recoded values for the new variable are now displayed in the cells.
MINITAB

Minitab features control statements that enable the user to set the order of commands in a macro. The IF command allows implementation of different blocks of commands, including IF, ELSEIF, ELSE, and ENDIF.

SAS

Within BASE SAS the IF, IF-THEN, and IF-THEN ELSE can be used to select cases with missing or out-of-range values. The LIST statement is useful for printing suspicious input lines. SAS Enterprise Guide allows the user to identify missing or out-of-range values with the FILTER DATA tab within the QUERY and FILTER DATA task. The procedures MI and MIANALYZE in SAS/STAT also offer the capability of imputing missing values when a more sophisticated approach is required.
Exhibit 15.2
Instructions for Running Computerized Demos

For best results while viewing the SPSS and Excel demonstrations, ensure that the “Display” resolution of your computer is set to 1280 by 1024 pixels. While we give instructions for running SPSS demonstrations, those for Excel are very similar.

For each procedure, you can download either a single Adobe Acrobat Document (.pdf) or a set of HTML files contained in a folder. Downloading the Adobe Acrobat Document (.pdf) is simple and you will need Acrobat 9 or higher to run it.

To download HTML files, pick the folder with the appropriate name. Each folder will have several files. It is important you download all the files in a folder and save them in one separate folder. All the files in a folder are required to run the demonstration.
Exhibit 15.2 (Cont.)

Instructions for Running Computerized Demos

The file that you should select to run the demonstration movie is the one with the "_.htm" extension appended to its name. For example, if you want to run a demonstration of variable respecification on the data of Table 15.1 using IBM SPSS 19, then double click the file “IBM SPSS 19 Demo Movie 15 Variable Respecification v2.htm.” Once you double-click, Internet Explorer (or, your default Web browser) will be loaded, and the demonstration movie will start automatically. Note that the other three files that also need to be in the same folder are IBM SPSS 19 Demo Movie 15 Variable Respecification v2.cptx, IBM SPSS 19 Demo Movie 15 Variable Respecification v2.swf, and standard.js.
Exhibit 15.2 (Cont.)
Instructions for Running Computerized Demos

If you want to stop the demonstration movie at any specific point in the demonstration, simply click the button. The demonstration stops at that point. That button now changes form, and looks like . To continue viewing the demonstration from that point on, simply click the button. To fast forward the demonstration, you can click the button. Click it multiple times if you need to fast forward through longer intervals. To rewind the demonstration, simply click the button. Click it multiple times if you need to rewind through longer intervals. At any time, if you want to replay the demonstration, right from the beginning, then simply click the button. Finally, you can also move the slide , left or right to navigate through the demonstration. The slider achieves the same purpose as that of the fast forward and rewind buttons.
Acronym: Data Prep

The data preparation process may be summarized by the acronym DATA PREP:
- **D** ata consistency checks
- **A** djusting the data for missing values
- **T** ranscribing
- **A** nalysis strategy
- **P** ost fieldwork questionnaire checking
- **R** ecording numerical or alphanumerical values:
  - Coding
- **E** diting
- **P** reliminary plan of data analysis