Chapter 2: Data

Mix and Match

1. Variable Name: brand of car; Type: categorical; Cases: drivers

2. Variable Name: household income; Type: numerical; Cases: households

3. Variable Name: color preference; Type: categorical; Cases: consumers in focus group

4. Variable Name: customer counts; Type: numerical; Cases: outlets of retail chain

5. Variable Name: item size; Type: ordinal; Cases: unknown (could be stocks in stores or purchase amounts)

6. Variable Name: shipping cost; Type: numerical; Cases: unknown (could be a time series or could be the costs for various items or destinations)

7. Variable Name: stock price; Type: numerical; Cases: companies (though the question is vague)

8. Variable Name: number absent; Type: numerical; Time Series Frequency: days

9. Variable Name: Sex; Type: categorical; Cases: respondents in survey

10. Variable Name: Education; Type: ordinal; Cases: customers

True/False

11. False. Zip codes are numbers, but these numbers are used only for identification and would not have any numerical meaning.

12. True.

13. False. Cases are another name for the rows in a data table.

14. True.

15. True.

16. False. A row holds an observation.

17. False. A Likert scale is used for ordinal data.

18. True.

19. False. Aggregation collapses a table into one with fewer rows.

20. True.

Think About It

21. (a) The data are cross sectional.
(b) The variables are Whether the employee opened an IRA (categorical) and the Amount saved (numerical with dollars as the units).
(c) Did employees respond honestly, particularly when it came to the amount they reported to have saved?

22. (a) The data are cross sectional.
(b) The variables are Reaction to increase (categorical, or perhaps ordinal if asked to rate the chance of moving to another bank), Current balance and Other aspects of the customer that would be useful additions to the data. (Bank may not care if it loses unprofitable customers.)
(c) How many customers responded to the questionnaire? Were their responses about leaving the bank sincere?

23. (a) The data are cross sectional.
(b) The variable is the Service rating (ordinal most likely, using a Likert scale).
(c) With only 500 replying, are the respondents representative of the other guests?

24. (a) The data are cross sectional.
(b) The variables are Whether a coupon was used (categorical) and Purchase amount (numerical with dollars as the units).
(c) How were these homes chosen? Was there a time limit on redemption?

25. (a) The data are a time series.
(b) The variable is the Exchange rate of the US dollar to the Canadian dollar (numerical ratio of currencies).
(c) Are the fluctuations in 2011 typical of other years?

26. (a) The data are a time series.
(b) The variable is the Average time spent on the lot for ten car models (numerical for each model).
(c) Did dealers accurately report this information? Were all dealers surveyed, or just some of them? If it's a survey, did it concentrate more in some regions than others?

27. (a) The data are cross-sectional.
(b) The variables are the Quality of the graphics (categorical, perhaps ordinal from bad to good) and the Degree of violence (categorical, perhaps ordinal from none to too much).
(c) Did some of the participants influence the opinions of others?

28. (a) The data are cross sectional.
(b) The variables are Income (numerical with units in dollars), Sex (categorical), Location (categorical), Number of cards (numerical count) and Profit (numerical with dollars as the units, derived from other data).
(c) Why were these accounts sampled and not all of them?

29. (a) The data are cross sectional (though they could be converted to a time series).
(b) The variables are Name (categorical), Zip code (categorical), Region (categorical), Date of purchase (categorical or numerical, depending on the context; the company could compute the average length of time since the last purchase), Amount of purchase (numerical with dollars as the units) and Item purchased (categorical).

(c) Presumably the region was recorded from the zip code.

30. (a) The data are a time series.
(b) The variable is Vehicle type (categorical or perhaps ordinal as compact, regular, large and SUV)
(c) The mix of cars on the weekend may not be the same as on a weekday. Do employees get an accurate count since they have other things to do as well?

31. 4M Economic Time Series

Motivation

(a) Answers will vary, but should resemble the following:

> By merging the data, we can see how sales of Best Buy move along with the health of the general economy. If sales at Best Buy rise and fall with disposable income, we might question the health of this company if the government predicts a drop in the amount of disposable income.

Method

(b) A row in the data from FRED2 describes the level of disposable income in a month whereas a row in the company-specific data is quarterly, summarizing a quarter (3 months).

(c) The columns are both numbers of dollars, but with different multipliers. The national disposable income is in billions (so the value for January 2010 means that consumers have $10.958 trillion annually to spend). The quarterly sales are in millions (so Best Buy's net sales in the first quarter of 2010 were $3.036 billion).

Mechanics

(d) We can aggregate the monthly numbers into a quarterly number such as by taking an average (FRED2 will do this for you if you want to return to the web site). Alternatively, we could take the quarterly number and spread it over the months. That's a bit hard to do, so the first path is more common.

(e) Name the columns Net Sales ($ billion) and Disp Income ($ trillion) and scale as shown previously. That avoids many extraneous zeros if you were, for example, to label them all as dollars. The dates might best be recorded in a single column as, say, 2010:1, 2010:2, and so forth, or in the style shown in the following table.

(f) Here's the merged data table for 2010:

Quarter	Net Sales ($ billion)	Disp Income ($ trillion)
Jan-2010	$3.036	$32.974
Apr-2010	$3.156	$33.451
Jul-2010	$3.233	$33.721
Oct-2010	$4.214	$34.010

Message

(g) Sales at Best Buy rocket up in the fourth quarter (30% higher during the holiday season), but consumers don't have that much more money to spend. Looks like some people spend a lot more during the holidays, no surprise there!

32. 4M Textbooks

Motivation

(a) Various sources report that books cost about $100 per class. In 2003, U.S. Senator Charles E. Schumer of New York released a study showing that the average New York freshman or sophomore pays $922 for textbooks in a year. So reducing the cost 5% would save $46.10 a year and by 10% would save $92.20 a year.

Method

(b) Your table should have headings like these. You should use the names of the stores you shopped at if different from these. The first two columns are categorical, with the first identifying the book and the second giving the label. The two columns of prices are both numerical.

Book Title	Type	Price at Amazon	Price at B&N

(c) These will vary. Presumably, you've got five textbooks from your current classes. Hopefully, you've also got some other personal books. For popular books, you might consider books on one of the best-seller lists or those at the top of the lists offered on-line.

Mechanics

(d) You may have to change the list of books, particularly for textbooks. Some on line sites have a limited selection of these.

(e) You should include all of the relevant costs. Some Internet retailers add high shipping costs.

Message

(f) Again, answers will vary depending on the choice of books and the choice of stores. The key to notice is the value of comparison. Because you've got two prices for the same books, you can compare apples to apples and see whether one retailer is systematically cheaper than the other