Chapter 2 – Data

SECTION EXERCISES

SECTION 2.1

1. a) Each row represents a different house that was recently sold. It is best described as a case.
b) Including the house identifier, there are seven variables in each row.

2. a) Each row represents a different transaction (not customer or book). It is best described as a case.
b) Including the transaction identifier, there are eight variables in each row.

SECTION 2.2

3. a) House _ID is an identifier (special type of categorical); Neighborhood is categorical (nominal); Mail_ZIP is categorical (nominal – ordinal in a sense, but only on a national level); YR_BUILT is quantitative (units – year), but could also be treated as categorical (ordinal); FULL_MARKET_VALUE is quantitative (units – dollars); SFLA is quantitative (units – sq. ft.).
b) These data are cross-sectional. Each row corresponds to a house that recently sold so at approximately the same fixed point in time.

4. a) Transaction ID is an identifier (special type of categorical); Customer ID is an identifier (special type of categorical); Date is categorical or may be treated as numerical if redefined as how many days ago the transaction took place; ISBN is an identifier (special type of categorical); Price is quantitative (units – dollars); Coupon is categorical (simply nominal); Gift is categorical (simply nominal); Quantity is quantitative (unit – counts).
b) These data are cross-sectional. Each row corresponds to a transaction at a fixed point in time. However the date of the transaction has been recorded. Consequently, since a time variable is included the data could be reconfigured as a time series.

SECTION 2.3

5. It is not clear that the real estate data of Exercise 1 are from a designed survey or experiment. Rather, the real estate major's data set was derived from transactional data (on local home sales). The major concern with drawing conclusions from this data set is that we cannot be sure that the sample is representative of the population of interest (e.g., all recent local home sales or even all recent national home sales).

6. The student is using a secondary data source (from the Internet). The main concerns about using these data for drawing conclusions is that the data were collected for a different purpose (not necessarily for developing a stock investment strategy) and information about how, when, where and why these data were collected may not be available.

CHAPTER EXERCISES

7. **The news.** Answers will vary.

8. **The Internet.** Answers will vary.

9. **Sales.** The description of the study has to be broken down into its components in order to understand the study. *Who*–months; *What*–money spent on advertising ($ thousand) and sales ($ million); *When*–monthly from 2010–2012; *Where*–United States; *Why*–to compare money spent on advertising to sales; *How*–not specified; *Variables*–there are two quantitative variables–the amount of money spent on advertising, and sales and one identifier variable (Month). The data form a time series.

10. **Food store.** *Who*– who or what was actually sampled–existing stores; *What*– what is being measured–weekly sales ($), town population (thousands), median age of town (years), median income of town($), and whether or not the stores sell beer/wine; *When*–not specified; *Where*–United States; *Why*–the food retailer is interested in understanding if there is an association amongst these variables to help determine where to open the next store; *How*–how was the study conducted–data collected from their stores; *Variables*–what is the variable being measured– sales ($), town population (thousands), median age of town (years), median income of town($) are all quantitative. Whether or not the stores sell beer/wine is

categorical. Store is an identifier variable; *Source* – data are not from a designed survey or experiment; *Type* –data are cross-sectional; *Concerns*–none.

11. **Sales II.** *Who*– who or what was actually sampled–quarterly data from a major U.S. company; *What*–what is being measured–quarterly sales ($ million), unemployment rate (%), inflation rate (%); *When*–quarterly from 2010–2012; *Where*–United States; *Why*–to determine how sales are affected by the unemployment rate and inflation rate; *How*–how was the study conducted–not specified; *Variables*–what is the variable being measured–quarterly sales ($ million), unemployment rate (%), and inflation rate (%) which are quantitative. Quarter is an identifier variable; *Source* – data are not from a designed survey or experiment; *Type* – data are time series; *Concerns*–none.

12. **Arby's menu.** *Who*–Arby's sandwiches; *What*–type of meat, number of calories (in calories), and serving size (in ounces); *When*–not specified; *Where*–Arby's restaurants; *Why*–assess the nutritional value of the different sandwiches; *How*–information gathered on each of the sandwiches offered on the menu; *Variables*–the number of calories and serving size (ounces) are quantitative, and the type of meat which is categorical; *Source* – data are not from a designed survey or experiment; *Type* – data are cross-sectional; *Concerns*–none.

13. **MBA admissions.** *Who*–MBA applicants; *What*–sex, age, whether or not accepted, whether or not they attended, and the reasons for not attending (if they did not accept); *When*–not specified; *Where*–a school in the Northeastern United States; *Why*–the researchers wanted to investigate any patterns in female student acceptance and attendance in the MBA program; *How*–data obtained from the admissions office; *Variables*—sex, whether or not the students accepted, whether or not they attended, and the reasons for not attending if they did not accept (all categorical) and age (years) which is quantitative; *Source* – data are not from a designed survey or experiment; *Type*–data are cross-sectional; *Concerns*–none.

14. **MBA Admissions II.** *Who*–MBA students; *What*–each student's standardized test scores and GPA in the MBA program; *When*–2007–2012; *Where*–London; *Why*–to investigate the association between standardized test scores and performance in the MBA program over five years (2007–2012); *How*–not specified; *Variables*—standardized test scores and GPA, both quantitative variables; *Source* – data are not from a designed survey or experiment, data are available from student records; *Type*–there are two quantitative variables: test scores and GPA. Although the data are collected over five years, the purpose is to examine them as cross sections rather than time series; *Concerns*–none.

15. **Pharmaceutical firm.** *Who*–experimental participants; *What*–herbal cold remedy or sugar solution, and cold severity; *When*–not specified; *Where*–major pharmaceutical firm; *Why*–scientists were testing the effectiveness of an herbal compound on the severity of the common cold; *How*–scientists conducted a controlled experiment; *Variables*—there are 2 variables. Type of treatment (herbal or sugar solution) is categorical, and severity rating is quantitative; *Source* – data come from an experiment; *Type* – data are cross-sectional; *Concerns*–the severity of a cold might be difficult to quantify (beneficial to add actual observations and measurements, such as body temperature). Also, scientists at a pharmaceutical firm could have a predisposed opinion about the herbal solution or may feel pressure to report negative findings about the herbal product.

16. **Start-up company.** *Who*–customers of a start-up company; *What*–customer name, ID number, region of the country, date of last purchased, amount of purchase ($), and item purchased; *When*–present day; *Where*–not specified; *Why*–the company is building a database of customers and sales information; *How*–assumed that the company records the needed information from each new customer; *Variables*—there are 6 variables: name, ID number, region of the country, and item purchased which are categorical and date and amount of purchase ($) are quantitative; *Source* – data are not from a designed survey or experiment; *Type* – data are cross-sectional; *Concerns*–although region is coded as a number, it is still a categorical variable.

17. **Vineyards.** *Who*–vineyards; *What*–size (acres), number of years in existence, state, varieties of grapes grown, average case price ($), gross sales ($), and percent profit; *When*–not specified; *Where*–could be assumed to be the United States as state is recorded but not specifically stated; *Why*–business analysts hope to provide information that would be helpful to grape growers in the United States; *How*–not specified;

Variables—size of vineyard (acres), number of years in existence, average case price ($), gross sales ($), and percent profit are 5 quantitative variables. State and variety of grapes grown are categorical variables; *Source* – data come from a designed survey; *Type* – data are cross-sectional; *Concerns*–none.

18. Gallup Poll. *Who*–1,180 American voters; *What*–region (Northeast, South, etc.), age (in years), party affiliation, whether or not the person owned any shares of stock, and their attitude (scale 1 to 5) toward unions; *When*–not specified; *Where*–United States; *Why*–the information was gathered as part of a Gallup public opinion poll; *How*–telephone survey; *Variables*— there are 5 variables. Region (Northeast, South, etc.), party affiliation, and whether or not the person owned any shares of stock are categorical variables. Age (in years), and their attitude (scale 1 to 5) toward unions are quantitative variables; *Source* – data come from a designed survey; *Type* – data are cross-sectional; *Concerns*–none.

19. EPA. *Who*–every model of automobile in the United States; *What*–vehicle manufacturer, vehicle type (car, SUV, etc.), weight (probably pounds), horsepower (units of horsepower), and gas mileage (miles per gallon) for city and highway driving; *When*–the information is currently collected; *Where*–United States; *Why*–the EPA uses the information to track fuel economy of vehicles; *How*– among the data EPA analysts collect from the automobile manufacturers are the name of the manufacturer (Ford, Toyota, etc.), vehicle type…."; *Variables*— there are 6 variables. Vehicle manufacturer and vehicle type (car, SUV, etc.) are categorical variables. Weight (probably pounds), horsepower (units of horsepower), and gas mileage (miles per gallon) for both city and highway driving are quantitative variables; *Source* – data are not from a designed survey or experiment; *Type* – data are cross-sectional; *Concerns*–none.

20. Consumer Reports. *Who*–78 refrigerators; *What*–brand, cost (probably $), size (cu ft), type (such as top-freezer), estimated annual energy cost (probably $), overall rating (good, excellent, etc.), and repair history (in percent requiring repair over the past five years); *When*–2011; *Where*–United States; *Why*–the information was compiled to provide information to readers of Consumer Reports; *How*–not specified; *Variables*— there are 7 variables. Brand, type (such as top-freezer), and overall rating (good, excellent, etc.) are categorical variables. Cost (probably $), size (cu ft), estimated annual energy cost (probably $), and repair history (in percent requiring repair over the past five years) are quantitative variables; *Source* – some data (overall rating and repair history) likely come from a designed survey; *Type* – data are cross-sectional; *Concerns*–they may be a representative sample of refrigerators, or all of them, we don't know.

21. Lotto. *Who*–states in the United States; *What*–state name, whether or not the state sponsors a lottery, the number of numbers in the lottery (counts), the number of matches required to win (counts), and the probability of holding a winning ticket; *When*–not stated; *Where*–United States; *Why*–not specified but likely that the study was performed in order to compare the chances of winning the lottery in each state; *How*–not specified but data could be gathered from a number of different sources, such as the state lottery; *Variables*— there are 5 variables. State name, whether or not the state sponsors a lottery are categorical variables. The number of numbers in the lottery, the number of matches required to win, and the probability of holding a winning tickets are quantitative variables; *Source* – data are not from a designed survey or experiment; *Type* – data are cross-sectional; *Concerns*–none.

22. L.L. Bean. *Who*–LL Bean catalog recipients; *What*–number of catalogs mailed out, square inches in catalog, and sales ($ million) in 4 weeks following mailing; *When*–this information is currently reported; *Where*–LL Bean (United States); *Why*–to investigate association among catalog characteristics, timing, and sales; *How*–collect internal data; *Variables*— there are 3 variables. Number of catalogs, square inches in catalog, and sales ($ million) are all quantitative; *Source* – data are not from a designed survey or experiment; *Type* – data are cross-sectional; *Concerns*–none.

23. Stock market. *Who*–students in an MBA statistics class; *What*–total personal investment in stock market ($), number of different stocks held, total invested in mutual funds ($), and the name of each mutual fund; *When*–not specified; *Where*–a business school in the Northeastern United States; *Why*–the information was collected for use in classroom illustrations; *How*–an online survey was conducted, participation was probably required for all members of the class; *Variables*— there are 4 variables. Total personal investment in stock market ($), number of different stocks held, total invested in mutual funds ($) are quantitative

variables. The name of each mutual fund is a categorical variable; *Source* – data come from a designed survey; *Type* – data are cross-sectional; *Concerns*–none.

24. Theme park sites. *Who*–potential theme park locations; *What*–country of site, estimated cost (€), potential population size (counts), size of site (hectares), whether or not mass transportation within 5 minutes of site; *When*–2008; *Where*–Europe; *Why*–to to present to potential developers on the feasibility of various sites; *How*–not specified; *Variables*— there are 5 variables. Country of site and whether or not mass transportation within 5 minutes of site are both categorical variables. Estimated cost (€), potential population size (counts) and size of site (hectares) are quantitative; *Source* – data are not from a designed survey or experiment; *Type* – data are cross-sectional; *Concerns*–none.

25. Indy 2011. *Who*–Indy 500 races; *What*–year, winner, car model, time (hrs), speed (mph), and car number; *When*–1911-2011; *Where*–Indianapolis, Indiana; *Why*–examine trends in Indy 500 race winners; *How*–official statistics kept for each race every year; *Variables*— there are 6 variables. Winner, car model, and car number are categorical variables. Year, time (hrs) and speed (mph) are quantitative variables; *Source* – data are not from a designed survey or experiment; *Type* – data are time series; *Concerns*–none.

26. Kentucky Derby. *Who*–Kentucky Derby races; *What*–date, winner, winning margin (in lengths), jockey, winner's payoff ($), duration of the race (minutes and seconds), and track conditions; *When*–1875-2011; *Where*–Churchill Downs, Louisville, Kentucky; *Why*–examine trends in Kentucky Derby winners; *How*–official statistics kept for each race every year; *Variables*— there are 7 variables. Winner, winning jockey, and track conditions are categorical variables. Date, winning margin (in lengths), winner's payoff ($), and duration of the race (minutes and seconds) are quantitative variables; *Source* – data are not from a designed survey or experiment; *Type* – data are time series; *Concerns*–none.

27. Mortgages. Each row represents each individual mortgage loan. Headings of the columns would be: borrower name, mortgage amount.

28. Employee performance. Each row represents each individual employee. Headings of the columns would be: Employee ID Number (to identify the row instead of name), contract average ($), supervisor's rating (1-10), and years with the company.

29. Company performance. Each row represents a week. Headings of the columns would be: week number of the year (to identify each row), sales prediction ($), sales ($), and difference between predicted sales and realized sales ($).

30. Command performance. Each row represents a Broadway show. Headings of the columns would be: the show name (identifies the row), profit or loss ($), number of investors and investment total ($).

31. Car sales. Cross-sectional are data taken from situations that vary over time but measured at a single time instant. This problem focuses on data for September only which is a single time period. Therefore, the data are cross-sectional.

32. Motorcycle sales. Time series data are measured over time. Usually the time intervals are equally-spaced (e.g. every week, every quarter, or every year). This problem focuses on the number of motorcycles sold by the dealership in each month of 2012; therefore, the data are measured over a period of time and are time series data.

33. Cross sections. Time series data are measured over time. Usually the time intervals are equally-spaced (e.g. every week, every quarter, or every year). This problem focuses on the average diameter of trees brought to a sawmill in each week of a year; therefore, the data are measured over a period of time and are time series data.

34. Series. Cross-sectional are data taken from situations that vary over time but measured at a single time instant. This problem focuses on data for attendance of the third World Series game. Therefore, the data are cross-sectional.

Brief Case – Credit Card Bank

List the W's for these data:

Who – company cardholders
What – offer status (type of offer made to cardholder), credit card charges made by cardholder in August 2008, September 2008, and October 2008, marketing segment, industry segment, amount of spend lift after promotion, average spending on card pre- and post- promotion, whether or not cardholder is a retail customer or enrolled in the program and whether or not the spend lift was positive.
Why – to determine what types of offers are most effective in increasing credit card spending
When – most likely in 2008
Where – although not specified, most likely national data collected in U.S.
How – demographic data most likely collected when credit card account was opened and spending data collected during transactions

Classify each variable as categorical or quantitative; if quantitative identify the units:

Variables:
Offer Status – categorical
Charges August 2008 – quantitative ($)
Charges September 2008 – quantitative ($)
Charges October 2008 – quantitative ($)
Marketing Segment – categorical
Industry Segment – categorical
Spend Lift After Promotion – quantitative ($)
Pre Promotion Avg Spend – quantitative ($)
Post Promotion Avg Spend – quantitative ($)
Retail Customer – categorical
Enrolled in Program – categorical
Spend Lift Positive – categorical