Chapter 1 – Data and Decisions

SECTION EXERCISES

SECTION 1.1

1. **a)** Each row represents a different house that was recently sold. It can be described as a case.
b) There are six quantitative variables in each row plus a house identifier.

2. **a)** Each row represents a different transaction (not customer or book). It can be described as a case.
b) There are six quantitative variables plus two identifiers in each row.

SECTION 1.2

3. **a)** House_ID is an identifier (special type of categorical, not ordinal); Neighborhood is categorical (nominal); Mail_ZIP is categorical (nominal – ordinal in a sense, but only on a national level); Acres is quantitative (units – acres); Yr_Built is quantitative (units – year); Full_Market_Value is quantitative (units – dollars); Size 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 (categorical, nominal, not ordinal); Customer ID is an identifier (categorical, nominal); Date can be treated as quantitative (how many days since the transaction took place, days since Jan. 1 2012, for example) or categorical (as month, for example); ISBN is an identifier (categorical, nominal); Price is quantitative (units – dollars); Coupon is categorical (nominal); Gift is categorical (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 1.3

5. It is not specified whether or not the real estate data of Exercise 1 are obtained from a survey. The data would not be from an experiment, a data gathering method with specific requirements. 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). Therefore, we should be cautious about drawing conclusions from these data about the housing market in general.

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. In addition, the companies may not be representative of companies in general. Therefore, the student should be cautious about using this type of data to predict performance in the future.

CHAPTER EXERCISES

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

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

9. **Survey.** The description of the study has to be broken down into its components in order to understand the study. *Who*– who or what was actually sampled–college students; *What*–what is being measured–opinion of electric vehicles: whether there will more electric or gasoline powered vehicles in 2025 and the likelihood of whether they would purchase an electric vehicle in the next 10 years; *When*–current; *Where*–your location; *Why*–automobile manufacturer wants college student opinions; *How*–how was the study conducted–survey; *Variables*–what is the variable being measured–there is one categorical variable–what

students think about whether or not there will be more electric or gasoline powered vehicles in 2025 and one ordinal variable–how likely, using a scale, would the student be to buy an electric vehicle in the next 10 years; *Source* –the data are not from a designed survey or experiment; *Type*–the data are cross-sectional; *Concerns*–none.

10. Your survey. Answers may vary.

11. World databank. Answers will vary but chosen from the following possible indicators:
- GDP growth (annual %)
- GDP (current US$)
- GDP per capita (current US$)
- GNI per capita, Atlas method (current US$)
- Exports of goods and services (% of GDP)
- Foreign direct investment, net inflows (BoP, current US$)
- GNI per capita, PPP (current international $)
- GINI index
- Inflation, consumer prices (annual %)
- Population, total
- Life expectancy at birth, total (years)
- Internet users (per 100 people)
- Imports of goods and services (% of GDP)
- Unemployment, total (% of total labor force)
- Agriculture, value added (% of GDP)
- CO2 emissions (metric tons per capita)
- Literacy rate, adult total (% of people ages 15 and above)
- Central government debt, total (% of GDP)
- Inflation, GDP deflator (annual %)
- Poverty headcount ratio at national poverty line (% of population)

12. Diets R Us menu. *Who*–Diet R US; *What*–Number of calories, amount of fat (in grams), and amount of protein (in grams); *When*–not given; *Where*–Service area of the company; *Why*–asses the nutritional value of the different meals, *How*–information gathered on each of the meals offered on the company website menu; *Variables*–there are three variables: the number of calories, the amount of protein, and the amount of fat; *Source*–data are not from designed survey or experiment; *Type*–data are cross sectional; *Concerns*–none.

13. MBA admissions. *Who*–MBA applicants (in France); *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*–there are 5 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 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 (in Northeast); *What*–each student's standardized test scores and GPA in the MBA program; *When*–2009 to 2014; *Where*–outside of Paris; *Why*–to investigate the association between standardized test scores and performance in the MBA program over five years (2009–2014); *How*–not specified; *Variables*–there are 2 quantitative variables: standardized test scores and GPA; *Source*–data are not from a designed survey or experiment, data are available from student records; *Type*–although the data are collected over 5 years, the purpose is to examine them as cross-sectional rather than as time-series; *Concerns*–none.

15. Pharmaceutical firm. *Who*–experimental volunteers; *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. **World values survey.** *Who*–world population; *What*–cultural region, self-expression score, traditional values score; *When*–not specified; Where–countries all over the world; *Why*–to create a research data-base to study changes in values; *How*–by surveys in all countries; *Variables*–there are four variables described here: country and cultural region, which are categorical and scores for self-expression values and traditional values are quantitative; *Source*–data are from a designed survey; *Type*–data are cross-sectional (as well as time-series, but that latter aspect is not elaborated here); *Concerns*–none.

17. **Olive oil growers.** *Who*–Olive growers; *What*–gross sales; *When*–not given; *Where*–not specified; *Why*–to provide better services to the clients, *How*–information gathered from each of the olive tree growers; *Variable*–there are 5 variables: gross sales, percent profit, unit price, varieties, age, locality, and average production per tree; *Source*–data comes from a designed survey; there are 2 categorical variables variety and state ; *Type*–data are cross sectional; *Concerns*–none.

18. **OECD better life initiative.** *Who*–adult population; *What*–well-being, in specific, the contribution of educational aspect to it; *When*–2013; *Where*–OECD countries plus key partners; *Why*–provide inside of a comparative evaluation of well-being across major countries; *How*–survey; *Variables*–educational data refer to 3 quantitative variables: educational attainment score, reading skills score, and score for years in education; *Source*–data from official statistics; *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 are categorical variables; weight, horsepower, and gas mileage 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*–46 models of smart phones; *What*–brand, price (probably dollars), display size (probably inches) operating system, camera image size (megapixels), and memory card slot (yes/no); *When*–2013; *Where*–United States; *Why*–the information was compiled to provide information to readers of Consumer Reports; *How*–not specified; *Variables*–there are a total of 6 variables: price, display size and image size are quantitative variables; brand and operating system are categorical variables, and memory card slot is a nominal variable; *Source*–not specified; *Type*–the data are cross-sectional; *Concerns*–this many or may not be a representative sample of smart phones, or includes all of them, we don't know. This is a rapidly changing market, so their data are at best a snapshot of the state of the market at this time.

21. **Zagat.** *Who*–restaurants; *What*–% of customers liking restaurant, average meal cost ($), food rating (0–30), decor rating (0–30), service rating (0–30); *When*–current; *Where*–United States; *Why*–service to provide information for consumers; *How*–not specified; *Variables*–there are 5 variables: % liking and average cost are quantitative variables; ratings (food, decor and service) are ordered categories, therefore, ordinal variables; *Source*–not specified; *Type*–the data are cross-sectional.

22. **L.L. Bean.** *Who*–catalog mailings; *What*–number of catalogs mailed out, square inches in catalog, and sales ($ million) in 4 weeks following mailing; *When*–current; *Where*–L.L. Bean (United States); *Why*–to investigate association among catalog characteristics, timing, and sales; *How*–collection of internal data;

Variables–there are 3 variables: number of catalogs, square inches in catalog, and sales are all quantitative variables; *Source*–not specified; *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 northeast US; *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.

24. Theme park sites. *Who*–potential theme park locations in Europe; *What*–country of site, estimated cost (probably €), potential population size (counts), size of site (probably hectares), whether or not mass transportation within 5 minutes of site; *When*–2013; *Where*–Europe; *Why*–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 is within 5 minutes of site are both categorical variables; estimated cost, potential population size and size of site are quantitative variables; *Source*–data are not from a designed survey or experiment; *Type*–data are cross-sectional.

25. Indy 2012. *Who*–Indy 500 races; *What*–year, winner, car model, time (hrs), speed (mph), and car number; *When*–1911–2012; *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 and speed are quantitative variables; *Source*–all race results; *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, winner's payoff, and duration of the race are quantitative variables (date could be categorical as well); *Source*–race results; *Type*–data are time-series; *Concerns*–none.

27. Mortgages. Each row represents each individual mortgage loan. Headings of the columns would be: borrower's 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. Education in better life. Each row represents a country. Headings of the columns would be: country (to identify each row), overall education topic score, educational attainment score, reading skills score, and score for years in education.

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 2014; therefore, the data are measured over a period of time and are time series data.

33. OECD well-being. Cross-sectional are data taken from situations that vary over time but measured at a single time instant. This problem focuses on well-being data measured in OECD countries in 2013. Therefore, the data are cross-sectional.

34. Developments in well-being. 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 well-being data for one country in two years; therefore, the data are measured over a period of time and are time-series data.

Ethics in Action

Sarah's dilemma: The company RSPT Inc. is having Sarah compare their strategies to other companies. However, they could influence the outcome by funding the research and providing free software. In addition, Sarah may feel obliged to favor RSPT Inc. because they were generous in providing her research tools and funding. The company may put pressure on her to favor their methods over others because of their close relationship. The undesirable consequences are that the results are not completely objective and bias exists due to the funding circumstances. One possible solution would be to find other grants outside of RSPT Inc. but not connected to any of the companies being compared. This might also be true of the software. It is important in a scientific study to be completely objective and not have any influence by one of the clients being examined.

Jim's dilemma: Statistics and data can often be manipulated to produce a desired result that can "fudge" results and present a more desirable outcome. The scientific method is constructed to be objective if the rules are followed. The objective of Jim's study was to increase the percentage of clients who viewed their advisory services as outstanding, not increase the overall satisfaction average. In presenting an increased average, Jim is not being honest about the specific results of his study with respect to his objective. He should be honest about the decrease in the "outstanding" category.
One possible solution might be to compare the number of responses in each survey to see if there is a discrepancy that could explain the change. In addition, he could point out the large increase in the "above average" category (10% to 40%) which shows a huge improvement. Many people may be unwilling to give the highest rating on an intermediate basis but would be willing to identify an improvement.

For further information on the official American Statistical Association's Ethical Guidelines, visit:
http://www.amstat.org/about/ethicalguidelines.cfm
The Ethical Guidelines address important ethical considerations regarding professionalism and responsibilities.

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:
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