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 six 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); Neighbourhood is categorical (nominal); 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—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. The real estate data in Exercise 1 are not 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, and why they were collected may not be available. Also the dataset probably doesn't contain information about companies that went bankrupt. Investors would want to know about them so as to avoid investing in similar companies.

CHAPTER EXERCISES

7. Canadian labour force.

a) Someone on vacation from a full-time job is employed (they are on the payroll of their employer).

b) Someone who is not working, has a job offer, but is trying to find a better offer is unemployed because they are not working but are available and searching for work.

c) Someone who looked for work until six months ago but then gave up looking is not in the labour force because they are not searching for work.

8. Non-employment in Canada

a) The labour force

b) Unemployed and non-employed are both "available" for work. Unemployed are the subgroup that are actively "searching." So the answer is yes.

9. Domestic credit in Canada. Imagine collecting the data and putting it in a table with one row for each year and two columns corresponding to the two variables: domestic credit and GDP. The rows identify the *Who* of the data (i.e., the years) and the column headings identify the *What* (domestic credit and GDP). *When* is recent years, *Where* is Canada, and *Why* is to investigate possible future trends. The two variables of domestic credit and GDP are both quantitative and measured in $ billion. *Concerns*—none.

10. Oil spills. The description of the study has to be broken down into its components in order to understand it. *Who*—50 tankers having recent major oil spills. *What*—what is being measured–: date, spillage amount (no specified unit), and cause of puncture. *When*—recent years. *Where*—United States. *Why*—not specified but probably to determine whether spillage amount per oil spill has decreased. *How*—how was the study conducted not specified, although it is mentioned that the data are online. *Variables*—what is the variable being measured There are three variables—the date, the spillage amount (which is quantitative), and the cause of the puncture (which is categorical). *Concerns*—more detail needed on the specifics of the study.

11. Sales. The description of the study has to be broken down into its components in order to understand it. *Who*—who or what was actually sampled: months at a major Canadian company. *What*—what is being measured: money spent on advertising ($ thousands) and sales ($ million). *When*—monthly for the past three years; *Where* Canada (assumed). *Why*—to compare money spent on advertising to sales. *How*—how was the study conducted: not specified. *Variables*—what is the variable being measured There are three variables—the date, the amount of money spent on advertising (which is quantitative), and sales (which are quantitative). *Concerns*—none.

12. 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 the stores sell beer/wine. *When*—not specified. *Where*—Canada. *Why*—the food retailer is interested in understanding if there is an association among these variables in order to 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 ($), town population (thousands), median age of town (years), and median income of town ($), which are all quantitative. Whether or not the stores sell beer/wine is categorical. *Concerns*—none.

13. Sales II. *Who*—who or what was actually sampled: quarterly data from a major Canadian company. *What*—what is being measured: quarterly sales ($ millions), unemployment rate (%), inflation rate (%). *When*—quarterly for the past three years. *Where*—Canada. *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 ($ millions), unemployment rate (%), and inflation rate (%), which are quantitative. *Concerns*—none.

14. Subway's menu. *Who*—Subway sandwiches. *What*—type of meat, number of calories, and serving size (in ounces). *When*—not specified. *Where*—Subway restaurants. *Why*—to 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 is categorical. *Concerns*—none.

15. MBA admissions. *Who*—MBA applicants. *What*—sex, age, whether or not accepted, whether or not attended, and reasons for not attending (if they did not attend). *When*—not specified. *Where*—the school. *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 students accepted, whether or not they attended, the reasons for not attending (all categorical), and age (years), which is quantitative. *Concerns*—none.

16. **Climate.** *Who*—385 species of flowers. *What*—date of first flowering (in days). *When*—data gathered over the course of 47 years. *Where*—southern England. *Why*—the researchers wanted to investigate if the first flowering is indicating a warming of the overall climate. *How*—not specified. *Variables*—date of first flowering is a quantitative variable. *Concerns*–date of first flowering should be measured in days from January 1 to address leap year issues.

17. **MBA admissions II.** *Who*—MBA students. *What*—each student's standardized test scores and GPA in the MBA program. *When*—the past five years. *Where*—London. *Why*—to investigate the association between standardized test scores and performance in the MBA program over the past five years. *How*—not specified. *Variables*—standardized test scores and GPA, both quantitative variables. *Concerns*–none.

18. **Canadian schools.** *Who*—students. *What*—age (years or years and months), number of days absent, grade level, reading score, math score, and any disabilities/special needs. *When*—ongoing and current. *Where*–a Canadian province. *Why*—keeping this information is a provincial requirement. *How*—data are collected and stored as part of school records. *Variables*—there are seven variables. Grade level and disabilities/special needs are categorical variables. Number of absences, age (years or years and months), reading scores, and math scores are quantitative variables. *Concerns*—what tests are used to measure reading and math ability and what are the units of measurement?

19. **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 a herbal compound on the severity of the common cold. *How*—scientists conducted a controlled experiment. *Variables*—there are two variables. Type of treatment (herbal or sugar solution) is categorical, and severity rating is quantitative. *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.

20. **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 ($), and item purchased. *When*—present day. *Where*—Canada (assumed). *Why*—the company is building a database of customers and sales information. *How*—assumed that the company records the needed information from each new customer. *Variables*—there are six variables: name, ID number, region of the country, and item purchased, which are categorical, and date and amount of purchase ($), which are quantitative. *Concerns*—although region is coded as a number, it is still a categorical variable.

21. **Cars.** *Who*—cars parked in executive and staff lots at a large company. *What*—make, country of origin, type of vehicle (car, van, SUV, etc.), and age of vehicle (probably in years). *When*—not specified. *Where*—a large company. *Why*—not specified. *How*—data recorded in executive and staff lots of a large company. *Variables*—make, country of origin, and type of vehicle are three categorical variables. Age is the single quantitative variable. Whether or not the vehicle is in an executive or staff lot is also a categorical variable. *Concerns*—none.

22. **Canadian vineyards.** *Who*—vineyards. *What*—size, number of years in existence, province, varieties of grapes grown, average case price, gross sales, and percent profit. *When*—not specified. *Where*—Canada. *Why*—business analysts hope to provide information that would be helpful to grape growers in Canada. *How*—not specified. *Variables*—size of vineyard (hectares), number of years in existence, average case price ($), gross sales ($), and percent profit are five quantitative variables. Province and variety of grapes grown are categorical variables. *Concerns*—none.

23. **Environment.** *Who*—streams. *What*—name of stream, substrate of the stream (limestone, shale, or mixed), acidity of the water (measured in PH), temperature (degrees Celsius), and BCI (a measure of biological diversity—unknown units). *When*—not specified. *Where*—Alberta. *Why*—research conducted for an ecology class. *How*—not specified. *Variables*—there are five variables. Name of stream and substrate of the stream (limestone, shale, or mixed) are categorical variables. Acidity of the water (PH), temperature (degrees Celsius), and BCI are quantitative variables. *Concerns*—none.

24. Canadian voters. *Who*—1180 Canadian voters. *What*—region, age, party affiliation, whether or not the person owned any shares of stock, and their attitude toward unions. *When*—not specified. *Where*—Canada. *Why*—the information was gathered as part of a Gallup public opinion poll. *How*—telephone survey. *Variables*—there are five variables. Region (East, West, Prairie, etc.), party affiliation, and whether or not the person owned any shares of stock are categorical variables. Age (in years) and attitude (scale of 1 to 5) toward unions are quantitative variables. *Concerns*—none.

25. CTA. *Who*—all airline flights in Canada. *What*—type of aircraft, number of passengers, whether departures and arrivals were on schedule, and mechanical problems. *When*—the information is currently recorded. *Where*—Canada. *Why*—the information is required by the CTA. *How*—data is collected from airline flight information. *Variables*—there are four variables. Type of aircraft, whether departures and arrivals were on schedule, and mechanical problems are categorical variables. Number of passengers is a quantitative variable. *Concerns*—none.

26. Mobile phones. *Who*—mobile phone manufacturers. *What*—sales. *When*—past three years. *Where*—worldwide. *Why*—to project the future of the mobile phone business. *Variable*—sales, which is a quantitative variable measured in $. *Concerns*—none

27. Canadian families. *Who*—all Canadians (since it is a census). *What*—family type. *When*—every five years. *Where*—Canada. *Why*—to investigate social trends. *Variable*—family type, which is a categorical variable. *Concerns*—none

28. Canadian oil and gas production. *Who*—crude oil, natural gas, natural gas by-products. *What*—value and volume. *When*—every year. *Where*—Canada. *Why*–not specified. *Variables*—value and volume, both quantitative (measured in $ and m3, respectively). *Concerns*—none

29. Overnight visitors to Canada. *Who*—overnight visitors to Canada. *What*—number of nights spent in Canada and money spent in Canada. *When*—each year. *Where*—Canada. *Why*—to provide information for the tourism industry. *Variables*—number of nights spent in Canada and money spent in Canada, both of which are quantitative variables (with no units and with units of $, respectively). *Concerns*—none

30. 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 Toronto. *Why*—the information was collected for use in classroom illustrations. *How*—an online survey was conducted, and participation was probably required for all members of the class. *Variables*—there are four variables. Total personal investment in stock market ($), number of different stocks held, and total invested in mutual funds ($) are quantitative variables. The name of each mutual fund is a categorical variable. *Concerns*—none.

31. Theme park sites. *Who*—potential theme park locations. *What*—country of site, estimated cost (in euros), potential population size within one hour drive of site (counts), size of site (hectares), whether or not mass transportation is within five minutes of site. *When*—2017. *Where*—Europe. *Why*—present to potential developers on the feasibility of various sites. *How*—not specified. *Variables*—there are five variables. Country of site and whether or not mass transportation is within five minutes of site are both categorical variables. Estimated cost (€), potential population size (counts), and size of site (hectares) are quantitative. *Concerns*—none.

32. Indy. *Who*—Indianapolis 500 races. *What*—year, winner, car model, time (hrs), speed (mph), and car number. *When*—1911–2012. *Where*—Indianapolis, Indiana. *Why*—examine trends in Indianapolis 500 race winners. *How*—official statistics kept for each race every year. *Variables*—there are six variables. Winner, car model, and car number are categorical variables. Year, time (hrs), and speed (mph) are quantitative variables. *Concerns*—none.

33. Kentucky Derby. *Who*—Kentucky Derby races. *What*—date, winner, winning margin (in lengths), jockey, winner's payoff ($), duration of the race (minutes and seconds), and track conditions. *When*—1875–2012. *Where*—Churchill Downs, Louisville, Kentucky. *Why*—examine trends in Kentucky Derby winners. *How*—official statistics kept for each race every year. *Variables*—there are seven variables. Winner, winning jockey, and track conditions are categorical variables. Date, winning margin (in lengths), winner's payoff ($), and duration of the race (minutes and seconds) are quantitative variables. *Concerns*—none.

34. Mortgages. Each row represents each individual mortgage loan. Headings of the columns would be: borrower name and mortgage amount ($).

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

36. 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 ($).

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

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

39. 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 last year; therefore, the data are measured over a period of time and are time series data.

40. 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.

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

42. Canadian immigrants.

a) *Who*: years; *What*: percentage unemployment rates; *When*: 2009–2013; *Where*: Canada; *Why*: to compare unemployment rates among demographic groups; *How*: complied from the Statistics Canada Labour Force Survey.

b)

i. Year and Unemployment Rate (%) are quantitative. Sex, Education Level, and Immigrant Status are categorical.

ii. For a given year, the data are cross-sectional. Overall the data are time series.

iii. All data are secondary, since they are derived from the primary interview data in the Canadian Labour Force Survey.

43. Interpreting Published Data. Answers will vary.

Mini Case: Ottawa Senators

PLAN	Setup Clarify the objective.	Identify the types of data in the data file.
DO	Mechanics Describe the W's for the data.	**WHAT:** The answer comes from the column headings. The shooter's first and last names, the team, the shooter's positions, total shots, total goals, shooting percentage, and the number of game-deciding goals. **WHO:** The shooter, since this is what each row of the table is about. **WHERE:** The data are from the NHL for games played across North America. **WHEN:** Data entries were recorded after each game in the 2007–08 season. **WHY:** The data has been collected to record the results of NHL shootouts.
	Identify the types of variables that the data consists of.	**Variable Type** The shooter's name, team, and position are primary data, since they have not been processed or summarized. The rest of the data are secondary, since they are a summary of the results of shots taken during multiple games and have therefore been processed. The whole table is cross-sectional, since it all applies to the 2007–08 season. The data on the shooter's name, team, and position are categorical and the rest of the data are quantitative.
REPORT	Conclusion Summarize the results.	We have identified the W's and the variable types of our data. The data is partly primary and partly secondary. The data is cross-sectional. The data is partly categorical and partly quantitative.

Mini Case: Credit Card Company

PLAN	Setup: State the objective	To gain a clear understanding of the data available.
DO	Mechanics: *List the W's for these data:* *Classify each variable as categorical or quantitative; if quantitative identify the units:*	Large format tables and graphs (if any) are placed below this PLAN/DO/REPORT table *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* – not specified *How* – demographic data most likely collected when credit card account was opened and spending data collected during transactions *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
REPORT	Conclusion: State the conclusion in the context of the original objective	We have clarified what the data consist of and the details are given above.

Mini Case: Canadian Immigrants

PLAN	State the objective in its context.	Clarify the implications of the differences in unemployment rates of immigrants and people born in Canada
DO	Mechanics (a) Which data was Anjali referring to? (b) What other explanations are there of the data other than "Canadian employers are against immigrants." (c) What additional data do you suggest Statistics Canada should collect in order to clarify this issue?	 (a) The data on Canadian-born males indicate a marked difference in unemployment from 6.2–8.5% for high school graduates to 2.9–3.7%, for university graduates. However for immigrant females, the difference is only from 8.8–11.8% to 7.7–9.2%. (b) Employers could have clear reasons for not employing immigrants, other than being against them— e.g., poor knowledge of English/French, graduation from an overseas university with lower standards than the average Canadian university. Also some immigrants, such as doctors, are required to re-qualify in order to work in Canada, and many become unemployed in the interim. (c) Statistics Canada could collect data on the universities attended by people born in Canada and by immigrants, and the factors employers consider when hiring employees (in addition to their academic qualifications)
REPORT	Conclusion. State the conclusion in the context of the original objective.	The data clearly show higher unemployment rates for immigrants than for people born in Canada even when they have the same level of academic qualifications. In order to draw conclusions from these data, more detail is needed on whether the academic qualifications are comparable and what other factors employers consider.