SOLUTIONS: Chapter 1 – Statistics, Data & Decisions

1. **The news.** Answers will vary.

2. **TFSA.** *Who*—who was actually sampled: Canadian residents; *What*—what is being measured: TFSA participation rates (in percent) and amounts of contributions (in dollars); *When*—each year from 2009; *Where*—Canada; *Why*—to assess the usefulness of this investment option as one of many government programs for retirement planning; *How*—not specified; *Variables*—participation (Yes or No) which is categorical, and amount of contribution ($) which is quantitative; *Concerns*—how will the information be collected; from an audit of government records or with a polling company study?

3. **Oil spills.** The description of the study must be broken down into its components in order to understand the study. *Who*—tankers having recent oil spills; *What*—date, spillage amount (no specified unit), and cause of puncture; *When*—recent years; *Where*—not specified; *Why*—not specified, but probably to determine whether or not spillage amount per oil spill has decreased as a result of improvements in new tanker design; *How*—not specified, although it is mentioned that the data are online; *Variables*—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.

4. **Sales.** The description of the study must be broken down into its components in order to understand the study. *Who*—months at a major Canadian company; *What*—money spent on advertising ($ thousand) and sales ($ million); *When*—monthly for the past three years; *Where*—Canada (assumed); *Why*—to compare money spent on advertising to sales; *How*—not specified; *Variables*—there are three variables: the date; the amount of money spent on advertising, which is quantitative; and sales, which is quantitative.

5. **Food store.** *Who*—existing stores; *What*—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*—Canada (assumed); *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*—data collected from their stores; *Variables*—sales ($), town population (thousands), median age of town (years), median income of town($), which are all quantitative. Whether or not the stores sell beer/wine is categorical.

6. **Sales, part 2.** *Who*—quarterly data from a major U.S. company; *What*—quarterly sales ($ million), 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*—not specified; *Variables*—quarterly sales ($ million), unemployment rate (%), and inflation rate (%) all of which are quantitative.

7. **Subway menu.** *Who*—Subway sandwiches; *What*—type of meat, number of calories (in calories), and serving size (in ounces); *When*—not specified; *Where*—Subway 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 (grams), which are quantitative; and the type of meat, which is categorical.

8. **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 business school in Canada; *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.

9. **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.

10. **MBA admissions, part 2.** *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 five years; *How*—not specified; *Variables*—standardized test scores and GPA, which are both quantitative variables.

11. **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 collected and stored as part of school records; *Variables*—there are six 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?

12. **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 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.

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

14. **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 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.

15. **Vineyards.** *Who*—vineyards; *What*—size (hectares), number of years in existence, province, varieties of grapes grown, average case price ($), gross sales ($), and percent profit; *When*—not specified; *Where*—assume Canada since province is recorded; *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 quantitative variables. Province and variety of grapes grown are categorical variables.

16. **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*—British Columbia; *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 (a measure of biological diversity—unknown units) are quantitative variables.

17. **Environics Poll.** *Who*—1180 Canadian voters; *What*—region (West, Prairies, 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*—Canada; *Why*—the information was gathered as part of an Environics public opinion poll; *How*—telephone survey; *Variables*—there are five variables. Region (West, Prairies, 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.

18. **TSB.** *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 (assumed); *Why*—the information is required by the TSB; *How*—data are 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.

19. **Cellphones.** *Who*—every cellphone user in Canada; *What*—cellphone manufacturer and model, and demographic characteristics of the users; *When*—the information is currently collected; *Where*—Canada; *Why*—all companies involved in telecommunications sales and marketing will use the information to design advertising campaigns, pricing, applications, etc.; *How*—the Print Measurement Bureau collects the data; *Variables*—there are six variables. Cellphone manufacturer and model are categorical variables. Some demographic characteristics will be categorical (e.g., gender, marital status, education, place of residence) and others will be quantitative (e.g., income, age). Others could be either depending on how they are collected (e.g., income).).

20. **Consumer Reports.** *Who*—46 smartphones; *What*—brand, price, display size, operating system, camera pixel count, and memory card slot; *When*—A recent year (not specified); *Where*—North America; *Why*—the information was compiled to provide information to readers of *Consumer Reports*; *How*—not specified; *Variables*—there are six variables. Brand, operating system, and memory card slot (yes or no) are categorical variables. Price (probably $), display size (mm), and pixel count (megapixels) are quantitative variables.

21. **Gasoline prices.** *Who*—major Canadian cities; *What*—city name, average price of unleaded gasoline ($ per litre) per city per month for a year; *When*—current year (assumed); *Where*—Canada; *Why*—not specified, but likely to examine geographical disparity in price and changes over time; *How*—not specified, but data were collected by Statistics Canada; *Variables*—City name is an identifier variable; each of the monthly prices ($ per litre) is a quantitative variable.

22. **Lands' End.** *Who*—Lands' End catalogue recipients; *What*—number of catalogues mailed out, square inches in catalogue, and sales ($ million) in four weeks following mailing; *When*—this information is currently reported; *Where*—North America; *Why*—to investigate association among catalogue characteristics, timing, and sales; *How*—collect internal data; *Variables*—there are three variables. Number of catalogues, square inches in catalogue, and sales ($ million) are all quantitative.

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 Ontario; *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 four 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.

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 five minutes of site; *When*—2008; *Where*—Europe; *Why*—to 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 transit is within five minutes of site are categorical variables. Estimated cost (€), potential population size (counts), and size of site (hectares) are quantitative variables.

25. **Indy.** *Who*—Indy 500 races; *What*—year, winner, car model, time (hrs), speed (mph), and car number; *When*—1911–2016; *Where*—Indianapolis, Indiana; *Why*—examine trends in Indy 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.

26. **The Queen's Plate.** *Who*—Queen's Plate races; *What*—year, winner, winning margin (in lengths), jockey, winner's payoff ($), duration of the race (minutes and seconds), and track conditions; *When*—1957–2015; *Where*—Woodbine Racetrack, Etobicoke, Ontario; *Why*—examine trends in Queen's Plate winners; *How*—official statistics kept for each race every year; *Variables*—there are seven variables. Winner, winning jockey, and track conditions are categorical variables. Year, winning margin (in lengths), winner's payoff ($), and duration of the race (minutes and seconds) are quantitative variables.

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.** Data taken from situations that vary over time but are measured at a single time instant are said to be a cross-section of the time series. 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 paced (e.g., every week, every quarter, or every year). This problem focuses on the number of motorcycles sold by the dealership in each month last year; therefore, the data are measurements 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 measurements over a period of time and are time-series data.

34. **Grey Cup.** Cross-sectional are data taken from situations that vary over time but measured at a single time instant is said to be a cross-section of the time series. This problem focuses on data for attendance at a single Grey Cup game. Therefore, the data are cross-sectional.

35. **Bullying.** "Bullying" is a broad term, and is not defined here. Two decimal places in the reported percentage is spurious accuracy; it would be sufficient to report the percentage to the nearest whole number (40%) for the newspaper's readers.

36. **University grades.** The discrepancy is likely due to bias from self-selection: students with low grades are less likely to respond to the online survey. As well, there may be misreporting due to faulty memory or misplaced optimism! Note also another instance of spurious accuracy; the average grade should be reported to the nearest whole number (74%).

37. **Financial data.** The student is using a secondary data source (from the Internet). The data are not from a designed experiment, a data gathering method with specific requirements. 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.

38. **Bookstore.**
    a. Each row represents a different transaction (not customer or book). It can be described as a case.
    b. There are 8 variables including two identifiers in each row, 6 of the variables are quantitative.
    c. 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 2017, 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).
    d. 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.

Mini-Case Study Project – Credit Card Company

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

Mini-Case Study Project—Edmonton Oilers

Identifier Variables
1. Uniform #
2. Player Name

Categorical Variables
1. Flag
2. Position
3. Shoots

Quantitative Variables:
1. Age (yrs)
2. Height (in). In the spreadsheet height is a "string" variable, e.g., 6–1 means 6 ft., 1 in. Some processing is required to turn the string into height in inches.
3. Weight (lb)
4. YrsExp (yrs). In the spreadsheet, the letter R is used instead of 0 for a player in his first year. All Rs will need to be replaced by 0s to make it a quantitative variable.
5. GP, G, A, PTS, PIM, +/-, Salary ($), EV, PP, SH, GW, S. All except Salary are reported as #s
6. S% (%)
7. TOI (min), ATOI (min)

Date or Time
1. Birth Date. This can be used to compute current age, which is quantitative.
2. ATOI. The spreadsheet records it as min:sec. Some processing is required to turn it into minutes.

String:
1. Summary. This is a text version of G, A, PTS (except for goalies), which are also listed separately as quantitative variables.

Mini-Case Study Project—Student Survey

Classify each variable as categorical or quantitative. If quantitative, identify units where possible.

Categorical Variables

1. Sex: Q2
2. Country of birth: Q4
3. Ethnicity: Q5
4. Any paid part-time work: Q6
5. Holiday spending: Q8
6. Alcohol use: Q9
7. Tattoos: Q11
8. Course expectation: Q14
9. Sleep duration: Q17
10. Personal electronics: Q19 (There are actually seven separate binary variables here, one for each device)
11. Facebook user: Q20
12. Facebook status: Q20b
13. Twitter user: Q21

Quantitative Variables:

1. Height: Q3 (cm or in.)
2. Part-time work earnings: Q6a ($)
3. Recreational spending: Q7 ($)
4. Beauty: Q10 (no clear units; perhaps units of beauty)
5. Grade in prerequisite: Q12 (%)
6. Expected grade in this course: Q13 (%)
7. Study time: Q15 (hrs)
8. Missed lectures: Q16 (#; here "count" is used as a quantitative variable, ranging from 0 up)
9. Physical activity: Q18 (hrs)
10. Facebook frequency: Q20a (#)
11. Online time: Q22 (hrs)
12. Cell phone usage: Q23 (#)—four variables, Q24 (#)—two variables

Notes: Q1 is neither, but can be used to compute Age, which is quantitative (yrs). It is acceptable to say that Q10 is categorical, but it will more likely be analyzed as quantitative.