

Chapter 2 – Data

1. 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 6 months ago, but then gave up looking is not in the labour force because they are not searching for work.

2. 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, WHY is to investigate possible future trends. The two variables, domestic credit and GDP, are both quantitative and measured in \$bn.

3. Oil spills. The description of the study has to be broken down into its components in order to understand the study. *Who*–50 tankers having recent 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 or not spillage amount per oil spill has decreased since Congress passed the 1990 Oil Pollution Act and use that information in the design of new tankers; *How*–how was the study conducted–not specified, although it is mentioned that the data is online; *Variables*–what is the variable being measured–there are 3 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 has to be broken down into its components in order to understand the study. *Who*– who or what was actually sampled–months at a major Canadian company; *What*–what is being measured–money spent on advertising (\$ thousand) and sales (\$ million); *When*–monthly for the last 3 years; *Where*–United States (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 3 variables–the date, the amount of money spent on advertising which is quantitative, and sales which is quantitative; *Concerns*–none.

5. 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(\$), which are all quantitative. Whether or not the stores sell beer/wine is categorical; *Concerns*–none.

6. 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 2004–2006; *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; *Concerns*–none.

7. Subway's menu. *Who*–Subway'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; *Concerns*–none.

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 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; *Concerns*—none.
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 II.** *Who*—MBA students; *What*—each student's standardized test scores and GPA in the MBA program; *When*—2000–2005; *Where*—London; *Why*—to investigate the association between standardized test scores and performance in the MBA program over the last five years; *How*—not specified; *Variables*—standardized test scores and GPA, both quantitative variables; *Concerns*—none.
11. **Schools.** *Who*—students; *What*—age (years or years and months), race or ethnicity, 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 state requirement; *How*—data collected and stored as part of school records; *Variables*—there are 7 variables. Race or ethnicity, 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 2 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 6 variables: name, ID number, region of the country, and item purchased which 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 3 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.
15. **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*—assume United States as state is recorded; *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; *Concerns*—none.

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*—Alberta; *Why*—research conducted for an ecology class; *How*—not specified; *Variables*— there are 5 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; *Concerns*—none.
17. **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; *Concerns*—none.
18. **FAA.** *Who*—all airline flights in the United States; *What*—type of aircraft, number of passengers, whether departures and arrivals were on schedule, and mechanical problems; *When*—the information is currently recorded; *Where*—United States (assumed); *Why*—the information is required by the CTA; *How*—data is collected from airline flight information; *Variables*— there are 4 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.
19. **Mobile Phones.** *Who*—mobile phone manufacturers; *What*—sales; *When*—last 3 years; *Where*—world wide; *Why*—to project the future of the mobile phone business. Variable—sales which is a quantitative variable measured in \$.
20. **Canadian Families.** *Who*—all Canadians (since it is a census); *What*—family type; *When*—every 5 years; *Where*—Canada; *Why*—to investigate social trends. Variable—family type which is a categorical variable.
21. **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.
22. **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 units of \$ respectively.
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 Toronto; *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; *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*—2010; *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 within 5 minutes of site are both categorical variables. Estimated cost (€), potential population size (counts) and size of site (hectares) are quantitative; *Concerns*—none.
25. **Indy.** *Who*—Indy 500 races; *What*—year, winner, car model, time (hrs), speed (mph), and car number; *When*—1911-2007; *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; *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*—1875-2004; *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; *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 is 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-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 measure 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 measure 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 is said to be a cross-section of the time series. This problem focuses on data for attendance of the third World Series game. Therefore, the data are cross-sectional.

MiniCase: Ottawa Senators

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