# Interpreting and Analysing Data

## Lesson objectives

- I can distinguish between primary and secondary sources of data
- I can distinguish between microdata and aggregate data
- I can collect and analyse data from primary and secondary sources
- I can collect and analyse data obtained through experimentation

1.1

Lesson objectives | Teachers' notes | Lesson notes

MHR Page 230 #s 1, 4 & 6abc

## Definitions

**Primary Source Data**

- Data that have been collected directly by the researcher and have not been manipulated or summarized

**Microdata**

- An individual set of data about a single respondent

**Secondary Source Data**

- Data used by someone other than those who actually collected them

**Aggregate Data**

- Data that are combined or summarized in such a way that the individual microdata can no longer be determined

## Example 1

**Interpreting Data from Statistics Canada**

The table shows the average domestic airfares for 10 Canadian cities.

| City | 2010 Dollars | 2011 Dollars | 2010 to 2011 % Change |
|---|---|---|---|
| Canada | 182.5 | 190.7 | 4.5 |
| Calgary | 165.5 | 176.2 | 6.5 |
| Edmonton | 160.8 | 170.0 | 5.7 |
| Halifax | 172.0 | 179.3 | 4.2 |
| Montréal | 191.1 | 194.1 | 1.6 |
| Ottawa | 196.0 | 194.8 | –0.6 |
| Regina | 168.1 | 177.8 | 5.8 |
| Saskatoon | 170.2 | 178.8 | 5.1 |
| Toronto | 205.2 | 214.9 | 4.7 |
| Vancouver | 199.2 | 206.7 | 3.8 |
| Winnipeg | 181.0 | 189.4 | 4.6 |

*Source*: Table 1 Average domestic air fares for 10 major Canadian cities, *The Daily*, Wednesday, January 9, 2013, Statistics Canada

a) Does this table show microdata or aggregate data? How do you know?

b) Is this table a primary or secondary source of data? Justify your answer.

c) Identify the independent and dependent variables.

d) What kind of story do the data in this table tell?

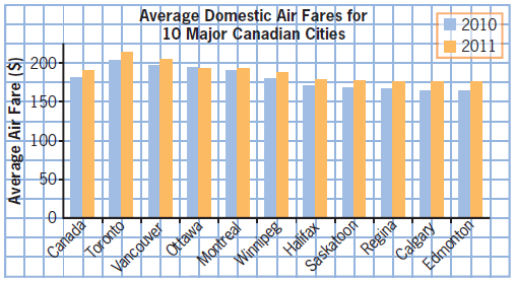
e) Locate this report from *The Daily* archive on the Statistics Canada website. What other types of things are mentioned in the report?

f) What type of sampling was used? Where did you find that information?

a) The data shows average domestic airfare, so this is aggregate data.

b) This is a primary source for Statistics Canada, as they collected it. Is a secondary source for anybody else.

c) Independent - city, Dependent - average airfare.




d) Creating a dual bar chart would help to compare the data. There is an increase for all average fares from 2010 to 2011 except for Ottawa. Most cities had a similar increase except for Montreal which had a smaller increase.

e) There is a link to the survey information.

f) In this case it was a stratified random sample.

## Your Turn

One of the tables that was used to collect the above data was CANSIM Table 401-0004. A condensed version of the table is shown.

a) Can you see any patterns in the data?

b) What kinds of comparisons could be made between 2008 and 2011?

**Average Domestic Fares for Canada and 10 Major Cities**

| Geography | 2008 | 2009 |
|---|---|---|
| Canada | 196.30 | 173.00 |
| Halifax | 197.20 | 170.80 |
| Montréal | 194.30 | 177.80 |
| Ottawa | 205.80 | 189.30 |
| Toronto | 219.80 | 194.40 |
| Winnipeg | 191.50 | 169.90 |
| Saskatoon | 184.00 | 160.50 |
| Regina | x | 160.00 |
| Calgary | 185.20 | 156.40 |
| Edmonton | 180.50 | 154.20 |
| Vancouver | 209.10 | 182.60 |

Symbol legend: x Suppressed to meet the confidentiality requirements of the *Statistics Act*

*Source*: CANSIM Table 401-0004, Average domestic fares for Canada and ten major cities, Statistics Canada, February 18, 2014

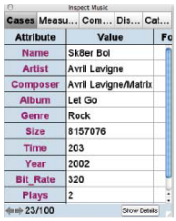
a) It appears that the average domestic fare decreased from 2008 to 2009 by around $30.

b) You could compare the different fares for 2008 to 2011 by city, by province, by cities within a province, or for all of Canada.

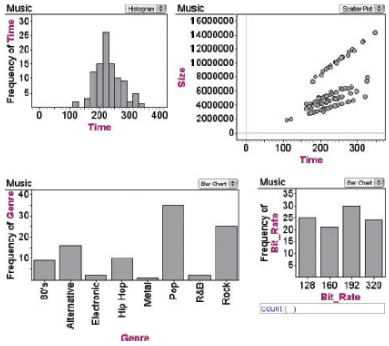
## Example 2

### Analysing a Database

A music library has the attributes shown in the table. Use the table and the graphs to answer the following questions:



| Attribute | Value |
|---|---|
| Name | Sk8er Boi |
| Artist | Avril Levigne |
| Composer | Avril Levigne/Matrix |
| Album | Let Go |
| Genre | Rock |
| Size | 8157076 |
| Time | 203 |
| Year | 2002 |
| Bit_Rate | 320 |
| Plays | 2 |




a) What type of data are these?

b) How many songs are in this library?

c) What is a more appropriate arrangement of the Genre graph?

d) What kind of story or stories do the data in the graphs show?

e) How does bit rate relate to the scatter plot of Size vs. Time?

a) These are microdata. As they are from a personal music library, they are also primary data.

b) There are 100 songs in the library.

c) The convention for categorical data is to arrange the bars from largest to smallest.

d) Using the genre bar chart we can see that the songs are predominantly pop or rock. Also the longer the song, the larger the file size.

e) Looking at the scatter plot, there seems to be different "lines" that represent the different bit rates. The steeper the line, the higher the bit rate.

## Your Turn

Using the data set from the example, determine the following:

a) What information is given about each song?

b) What kind of story does the Time histogram tell you?

c) Are each of the artist genres equally distributed in terms of the bit rate?

d) If you knew how long a song was, could you determine its size? Explain.

a) For each song, you are given its name, composer, album, genre, size of file, length, year, sampling bit rate, number if times played.

b) The time (assuming in seconds) histogram tells us that most common song length was between 210 and 230 seconds.

c) We would need access to the microdata to determine this.

d) We could find an approximate file size using the scatter plot.

## Example 3

### Interpreting a Trend on a Time Graph

One of the gases directly linked to the greenhouse effect and climate change is carbon dioxide ($CO_2$). Keeping track of the amount of carbon dioxide in the atmosphere is an indirect method of measuring the potential for climate change. At the Mauna Loa Observatory in Hawaii, scientists have been collecting atmospheric data since the 1950s. Due to its remote location and minimal influence from vegetation or human activity, this site has become important in the collection of many types of atmospheric data.

a) This data set shows the average monthly carbon dioxide levels (in parts per million) in January of each year for two decades. What type of data are these? Explain.

b) Without graphing, suggest what appears to be happening over time.

c) Create a graph showing these data. What story do the data tell?

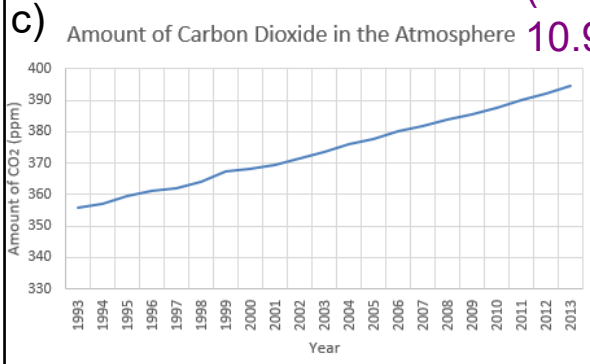
| Year | $CO_2$ Reading (ppm) |
|---|---|
| 1993 | 355.97 |
| 1994 | 356.90 |
| 1995 | 359.57 |
| 1996 | 361.02 |
| 1997 | 362.02 |
| 1998 | 364.16 |
| 1999 | 367.52 |
| 2000 | 368.36 |
| 2001 | 369.61 |
| 2002 | 371.43 |
| 2003 | 373.65 |
| 2004 | 376.03 |
| 2005 | 377.61 |
| 2006 | 380.10 |
| 2007 | 381.80 |
| 2008 | 384.06 |
| 2009 | 385.66 |
| 2010 | 387.51 |
| 2011 | 390.06 |
| 2012 | 392.17 |
| 2013 | 394.66 |

*Source:* National Oceanic and Atmospheric Administration, August 28, 2013

a) Since these are monthly averages, they are aggregate data. Although the researchers collected the readings and therefore used primary data to create the table, for anybody else using the data, it will be secondary data.

b) The level of $CO_2$ appears to be rising at an increasing rate. The total rise from 1993 to 2013 is 38.69 ppm (394.66-355.97). That is an increase of 10.9%.

c)




The graph confirms that the rise is real.

## Your Turn

A student drops a ball from various heights and measures the height of the bounce. The table shows the results.

a) Is this a study or an experiment? Explain.

b) What is being controlled and what is being changed in this activity?

c) What is the relationship between the drop height and the bounce height?

d) Create a graph of the data. Describe the relationship in the graph.

e) If you dropped the ball from a height of 130 cm, could you use this information to predict its bounce height? If so, explain how.

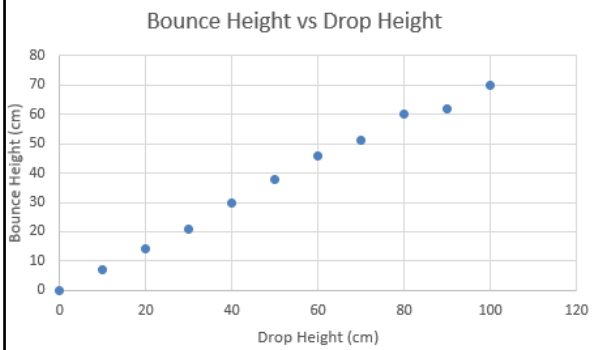
| Drop Height (cm) | Bounce Height (cm) |
|---|---|
| 0 | 0 |
| 10 | 7 |
| 20 | 14 |
| 30 | 21 |
| 40 | 30 |
| 50 | 38 |
| 60 | 46 |
| 70 | 51 |
| 80 | 60 |
| 90 | 62 |
| 100 | 70 |

a) This is an experiment, because the student is controlling the height that the ball is dropped from.

b) The height that the ball is dropped from is being controlled, and the bounce height is recorded for that particular height.

c) The higher the ball is dropped from, the higher the bounce height. It appears to be about 70% of the original height.

d) The relationship appears to be linear.




e) You could use this data to extrapolate to find a predicted value

## Key Concepts

- Primary sources of data are collected directly from the source and are not manipulated or summarized in any way.
- Microdata are the individual pieces of data that make up all of the primary data.
- Secondary sources of data are used by someone who did not collect them. Often these data have been manipulated and summarized. Data found in the media often are secondary data.
- Data that are summarized in some way are called aggregate data.
- Large sources of data are available for analysis on the Internet.
- Sources of data also are hidden in digital items like songs and photos.

**R1.** How are websites for the NHL, MLB, NBA, and NFL like databases?

These websites are like databases because they provide an organized store of records containing information on players, teams, and games.

**R2.** Which is more visible to the public, primary or secondary data? Why do you think this is?

Secondary data is more visible to the public because we typically are seeing other people's primary data.