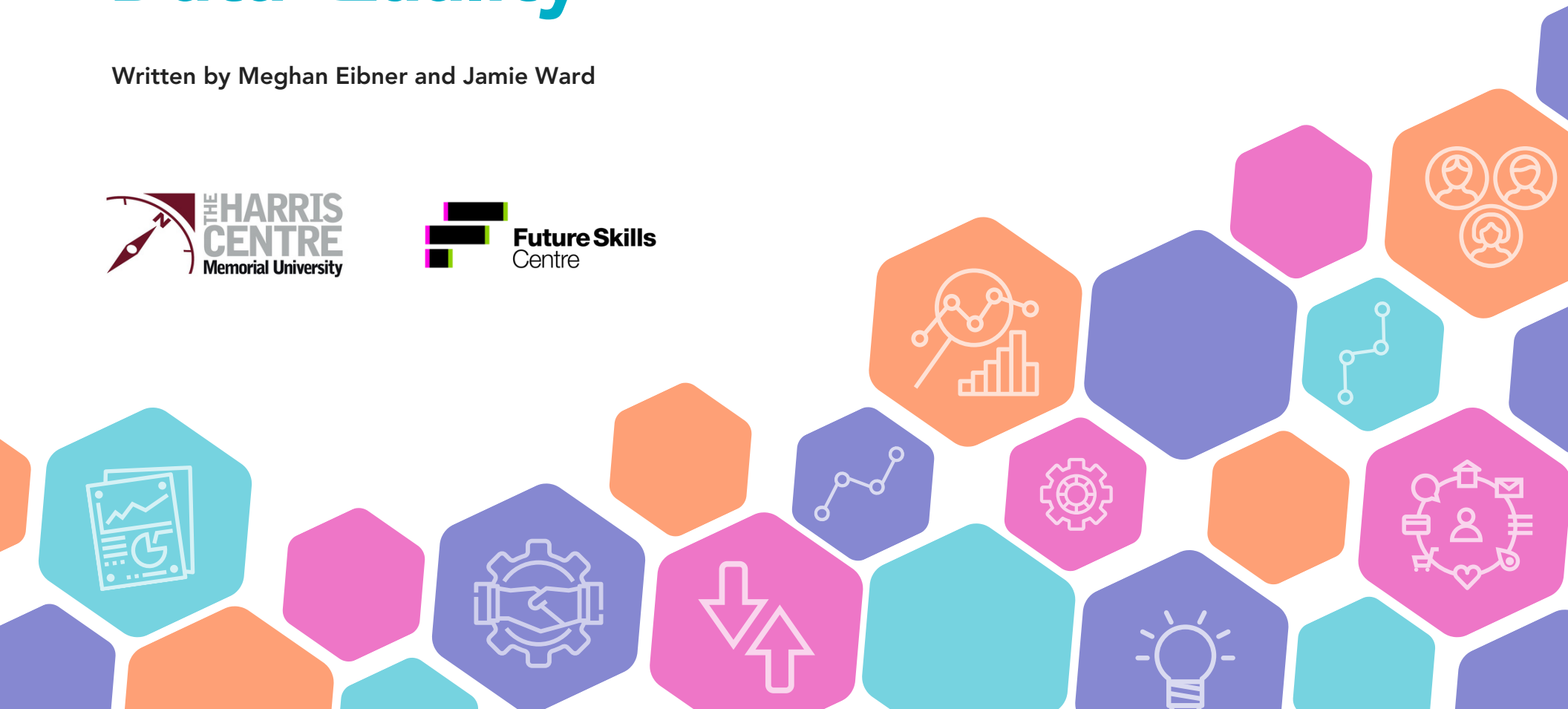
Regional Analytics Laboratory **PRESENTS**

December 2021

DATA TO DECISIONS:
Data Quality

Written by Meghan Eibner and Jamie Ward



Foreward

Data to Decisions is an initiative of the Regional Analytics Laboratory (RAnLab) and is sponsored in part by Future Skills Centre.

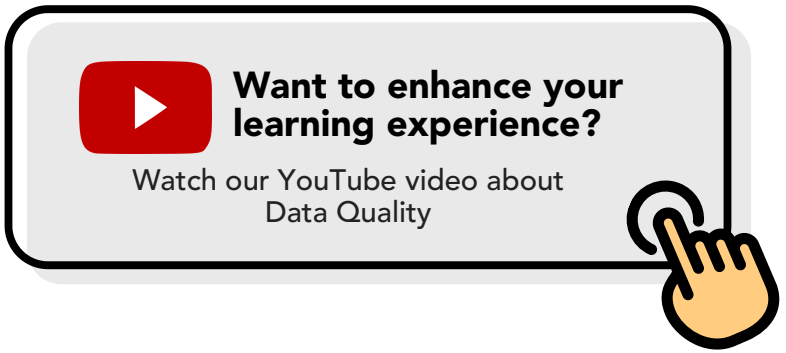
RAnLab is the data and analytics unit of Memorial University's Leslie Harris Centre of Regional Policy and Development. RAnLab analyzes data and geography to provide insight into, and projection modeling for, things like demographics, labour supply, service demands, commodity prices, and other socio-economic indicators. Some examples of their work include producing long-term community and regional population projections, assessing local housing demand, and providing detailed local data analysis and modelling to municipalities and regions–providing critical information for evidence-based decision-making.

The purpose of Data to Decisions is to help Canadians from varying backgrounds learn how to apply data to their own projects. Data to Decisions uses plain language, examples, and video presentations to enhance the learning experience.

If you have any questions, please contact Meghan Eibner (meibner@mun.ca) or Jamie Ward (jward@mun.ca).

Throughout this chapter, you will see clickable buttons that link to webpages with additional information.

The button on the right will lead you to the Data to Decisions presentation on Data Quality.



Contents

Data Quality

When gathering and using data it is important to be aware of various issues that may impact the quality, and therefore the interpretation, of the data.

Some of the most common data quality issues are related to data availability, consistency, and accuracy.

It's important to conduct quality control checks on any gathered data and document any data issues you identify (and any actions taken to address the issues). This helps ensure analysis and interpretation is as transparent as possible.

When examining a data set to identify potential issues, things to keep an eye out for include:

- big shifts in the data;
- outlier data points;
- breaks or gaps in the data.

Examples of these data quality issues, using hypothetical data, are shown on the following pages.

Data Availability

The degree to which data is readily available to users when and where they require it.

Data Consistency

The usability of applicable data. The process of keeping data uniform as it moves across, and between, various applications.

Data Accuracy

The degree to which data is error-free and can be used as a reliable source of information.

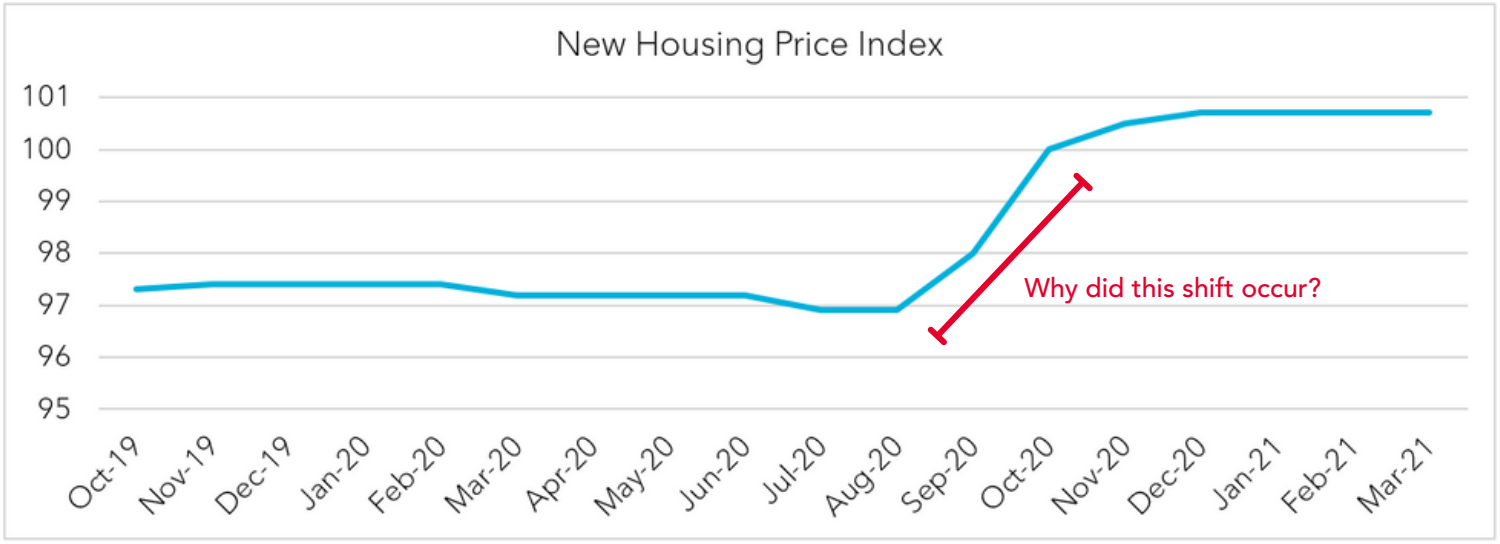
Shift in Data:

The graph below shows a time series data set with an observable shift in data at the Aug-20 mark. Prior to this date, the data had been relatively stable for quite some time. After a steep increase between Aug-20 and Nov-20, the data levels out again at a much higher level.

Such a drastic shift in data must be investigated. We need to ask questions to try determine the cause of the shift, such as:

- Did something happen in the housing market or economy to justify a legitimate increase in data?
- Was there a change in the definition of New Housing Price Index around Aug-20 to result in a shift in the data set?

Data Shift
A change in data distribution (for example, a change in classification or geography).



Outlier Data:

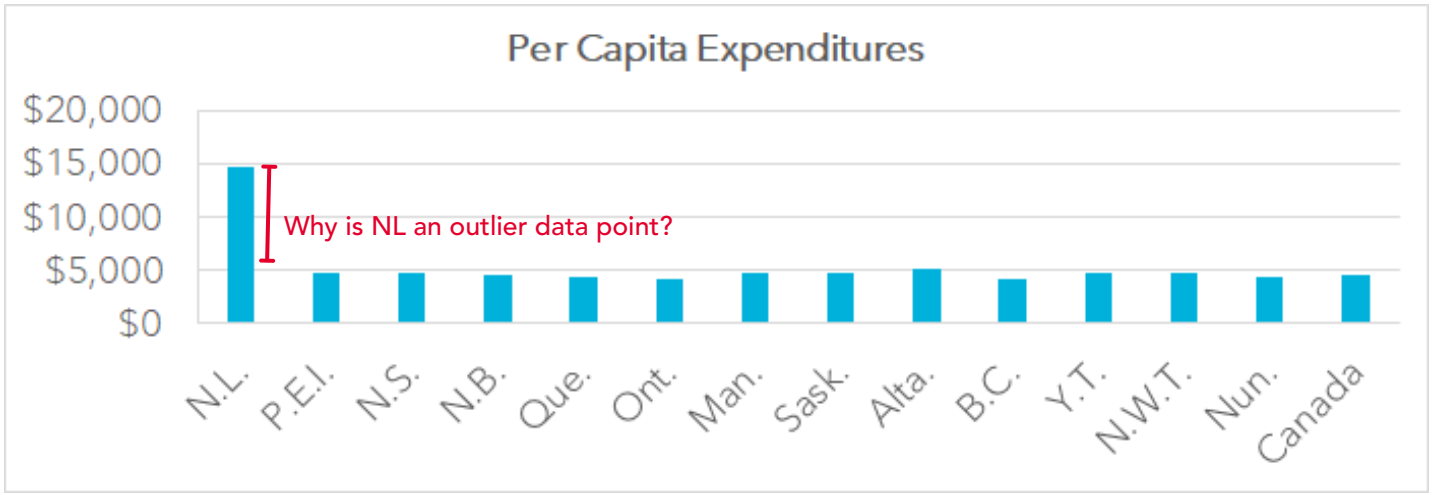
The graph below is an example of a data set with an outlier data point. All data points, except for NL, follow a similar pattern.

The outlier data point should be researched to see whether it is the result of a mistake or accurately represents what it is measuring.

For example, we must ask ourselves:

- Did something happen in NL at the time that resulted in significantly higher expenditures than the rest of the country?
- Was the NL data incorrectly entered into the database resulting in a data point much different than all the rest?

Outlier Data Point
A data point that differs significantly from other observations.



Data Gaps:

The table below is an example of a data set with a gap in the data; the data for the year 2006 is missing. The reason for this should be investigated:

- Is it simply a mistake – was the data for that year accidentally omitted?
- Is the data for that year unavailable? If so, why?
- Was the survey not conducted for that year?
- Did confidentiality practices require that year of data be suppressed?

Data Gap
Data for particular elements or data sets are knowingly or unknowingly missing.

Year	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Population	366	376	388	400	412		425	429	438	444	454

Why is there a data gap?

If questions about the data can't be answered, using that data should be done with extreme caution.

- Data published without documentation or with unanswered questions may say something about the credibility/reliability of the source.
- If you don't understand your input data then it's unlikely that you are going to understand your results.

Data Availability

Data availability can vary over time periods, geography, and topic. For example, a data set from a survey introduced in the 1980s will not exist prior to that time period because that data had not been collected yet. The format in which data is available can also change over time. Most Statistics Canada data in recent history is available online for download in a useable format. However, some older historical data (historical census data, for example, pre-1970s) may only be available in scanned PDF form which can make data selection and use challenging.

If the geographic or organizational level of a data set is small enough that it's possible to identify individual survey responses, the Statistics Act requires that this data be suppressed. This is to protect individual privacy and ensure data collected is private, secure, and confidential.

This results in occasional gaps existing where data is unavailable due to suppression. For example, a particular data set may be available for all municipalities in an area except for one, due to confidentiality requirements.

Alternatively, years of data may be missing from a time series data set. In this case, if the data gap is relatively small, it may be possible to fill in by making an assumption about what the data within the gap looks like.

Statistics Canada will note when data has been suppressed. You may see a superscript number next to a data point which corresponds to a note below the data set. See page 14.

Interpolation

The process of filling in a data gap based on an assumption is called interpolation. The following examples show three different ways to fill in a missing year of data within a data set by using three different assumptions. Note that this is a simple example with a single year of missing data. The more gaps there are within a data set, the more complicated the process of attempting to fill them in will be.

Interpolation
The process of filling in a data gap using reasonable, justifiable, and well-documented assumptions

Year	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Population	366	376	388	400	412		425	429	438	444	454

How can we fill the gap?

Assumption 1: Two Point Average

The data for the year 2007 is missing. One way of filling in the missing year is to simply take the average between the two data points on either side of the missing year. The population for 2006 was 412, and the population for 2008 was 425, so one could assume that the population for 2007 was 419.

$$\frac{412 + 425}{2} = \frac{837}{2} = 419$$

Assumption 2: Average Annual Percentage Change

The second way of filling in the missing year is to determine the average annual percentage change in population and use that to calculate the data for 2007.

Year	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Population	366	376	388	400	412		425	429	438	444	454
Annual % Change		2.7%	3.2%	3.1%	3.0%			1.7%	2.1%	1.4%	2.3%

Calculate the annual percentage change for every year where data is available (see above table), then take the average of that calculated data set. In this case, the average annual percentage change is 2.3 percent.

$$average(2.7\%, 3.2\%, 3.1\%, 3.0\%, 1.7\%, 2.1\%, 1.4\%, 2.3\%) = 2.3\%$$

It can then be assumed that the 2006 population increased by 2.3 percent to result in a 2007 population of 422.

$$412 * (1 + 2.3\%) = 412 * (1 + 0.023) = 412 * 1.023 = 422$$

Assumption 3: Straight Line Percentage Change

This approach is similar to the previous example except, instead of calculating the percentage change for every year, a single percentage change over the entire time period is assumed. To start, calculate the percentage change between the first year of data and the last year of data in the data set. The percentage change between 366 (in 2002) and 454 (in 2012) is 24.0 percent.

$$\frac{454 - 366}{366} = \frac{88}{366} = 0.24 \quad or \quad 24\%$$

Now, divide that percentage by the number of years in the data set (in this example, 11 years) to get the assumed straight line percentage change. The result is 2.2 percent.

$$\frac{0.24}{11} = 0.022 \quad or \quad 2.2\%$$

This 2.2 percent change can now be to applied to the 2006 population to calculate the missing 2007 data point. This results in an assumed 2007 population of 421.

$$412 * (1 + 2.3\%) = 412 * (1 + 0.022) = 412 * 1.022 = 421$$

The method used to fill in a data gap depends on the characteristics of the data within the data set:

- The straight line assumption works well with consistent data that follows a predictable pattern, making it reasonable to assume that the annual percentage change between the first and last year of data is relatively stable.
- The average annual percentage approach may better reflect changes within a data set that is less consistent but still relatively predictable (ie, no outliers or big shifts).
- The two point average approach may work best with more volatile data in order to capture year-to-year changes.

Data Accuracy

Data accuracy can become an issue when entering, copying, converting, or transferring data. When errors occur in data it may no longer be a reliable source of information so conducting quality control checks on any data that has been manipulated is critical. An example of human error resulting in data inaccuracy might be a simple typo when transferring scanned PDF data into a digitized format (e.g. excel).

Data Accuracy
The degree to which the data correctly describes what it was designed to measure

TIP Outlier data points can be an indicator of data accuracy issues.

Data Consistency

Consistency issues can occur when collecting data from different sources, jurisdictions, and time periods. It is important to look at the definition and description of the data sets being collected in order to determine if they are comparable and can be used as a valid, consistent data set. Even though a data set may have the same name (e.g. "Net Debt"), it may be defined differently across geographic areas, jurisdictions, and sources.

Data Consistency
The usability of related data. The process of keeping data uniform as it moves across and between various applications.

TIP Big shifts in data can be an indicator of data consistency issues.

A data set collected from a single source can also have definition changes over time. For example, population data for Town X may be available for the past several decades, but, during this time period, the town amalgamated with surrounding communities: Town Y and Town Z. This would result in both a shift in the data, as the populations numbers are now combined into one, as well as a change in the definition of the data set itself.

Data Appropriateness

Once you've done your quality control check on your data and you're ready to use it, you also need to make sure you're using it appropriately. One of the most common issues that arises from inappropriate use of data is called the Ecological Fallacy.

There are several spatial fallacies that can occur through the inappropriate use of data, but the Ecological Fallacy tends to be the most common.

When data at a specific geographic level isn't available, it can be tempting to apply data from other geographic areas/levels. for example, applying provincial or national data at a local level. For the majority of localities, the picture painted by national data is quite different than the local reality. Using data in this way will almost certainly lead to a misleading interpretation. This is called Ecological Fallacy.

Interested in learning more about fallacies? Try researching the following:

- Known Length
- Locational Fallacy
- Atomic Fallacy
- Modifiable Areal Unit Problem

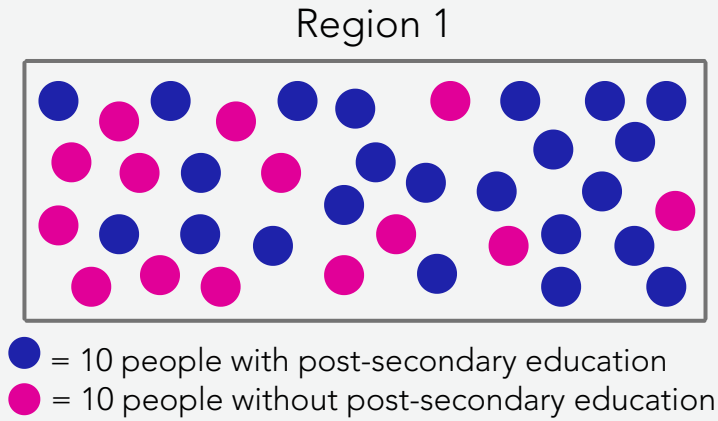
We'll cover ecological fallacy in the next section.

Ecological Fallacy

The Ecological Fallacy is an error that occurs when conclusions are drawn about individuals or small areas based on data relating to a larger group or geographic area. Assuming that what is true for a group is also true for individual members of the group can result in misleading or inaccurate data analysis and interpretation. This is why it is critical to use data at the appropriate geographic level.

For instance:

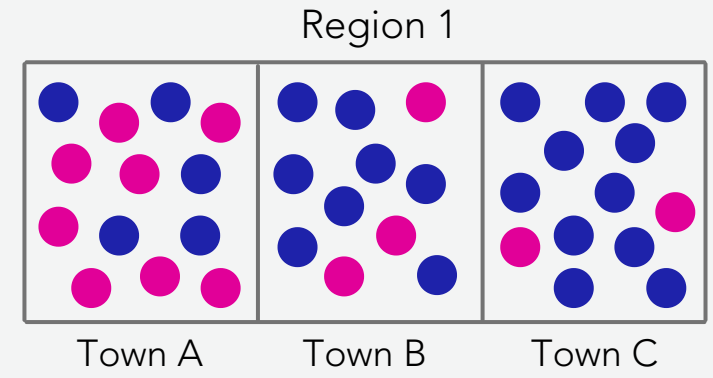
- If a province has a declining population, it is incorrect to assume that every community within that province is experiencing population decline.
- If a community is composed of a large highly skilled population, it is incorrect to assume that every person within that community is highly skilled.



What we know:

- Statistics show that the majority of people living in Region 1 have attended post-secondary education
- Region 1 is made up of three towns

Can we assume that the majority of people living within each town have attended post-secondary education?
No, this reasoning is an ecological fallacy.



By looking at data at the correct geographic level we can see that the majority of people living in Town A have not attended post-secondary education.

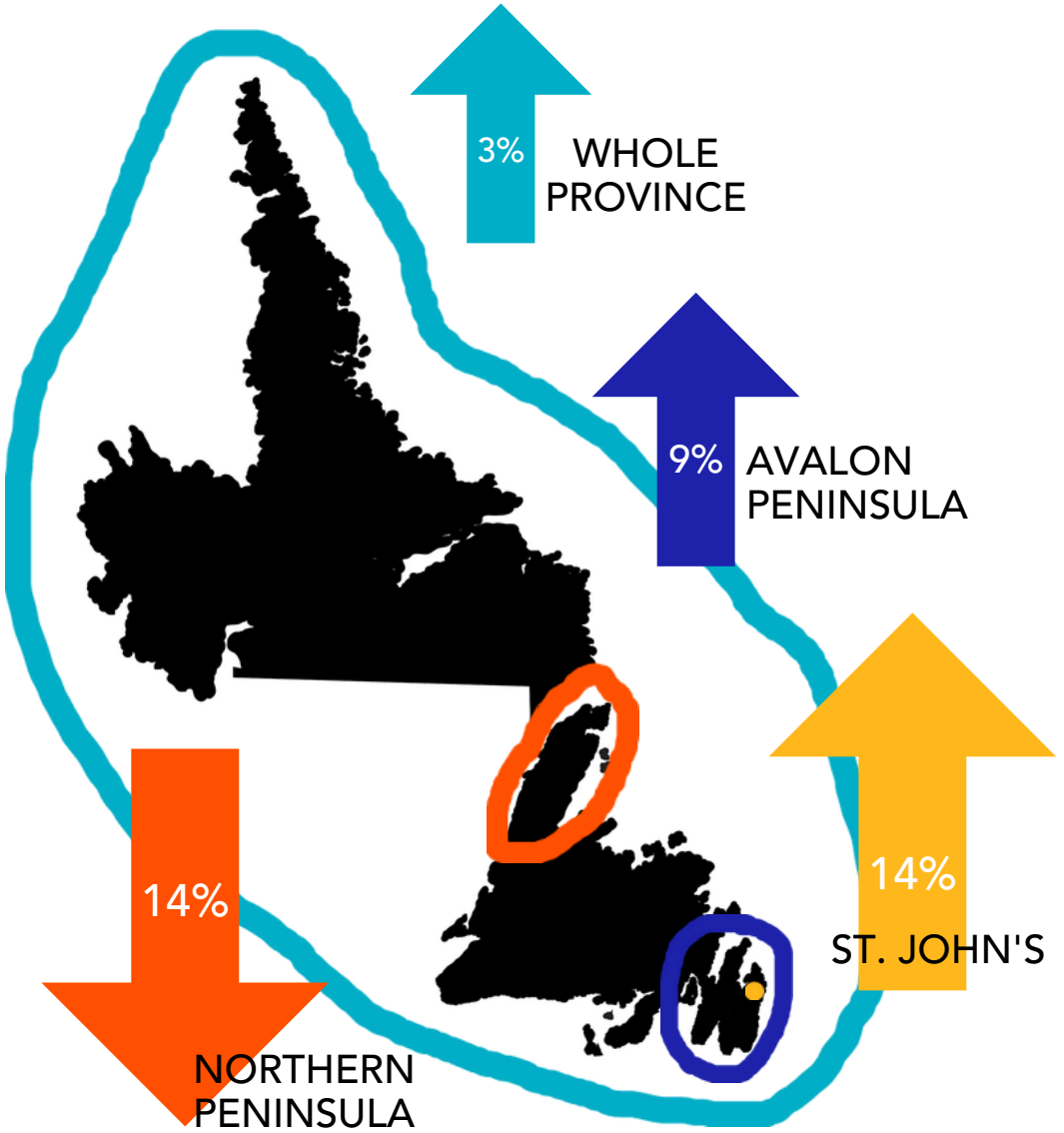
An example of ecological fallacy:

From 2006 to 2016, the province of Newfoundland and Labrador experienced slow growth, with the total population increasing only 3% over that ten-year time period.

The Avalon Peninsula, however, experienced 9% population growth over that same time frame and the St. John's population grew 14%.

Meanwhile, the population on the Northern Peninsula declined by 14%.

This example demonstrates how provincial data may not accurately reflect the more nuanced experiences of smaller geographic areas within the province.



Extra Resources

Statistics Canada has developed their Six Dimensions of Data Quality which goes into more detail on the factors they account for when determining data quality.

Click here for more detail on the Six Dimensions of Data Quality

Based on these criteria, Statistics Canada will flag data sets or points based on their "fitness of use." For example, a data point may be flagged as 'use with caution' or 'unreliable'. These notes should be taken into consideration, and documented, when collecting and using flagged data.

Click here for Statistic Canada's Data quality toolkit

Statistics Canada has also developed a Data quality toolkit which aims to raise awareness about data quality assurance practices.

Chapter Summary

Data to Decisions:
Data Quality

- Quality data analysis relies on the use of quality data. Therefore, it is important to ensure data is available, consistent, and accurate.
- Issues with data quality to look out for include big shifts in data, outlier data points, and data gaps.
- In some cases, these issues can be resolved through interpolation – using documented, defensible assumptions to fill in missing data.
- Some methods of interpolation include two point average, average annual percentage change, and straightline percentage change.
- Ecological Fallacy which is an error that occurs when assuming that what is true for a group is true for an individual within that group and can result in misleading analysis and data interpretation.