DATA ANALYTICS LAB: FRAMEWORKS FOR DOCUMENTING,

ASSESSING AND ANALYZING DATASETS

April, 2017

Table of Contents

List of Tables

List of Appendix Figures

Executive Summary

Like other departments within the Government of Canada, the Strategic Research and Statistics Department of Indigenous and Northern Affairs Canada lacks a strategic approach for documenting, finding and using data that is required to meet the increasing demand for quality and timely data. Furthermore, issues resulting from siloed data systems and inadequate documentation make it difficult to use datasets to their full potential which affects research and the use of statistics.

This report introduces a dataset tracking system and assessment methodology framework, discusses strategies for searching for data, identifies sources of data and presents the main results of a statistical analysis run using selected datasets. It also examines common problems with datasets available to the department, such as those in the Open Government Data Portal.

The report concludes with a discussion on the adequacy of the templates and systems for data documenting, and describes challenges that still remain.

1.0 Introduction

As part of the Government of Canada's results and delivery agenda, departments face an increasing demand for timely and quality data, and must develop processes that support transparency, usability, as well as replicability in finding and using data. However, like many other departments within the Government of Canada, the Strategic Research and Statistics Directorate (SRSD) branch of Indigenous and Northern Affairs Canada (INAC) lacks the strategic approach for documenting and searching for data that is required to meet the demand. Furthermore, there are many problems with data gaps and inadequate documentation that hamper the linking and sharing of data, which prevents the datasets from being used to their full analytical potential (Indigenous and Northern Affairs, 2016). This leads to the inappropriate use of statistics and ultimately affects decision making.

The purpose of this report is to introduce the implementation of an analytics lab within INAC's Strategic Research and Statistics department that aims to address the challenges the department faces with documenting, finding and using data.

This report considers systems for data documenting, discusses strategies for searching for data, identifies sources of data and potential datasets for the analysis, evaluates datasets using the current assessment methodology, and

presents the main findings of a regression analysis run using final selected datasets.

2.0 Analysis

2.1 Current Data Challenges

As technology advances and the Government of Canada's demand for timely and quality data increases, the ability to document, find and use data in ways that support usability, transparency and replicability of data is of utmost importance for INAC's Strategic Research and Statistics Directorate. However, systems that support this have yet to be implemented, and issues with inadequate metadata, data gaps and siloed data systems result in inconsistencies in datasets across departments. This makes it difficult to efficiently link datasets since common unique IDs do not exist, which in turn affects the accuracy and quality of analyses that are performed using those datasets. By utilizing a framework for dataset assessment methodology, this analytics lab guides the department in what types of analysis selected datasets can support. Also, the discussion on sources of data, such as the Open Government Data portal, gives the department more of an idea of what kinds of data are available.

2.2 Documenting Data: Project Dashboard

A large portion of this data analytics lab was devoted to searching for datasets to determine what kinds of analysis the data can support, and to finding data that can be analyzed in a systematic way. To keep track of all potential datasets, a Microsoft Excel spreadsheet was created. This spreadsheet records the dataset name, source (organization or website), date retrieved, and main variables of interest. Using a tracking system such as an Excel spreadsheet to document potential datasets is necessary as it allows for replicability of data and allows users to search for key terms using the filter feature.

While regular Microsoft Excel was used for the initial implementation of the tracking system, project management software that would be good for future uses of the system is Microsoft Project. Microsoft Project would be effective because it combines project management features such as customizable project timelines with functions that allow reports to be created and shared (Microsoft, 2017).

2.3 Data Acquisition

In addition to the logging component of the analytics lab, an important element involves the acquisition of data, specifically the identification of data sources, strategies for searching for data within those sources, a template for recording the datasets details, as well as a dataset maintenance strategy.

2.3.1 Sources of Data

To determine what kinds of data are available to the department, different sources were explored. The main source of data was the Open Government Data Portal, on the Government of Canada's website (http://open.canada.ca/data/en/dataset). The portal contains both geospatial and non-spatial data on many different topics from various governmental organizations including Statistics Canada and Health Canada, and indicates what format the datasets are in (such as .csv, which is the most commonly used file type for statistical analysis).

From the Open Data portal, the first datasets identified with analysis in mind were the Canadian Disaster Database from Public Safety Canada, and the Community Well-Being Index (CWB) from INAC. Both are available in .csv format. While the search for data sources was open-ended, we knew that we wanted to use the CWB, which uses census subdivision code as the unique primary key. The CWB is "a means of examining the well-being of individual Canadian communities" and assigns scores to each community by combining indicators for income, housing, labour force activity and education (Indigenous and Northern Affairs Canada, 2016). Hence, datasets with census subdivision as the unit of analysis are the most compatible with the CWB, and were therefore the focus of the search for data going forward.

Finding data from primary sources is preferred since it is more reliable than data from secondary sources; however, if a secondary data source has sufficient documentation indicating where the data comes from and how it is compiled, it can be used. An example of data from a secondary source that was considered for use in this analytics lab is a list of all drinking water advisories currently in place in Canada. This data is from www.watertoday.ca and is gathered daily from official provincial and municipal water advisory sources (Water Today, 2017). While this data includes both Indigenous and non-Indigenous communities in Canada, it does not have a variable that clearly indicates whether or not the community is a First Nation, so there is no effective way to filter through the data. Also, although the main identifier in the dataset is the community name, it is not unique (like a census subdivision code is). Thus, if this dataset were to be merged with the CWB, census subdivision codes would have to be added manually. This is quite inefficient and would take a long time since the dataset contains 950 rows.

This problem was solved by retrieving lists of water advisories in First Nation communities from the Health Canada and British Columbia First Nations Health Authority sites. These datasets contain the same variable names, so it was simple to combine them to create a dataset of all First Nation communities in Canada with water advisories.

In addition to the CWB, INAC is the source for ‘List of Indian band areas and the census subdivisions they include’, which is a table that lists the province/territory, band name, census subdivision name, and any common or alternative names of the First Nation. This was used along with the CWB to assign census subdivision codes to the communities in the First Nations water advisory dataset.

2.3.2 Strategies for Searching for Data within Sources

On the Open Data portal, there is a search bar as well as several filters that allow the user to search for datasets more effectively. Users can choose from many different organizations, keywords, subjects, and collection types (geospatial, non-spatial or open maps). The number of datasets that match each search criteria are listed beside the filters.

Within the datasets themselves, sorting and filter functions can be used since the datasets are available as Excel .csv files. This helps the analyst quickly locate entities with certain attribute values. One useful strategy to use in Excel to search within datasets is to use the ‘contains’ option in the text filters.

Using these strategies when searching for and within data makes the process more efficient and precise.

2.3.3 Final Selection of Datasets for Regression Analysis

The datasets that were considered for use in the analysis are the Canadian Disaster Database (CDD), CWB, and Water Advisories in First Nations Communities. The initial idea for the regression analysis was to see if water advisories and disasters such as floods or fires are associated with differences in Community Well-Being Index scores in First Nation communities. Each of these datasets has the community/city as the unit of analysis, so there is the possibility of merging the files by census subdivision code.

Although the CDD would be interesting to use, the locations of many of the incidents are inexact and there are no unique keys, which makes it difficult to accurately and efficiently compare the CDD with the CWB and water advisory dataset. For these reasons, and to ensure that the data used are as accurate as possible, the datasets that were selected for the regression analysis are INAC's CWB and the list of water advisories in First Nation communities from Health Canada and British Columbia's First Nations Health Authority.

In order to merge the CWB and water advisory data, census subdivision codes were added to the water advisory dataset. This was accomplished by using INAC's 'List of Indian band areas and the census subdivisions they include' to find the official census subdivision names for the First Nation communities with water advisories. Once the census subdivision names were found, the census

subdivision codes were assigned easily. The two files were merged using SAS (code available in the Appendix).

In the next section, a framework for dataset assessment methodology is used to provide more detail on the two selected datasets and their variables.

2.3.4 Dataset Template: Acquisition Detail and Metadata

Without proper documentation, datasets are difficult to understand and use. For example, many datasets in the Open Government Data Portal contain poorly named variables, but do not have data dictionaries or glossaries. The assessment methodology used in this section comes from INAC's Data Assessment Methodology (INAC, 2016) and provides a clear framework for documenting dataset details including coverage and variable types. The assessments of the CWB and Water Advisory datasets using this framework are displayed in the following tables.

Table 1: Dataset Assessment Framework for the Community Well-Being Index

Dataset and Variable Details	Description
Variable Names	The variables in this dataset are: CSD Code, CSD Name, 2A Pop 11, CSD Type, FN, Inuit, Incompletely Enumerated 2011, CSD Code Change 2006-2011, Global Non-response Rate 2011, Income, Education, Housing, LFA, CWB 2011
Unique Identifier	The unique identifier is the census subdivision (CSD) code
Unit of Analysis	The unit of analysis is census subdivision
Variable Definition	The variables that were kept for the regression analysis are: CSD Code, FN, Inuit, Global Non-response, Income, Education, Housing, LFA and CWB and CSD Name. Their definitions are available in the spreadsheet "CWB FINAL SMD WORKING FILE WITH 1981 AND 1991 TO 2011 WORKSHEETS - NO 2011 GNR SUPPS"
Variable Type	CSD Code, FN, Inuit, Global Non-response, Income, Education, Housing, LFA and CWB are numeric while CSD Name is a string
Universe	The universe is all census subdivisions in Canada
Coverage	Some communities chose not to participate in parts of the census, so some component scores for those communities are missing. Also, communities with a population of less than 65 or a global non-response rate of over 50% are excluded
Periodicity	Every five years
Time Lag	The average period of time between data collection and availability of the dataset is three years

Table 2: Assessment Framework for First Nation Water Advisories Dataset

Dataset and Variable Details	Description
Variable Names	The variables in this dataset are: First Nation, Community, Water System Name, Advisory Type, Date Set, Date Revoked, Population
Unique Identifier	Although the community name is the main identifier, none of the variables are unique identifiers. The census subdivision code became the unique identifier after it was added as a variable
Unit of Analysis	The unit of analysis in this dataset is census subdivision
Variable Definitions	First Nation is the band name; Community is the census subdivision name; Advisory Type indicates whether the advisory is a boil water or do not consume advisory; Population is the number of people affected by the advisory
Variable Type	First Nation, Community, Water System Name, and Advisory Type are strings while Date Set, Date Revoked and Population are numeric
Universe	The universe is all First Nation communities
Coverage	The coverage is all First Nation communities south of the 60^{th} parallel
Periodicity	Monthly
Latest Availability	Date of most recent data entry: Feb. 24, 2017
Earliest Availability	Date of earliest data entry: Feb. 1, 1995

2.3.5 Maintenance Strategy

Keeping datasets updated is important as it ensures the data remain relevant, which is one of the dimensions of quality for administrative data (Statistics Canada, 2002). The current approach to maintaining datasets is to check back to the source for updates. This is fine for smaller datasets such as the Health Canada/British Columbia First Nation Health Authority water advisory lists since they are updated monthly, but for large datasets that are updated more frequently this approach is inefficient and time consuming.

While an extensive maintenance strategy does not yet exist within SRSD, using the data assessment methodology in the previous section tells researchers how often datasets are updated and thus how frequently they should check back. Furthermore, the Open Data portal features a "maintenance and update frequency" filter that indicates whether the datasets are updated annually, monthly, irregularly, as needed or unknown. Developing a concrete strategy for maintaining datasets will ensure the data used by the department remains relevant and reliable.

2.4 Data Use: Template for Documenting Analyses

The purpose of this portion of the analytics lab is to present the main findings of the regression analysis performed using the final selected datasets and to describe

the methodology used. The goal of the Community Well-Being Index and First Nation Water Advisories study was to determine if there is an association between differences in CWB scores and the presence of water advisories in First Nation communities. To do this analysis, the software program SAS was used; simple and multiple linear regression models were fit, boxplots for each CWB component score were created and statistical significance tests were run. See the appendix for the full details of the analysis, including coding and output.

The boxplots (Appendix B) show that each mean CWB component score is higher for communities with no advisories. The regression analysis (Appendix C) reveals if the presence of a water advisory actually has a significant impact on communities' scores. For the simple linear regression, each CWB component score was fit against Advisory (0 if no advisory and 1 if the community has an advisory). The main result for the simple linear regression analyses was that the presence of a water advisory is significantly associated with a decrease in all CWB component scores (housing, income, education, labour force activity, and overall CWB score) at the five percent level. For example, the expected difference in income score between a community without an advisory and one with one is approximately seven points. That is, the expected income score drops seven points when an Indigenous community has a water advisory.

I also wanted to see if location is a potential confounder for the association between component scores and the presence of an advisory. To do this, five multiple linear regression models were fit; each score was fit against Advisory and ProvCode (the indicator variable for province). In each case, location did not distort the association between scores and the presence of an advisory. From the SAS output, there is no instance with both the p-value for Advisory > 0.05 and the p-value for ProvCode < 0.05. Therefore, province is not a confounder so there is no need to include it in the regression model.

This analysis provides quantitative evidence of a significant association between the presence of water advisories and poorer socio-economic conditions in First Nation communities. The results of the analysis suggest that the Government of Canada's commitment to end all long-term drinking water advisories on reserves funded by INAC within five years will not only improve water quality and infrastructure; improvements in housing, income, education, employment and overall CWB scores may also be seen.

3.0 Conclusions

The systems used in this data analytics lab to document, find, describe and analyze datasets provide a good foundation for the department to build upon. The Excel tracking system is straightforward but can be improved by incorporating it into a project management model. The dataset assessment framework is comprehensive and helps the researcher understand the dataset's contents and purpose. The discussion on the Open Data Portal gives the department more of an idea of what types of data are available, as well as the main issues that datasets within the portal have.

While the systems described in this analytics lab successfully address some of the challenges the department faces, issues such as the inability to accurately link datasets still remain. Better collaboration between organizations to create datasets with common identifiers would allow for more efficient research and analysis. In addition, further development of a dataset maintenance strategy will allow the data used by the department to remain reliable and relevant.

References

First Nations Health Authority. (2017). Drinking Water Advisories. Retrieved February 27, 2017, from http://www.fnha.ca/what-we-do/environmental-health/drinking-water-advisories

Government of Canada (2015, November 13). List of Indian band areas and the census subdivisions they include. Retrieved January 24, 2017, from http://www12.statcan.gc.ca/nhs-enm/2011/dp-pd/aprof/help-aide/a-tab.cfm?Lang=E&wbdisable=true

Government of Canada, Public Safety Canada. (2013, September 12). Canadian Disaster Database. Retrieved January 19, 2017, from http://cdd.publicsafety.gc.ca/srchpg-eng.aspx?dynamic=false

Health Canada (2017, February 23). Drinking water advisories: First Nations south of 60. Retrieved April 11, 2017, from https://www.canada.ca/en/health-canada/topics/health-environment/water-quality-health/drinking-water/advisories-first-nations-south-60.html

Indigenous and Northern Affairs. (2016). Data Mobilization for Performance Measurement.

Indigenous and Northern Affairs. (2016). Database Assessment Methodology.

Indigenous and Northern Affairs. (2016, February 08). The Community Well-Being (CWB) Index. Retrieved April 11, 2017, from https://www.aadncaandc.gc.ca/eng/1100100016579/1100100016580

Microsoft. (2017). Project Management. Retrieved April 17, 2017, from https://products.office.com/en-ca/project/project-management

Open Government Portal. (2017). Retrieved January 19, 2017, from http://open.canada.ca/data/en/dataset

Statistics Canada. (2002), "Statistics Canada's Quality Assurance Framework", Catalogue no. 12-585-XIE, Ottawa, Canada: Statistics Canada.

Water Today. (2017). Water Advisory Information. Retrieved February 27, 2017, from http://www.watertoday.ca/bwa.asp

Appendix A: SAS Code – Merging and Creation of Final Dataset

```
DATA FNWater;
      INFILE 'FNWater.csv' DLM=',' FIRSTOBS=2 TERMSTR=cr;
      INPUT Province$ FirstNation$ Community$ SystemName$
      TypeAdvisory$ DateSet$ DateRevoked$ Population$
      CSDCode NumAdvisories;
RUN; /* Import First Nations Water Advisory Dataset */


DATA CWB2011;
      INFILE 'CWB2011.csv' DLM=',' FIRSTOBS=2 TERMSTR=cr;
      INPUT Col$ CSD FN Inuit GlobalNonResp
      Income Education Housing LFA CWB CSDName2011$;
RUN; /* Import CWB */


PROC SQL;
      CREATE TABLE all AS
      SELECT *
      FROM CWB2011 LEFT OUTER JOIN FNWater
      ON CWB2011.CSD=FNWater.CSDCode
      ;
QUIT; /* Merge CWB and Water Advisories */


PROC SQL;
      CREATE TABLE Final AS
      SELECT CSD, CSDName2011, FN, Inuit, GlobalNonResp,
      Income, Education, Housing, LFA, CWB, TypeAdvisory,
      DateSet, NumAdvisories, Province
      FROM all
      WHERE FN=1 OR Inuit=1;
QUIT; /* Prepare Final Dataset */


DATA FinalSet;
      INFILE 'FNAdvisoriesCWB.csv' DLM=',' FIRSTOBS=2
      TERMSTR=cr;
      INPUT CSD$ CSDName2011$ FN Inuit GlobalNonResp Income
      Education Housing LFA CWB Advisory DateSet$
      NumAdvisories Province$ ProvCode;
RUN; /* Import Final Dataset */
```

Appendix B: SAS Code and Output for Boxplots

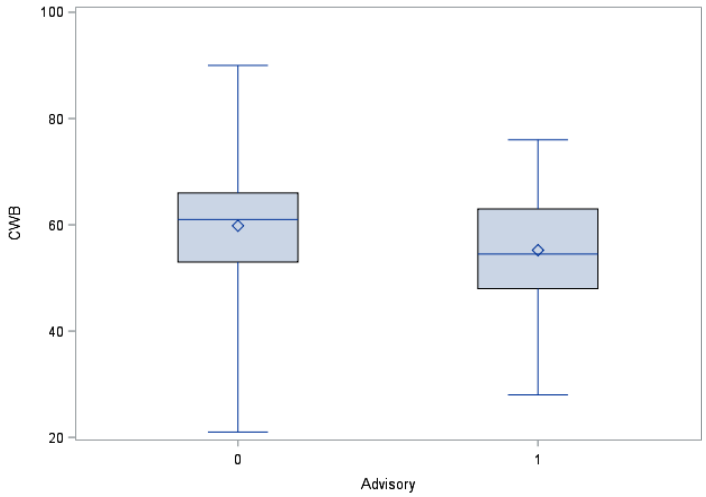
```
/* plot boxplots of CWB component scores by advisory */


PROC SORT DATA=FinalSet;
      BY Advisory;


/* CWB */


PROC BOXPLOT DATA=FinalSet;
      PLOT CWB*Advisory;
RUN;
```

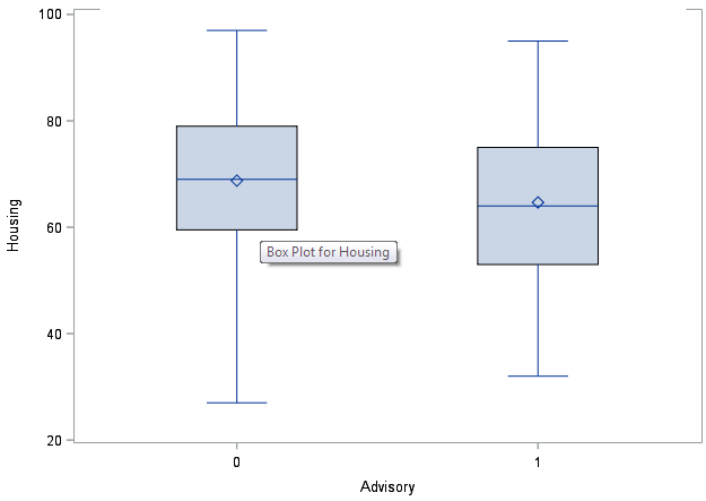
Figure B1: Boxplot of CWB Scores by Advisory




```
/* Housing */

PROC BOXPLOT DATA=FinalSet;
      PLOT Housing*Advisory;
RUN;
```

Figure B2: Boxplot of Housing Scores by Advisory



```
/* Income */


PROC BOXPLOT DATA=FinalSet;
      PLOT Income*Advisory;
RUN;
```

Figure B3: Boxplot of Income Scores by Advisory

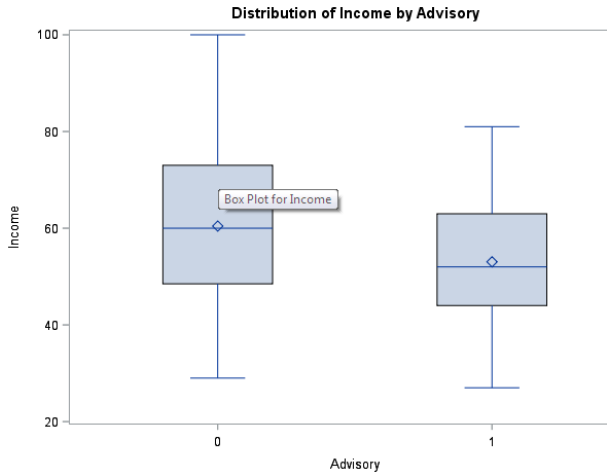


Figure B4: Boxplot of LFA Scores by Advisory

```
/* LFA */


PROC BOXPLOT DATA=FinalSet;
      PLOT LFA*Advisory;
RUN;
```

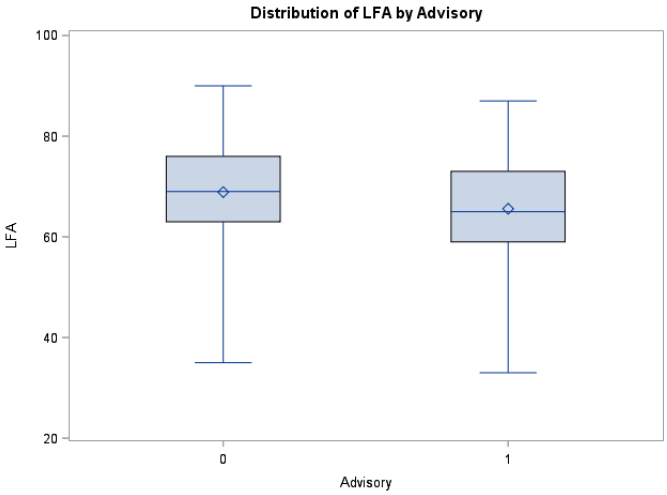
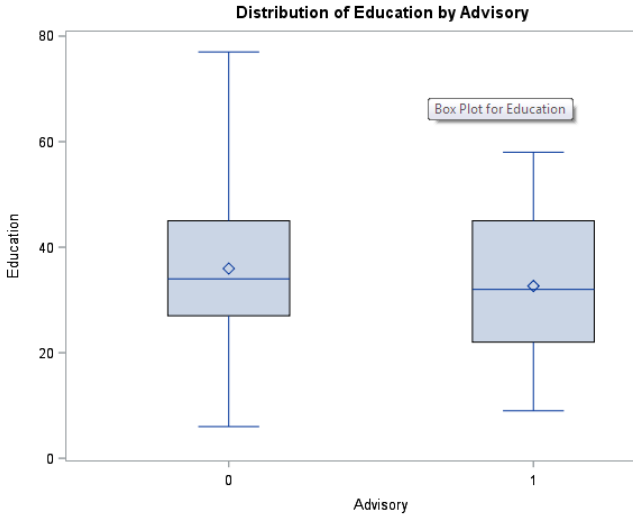


Figure B5: Boxplot of Education Scores by Advisory

```
/* Education */


PROC BOXPLOT DATA=FinalSet;
      PLOT Education*Advisory;
RUN;
```



Appendix C: Regression SAS Code and Output

Simple Linear Regression

```
PROC REG DATA=FinalSet PLOTS=none;
      MODEL Housing=Advisory;
RUN; /* Housing = B0 + B1*Advisory */
```

Figure C1: Regression Output for Housing

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	68.75000	0.81383	84.48	<.0001
Advisory	1	-4.08333	1.83766	-2.22	0.0268

```
PROC REG DATA=FinalSet PLOTS=none;
      MODEL Education=Advisory;
RUN; /* Education = B0 + B1*Advisory */
```

Figure C2: Regression Output for Education

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	35.95181	0.68435	52.53	<.0001
Advisory	1	-3.29749	1.54530	-2.13	0.0334

```
PROC REG DATA=FinalSet PLOTS=none;
      MODEL LFA=Advisory;
RUN; /* LFA = B0 + B1*Advisory */
```

Figure C3: Regression Output for LFA

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	68.88554	0.51447	133.90	<.0001
Advisory	1	-3.30530	1.16169	-2.85	0.0047

```
PROC REG DATA=FinalSet PLOTS=none;
      MODEL Income=Advisory;
RUN; /* Income = B0 + B1*Advisory */
```

Figure C4: Regression Output for Income

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	60.43373	0.80579	75.00	<.0001
Advisory	1	-7.39670	1.81951	-4.07	<.0001

```
PROC REG DATA=FinalSet PLOTS=none;
      MODEL CWB=Advisory;
RUN; /* CWB = B0 + B1*Advisory */
```

Figure C5: Regression Output for CWB

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	59.82008	0.44021	135.89	<.0001
Advisory	1	-4.59594	1.03722	-4.43	<.0001

In each simple linear regression model, Advisory is significant at the five percent level. This means that we are in favour of saying that the presence of a water

advisory is associated with a decrease in CWB component scores for Indigenous communities.

To see if location is a potential confounder for the association between component scores and the presence of a water advisory, five multiple linear regression models were fit; each score was fit against Advisory and ProvCode (the indicator variables for the provinces).

Multiple Linear Regression

```
PROC REG DATA=FinalSet PLOTS=none;
      MODEL Housing=Advisory ProvCode;
RUN; /* Housing = B0 + B1*Advisory +
      B2*ProvCode */
```

Figure C6: Multiple Regression Output for Housing

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	65.26936	2.04167	31.97	<.0001
Advisory	1	-3.44227	1.86441	-1.85	0.0656
ProvCode	1	0.46186	0.24860	1.86	0.0639

```
PROC REG DATA=FinalSet PLOTS=none;
      MODEL Income=Advisory ProvCode;
RUN; /* Income = B0 + B1*Advisory +
      B2*ProvCode */
```

Figure C7: Multiple Regression Output for Income

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	52.67692	1.98661	26.52	<.0001
Advisory	1	-5.96806	1.81414	-3.29	0.0011
ProvCode	1	1.02928	0.24190	4.26	<.0001

```
PROC REG DATA=FinalSet PLOTS=none;
      MODEL LFA=Advisory ProvCode;
RUN; /* LFA = B0 + B1*Advisory +
      B2*ProvCode */
```

Figure C8: Multiple Regression Output for LFA

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	71.06446	1.29076	55.06	<.0001
Advisory	1	-3.70661	1.17869	-3.14	0.0018
ProvCode	1	-0.28913	0.15717	-1.84	0.0665

Figure C9: Multiple Regression Output for Education

```
PROC REG DATA=FinalSet PLOTS=none;
      MODEL Education=Advisory
      ProvCode;
RUN; /* Education = B0 + B1*Advisory +
      B2*ProvCode */
```

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	34.10074	1.72118	19.81	<.0001
Advisory	1	-2.95656	1.57175	-1.88	0.0607
ProvCode	1	0.24563	0.20958	1.17	0.2419

Figure C10: Multiple Regression Output for CWB

```
PROC REG DATA=FinalSet PLOTS=none;
      MODEL CWB=Advisory ProvCode;
RUN; /* CWB = B0 + B1*Advisory +
      B2*ProvCode */
```

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	57.95455	1.13244	51.18	<.0001
Advisory	1	-4.19014	1.06005	-3.95	<.0001
ProvCode	1	0.22607	0.12648	1.79	0.0743

In each case, province does not distort the association between scores and the presence of a water advisory. There is no instance with both the p-value for Advisory > 0.05 and the p-value for ProvCode <0.05. Therefore, province is not a confounder so there is no need to include it in the regression model.