LAB 1 ASSIGNMENT

DISPLAYING AND DESCRIBING DISTRIBUTIONS

In this lab assignment, you will use graphical and numerical tools in StatCrunch to explore the data about greenhouse gas (GHG) emissions reported from more than one hundred Alberta facilities in a ten-year period (2004-2013). You will examine the distribution of GHG emissions in 2013 as well as analyse the trend and changes in the distribution of GHG emissions over the aforementioned period. You will also identify the largest emitters and compare Alberta industrial sectors in terms of their total GHG emissions. In particular, you will assess the impact of Oil Sands on GHG emissions. The questions in the lab assignment refer to the StatCrunch tools discussed in the *Lab 1 Instructions*.

Alberta Facility Greenhouse Gas Emissions

Global warming is driven mainly by the following greenhouse gases (GHG): carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), hydrofluorocarbons (HFC), perfluorocarbons (PFC), and sulphur hexafluoride (SF_6). GHGs are not equal in their effect on the atmosphere, however; each GHG has a unique average atmospheric lifetime and heat-trapping potential. GHG emissions are often calculated and reported in terms of how much CO_2 would be required to produce a similar warming effect over a given time horizon (usually 100 years). This is called the CO_2 equivalent (CO_2e) value and is calculated by multiplying the amount of the gas by its associated global warming potential. The idea is to express the impact of each different greenhouse gas in terms of the amount of CO2 that would create the same amount of warming.

Since 2004, the Government of Alberta has required industrial facilities across the province to report their GHG emissions annually through the *Alberta Specified Gas Reporting Regulation* program. Initially, the program had applied to all facilities that emit the equivalent of 100,000 tonnes (100 kilotonnes, or 100 kt) or more of GHGs in CO_2 per year. This threshold was lowered to 50 kt for the 2010 calendar year and subsequent reporting years. Note that the program just covers industrial sources; emissions from other sources including transportation, agriculture, commercial, and residential are not included in the inventory. In 2013, reported industrial emissions accounted for 49.7 per cent of Alberta's total emissions.

In this lab assignment, you will examine the greenhouse gas data from up to 167 facilities in Alberta in a ten-year period (2004-2013). The data are posted under *Open Government License- Alberta* at https://open.alberta.ca/dataset/specified-gas-reporting-regulation-summary.

The data are also available in the StatCrunch file *lab1.txt* located in eClass or on the *STAT 151* Laboratories web site at http://www.stat.ualberta.ca/statslabs/stat151/index.htm (click *Stat 151* link, and *Data* for *Lab 1*). The following is a description of the variables in the data file:

Variable Name	**Description of Variable**
YEAR	Year in which the values of the emissions were obtained,
SECTOR	Industry type,
LEGAL NAME	Legal name of a corporation.
FACILITY	Facility Name,
CO2	CO_2 emissions (in tonnes),
NH4	NH_4 emissions (in tonnes, CO_2 equivalent),
N2O	N_2O emissions (in tonnes, CO_2 equivalent),
HFC	HFC emissions (in tonnes, CO_2 equivalent),
PFC	PFC emissions (in tonnes, CO_2 equivalent),
SF6	SF_6 emissions (in tonnes, CO_2 equivalent),
TOTAL	Total CO_2 equivalent emissions (in tonnes).

1. First examine the data collection process.

   (a) What assumptions must be made about the data collection process to ensure that valid conclusions about the trends in emissions can be reached based on the data? What are the limitations of the data?

   (b) The number of facilities reporting their annual emissions varies from year to year. What are possible direct reasons for the variability? How may this affect the conclusions about the data?

   (c) Some facilities that do not exceed the threshold of 50,000 tonnes (50 kt) are voluntary participants in the program and are included in the inventory. Provide an advantage and disadvantage in how the voluntary-reported greenhouse gas emissions may affect conclusions for the data.

   (d) In the 2004 – 2009 period, the mandatory threshold for reporting emissions was 100 kt. This threshold was lowered to 50 kt for the 2010 calendar year and subsequent reporting years. How may this change affect any comparisons in total emissions for years before and after the threshold change?

2. Now examine the reported Alberta greenhouse gas emissions for 2013.

   (a) How many facilities reported GHG emissions in 2013 and what were their total CO2e emissions? What are the mean, standard deviation, and maximum value of the total CO2e emissions? What are the three quartiles? (Consider reporting values in megatonnes (Mt), where 1 Mt = 1 million tonnes or 1,000 kilotonnes, for better clarity.)

   (b) How many facilities reported their emissions voluntarily in 2013? What is the percentage contribution of total reported greenhouse gas emissions by facilities reporting their emissions voluntarily in 2013?

   (c) Which facility in Alberta had the largest total CO2e emissions in 2013? What were its total emissions in that year? Compare the total emissions from the single facility with the total emissions of the 110 lowest CO2e emitters for that year (their emissions are below 300,000).

   (d) What is the 95^{th} percentile of the total CO2e emissions and what is the percentage contribution of the remaining 5% of the facilities in total CO2e emissions? Report the appropriate values from the output. There is no need to paste the outputs into your report.

   (e) What were the total CO2e emissions for each of the 16 sectors in that year? Which sector reported the largest share of 2013 total CO2e emissions in Alberta and what was its percentage contribution? Paste the appropriate output into your report. (Make sure to use the "*Store in data table*" option for later parts of the lab assignment.) Also obtain a pie chart to display the contribution of each sector in total CO2e emissions in 2013. Paste the chart into your report. Comment briefly on the largest and smallest sectors.

   (f) What was the percentage contribution of each gas type in total reported GHG emissions in 2013? In order to answer the question, obtain the total emissions for each gas type and a pie chart displaying the contribution of total reported GHG emissions by gas type and comment briefly. (Make sure to use the "*Store in data table*" option.)

   (g) Obtain a frequency histogram of total CO2e emissions in 2013 with bins starting at 0 and a width of 1,000,000. The format of your histogram should be the same as the format of the histogram in the *Lab1Instructions* (title and axis labels). Describe the shape of the histogram. What do you conclude about the distribution of total CO2e emissions in 2013?

3. Now explore long-term trends in reported GHG emissions in Alberta. In order to meaningfully compare the data over time, emissions from facilities reporting their emissions voluntarily must be excluded. Therefore, for the comparisons over the entire period (2004 – 2013), consider only facilities whose emissions are 100 kt CO2e or more.

   (a) Obtain the total CO2e emissions for all facilities with total CO2e emissions of 100 kt or more for each year of the period (2004 – 2013). Paste the summaries into your report. (Make sure to use the "*Store in data table*" option.) Based on the summaries, calculate the percentage increase in total emissions over the entire period.

   (b) Obtain a timeplot for the summaries obtained in part (a). Paste the plot into your report. Comment about the change in total CO2e over the period.

   (c) Obtain side-by-side boxplots of total CO2e emissions by *Year* for all facilities with total emissions of 100 kt or more for each year of the period (2004 – 2013). Check the "Use fences to identify outliers" option. Paste the graph into your report. Compare the centres, spreads, and shapes of the ten distributions and provide a brief summary. Identify the largest emitters over the period (hover the cursor over the outliers). If the same facility appears more than once as the largest emitter in the period, assess progress in reducing emissions in the facility over time. (Note that some facilities may have slightly different names from year to year.)

   (d) Compare the information about emissions derived from the timeplot in part (b) and the side-by-side boxplots in (c). What information can be derived from one plot but not from the other? Explain briefly.

4. In this question, assess the impact of Oil Sands on total CO2e emissions in the period (2004 – 2013).

   (a) Obtain the summary statistics (number of facilities, mean, standard deviation, median, sum, and IQR) for total CO2e emissions for the combined "Oil Sands In Situ Extraction" and "Oil Sands Mining and Upgrading" sectors. (Make sure to use the check "*Store in data table*" option for later parts of the lab assignment.) Include only facilities with total emissions of 100 kt or more for each year in the period (2004 – 2013). Paste the output into your report. Comment briefly about the change in the total emissions by the industry in the period.

   (b) Obtain a timeplot for the summaries obtained in part (a). Paste the plot into your report. Comment about the change in total CO2e in the Oil Sands sectors over the period.

   (c) Obtain a new variable "Alberta minus Oils Sands", which is the difference of "Total CO2e for all facilities with emissions 100 kt or more" (the sums obtained in Question 3a) and the emissions of the combined Oil Sands categories (the sums obtained in Question 4a). Paste the summaries into your report. To obtain the variable, use the "*Store in data table*" option or the expression builder found using the following path: *Data → Compute → Expression*

   Comment about the change in total CO2e emissions for all Alberta facilities with Oil Sands facilities and without Oil Sands facilities in the period (2004 – 2013). Then use the summaries to obtain a timeplot with two lines over the period: 1) total CO2e emissions for all Alberta facilities without Oil Sands facilities and 2) Oil Sands emissions only. Paste the chart into your report. Comment briefly.

LAB 1 ASSIGNMENT: MARKING SCHEMA

Proper cover page (see example on eClass for proper format) and appearance (lab reports must be **typed)**: 10 marks

Question 1 (12)

(a) Assumptions: 2 marks
Limitations: 2 marks
(b) Reasons for variability in the number of facilities: 2 marks
Impact on the conclusions about the data: 2 marks
(c) Advantage and disadvantage of the voluntary submissions: 2 marks
(d) Effect of change in the threshold value: 2 marks

Question 2 (54)

(a) Total number of facilities: 2 marks
Total emissions: 2 marks
Mean, standard deviation, and maximum value: 3 marks (1 each)
Quartiles: 3 marks (1 each)

(b) Number of voluntary facilities in 2013: 2 marks
Percentage of total emissions of voluntary facilities: 2 marks
(c) Largest emitter in 2013: 2 marks
Total emissions of the largest emitter in 2013: 2 marks
Comparison with 110 lowest emitters: 2 marks
(d) 95^{th} percentile: 2 marks
The contribution of the top 5% in total CO2e emissions: 2 marks
(e) Total CO2e emissions for each of the 16 sectors in 2013: 4 marks
Sector with the largest total CO2e emissions in 2013: 2 marks
Contribution of the sector with the largest total CO2e emissions: 2 marks
Pie chart for total CO2e emissions for 16 sectors: 4 marks
Comments: 2 marks
(f) Total emissions for each gas type: 2 marks
Pie chart for the contribution of each gas in total emissions: 4 marks
Comments: 2 marks
(g) Properly formatted frequency histogram of total CO2e emissions in 2013: 4 marks
Shape: 2 marks
Conclusions: 2 marks

Question 3 (28)

(a) Total CO2e emissions for facilities 100 kt or more: 2 marks
Percentage increase in total emissions: 2 marks
(b) Timeplot: 4 marks
Comments: 2 marks
(c) Side-by-side boxplot (with proper format) of total CO2e emissions by *Year*: 4 marks
Summary comments about the centres, spreads and shapes of the 10 distributions: 4 marks
Largest emitter(s): 2 marks
Progress in reducing emissions by the largest emitter(s) over the period: 2 marks
(d) Comparison: 3 marks
Information derived from one plot but not the other: 3 marks

Question 4 (26)

(a) Summary statistics: 4 marks
Comments about change in total emissions for Oil Sands: 2 marks

(b) Timeplot: 4 marks
Comments: 2 marks

(c) Summaries: 4 marks
Comments: 3 marks
Timeplot: 4 marks
Timeplot interpretation: 3 marks

TOTAL = 130

Created by: Henryk Kolacz;
Edited by: Paul Cartledge
University of Alberta
June, 2017