LAB 1 ASSIGNMENT SOLUTIONS

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 154 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?

   In order to reach valid conclusions about the data, each facility must use the most appropriate and common GHG estimation methodology for their particular industry. The methodology should be consistent over the reporting period to allow meaningful comparisons of emissions from one year to another. Control equipment must be functioning properly and its accuracy and reliability be systematically assessed. Moreover, strict verification protocols and estimations by independent third-party verifiers must be in used to validate the data and make compliance determinations.

   As the methods to estimate GHG emissions are constantly being refined within each industry, the data may be limited by the lack of consistency and also variation in the annual facility list (facilities may exceed or fall below the reporting threshold in subsequent years).

   (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?

   Changes in production levels, processes, technologies, and the types of fuels used at a facility can all result in a change in the annual emissions reported, such that a facility may fall below or conversely attain the reporting threshold of 50 kt CO2e from one year to the next. The number of voluntary reporters may also change each year, affecting the number of reporting facilities.

   The fact that emissions from different facilities may be included in the inventory year after year makes it harder to compare the total CO2e emissions if the number of those facilities is large.

   (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.

   Voluntary participants provide a more accurate picture of the sources and amounts of emissions from Alberta facilities because the more facilities reporting their emissions (voluntary or not), the better the estimate of the total emissions in the province.

   On the other hand, voluntary participants may affect the summaries for the data, such as lowering the mean and median of total emissions. Note that some manufacturing or industrial operations that use many scattered facilities over the province, each with total annual emissions just below the threshold of 50 kt will likely be excluded from the inventory (if not voluntary participants).

   (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?

   The decrease in the threshold level from 100 kt to 50 kt has increased the number of reporting facilities in the subsequent years. The number of reporting facilities prior to the change varied between 93 in 2009 and 109 in 2008. This makes it harder to compare the trends in the total emissions between 2004 – 2009 and 2010 – 2013.

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.)

The requested summaries are provided in the output below:

Options

Summary statistics:
Where: Year = 2013

Column	n	Mean	Std. dev.	Median	Q1	Q3	Max	Sum
Total (tonnes CO2 eq)	167	794341.38	1906664.2	165157	70983	493288	12548723	1.3265501e8

For the 2013 calendar year, 167 facilities in Alberta reported GHG emissions. The total CO2e emissions for these facilities equalled 132.655 Mt. This number includes facilities that were below the 50 kt threshold, but voluntarily reported.

The average total CO2e emissions were 794,341.38 tonnes (0.794 Mt) with a standard deviation of 1,906,664.2 tonnes (1.906 Mt) and a maximum of 12,548,723 tonnes (12.549 Mt). The median (or second quartile) is 165,157 tonnes (which is much smaller than the mean). The first and the third quartiles are 70,983 and 493,288 tonnes, respectively.

(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?

Options

Summary statistics:
Where: Year = 2013 and "Total (tonnes CO2 eq)" < 50000

Column	n	Sum
Total (tonnes CO2 eq)	20	489762

Twenty facilities reported their emissions voluntarily (total CO2e emissions below 50 kt threshold) with the total of 489,762 tonnes. The percentage contribution of the facilities with the total emissions not exceeding the threshold of 50 kt is 489,762/132,655,010 = 0.00369, which is less than 0.4%. The relatively small percentage of emissions from the voluntary facilities does not greatly change the above summaries for all 167 facilities. See the output below (not required).

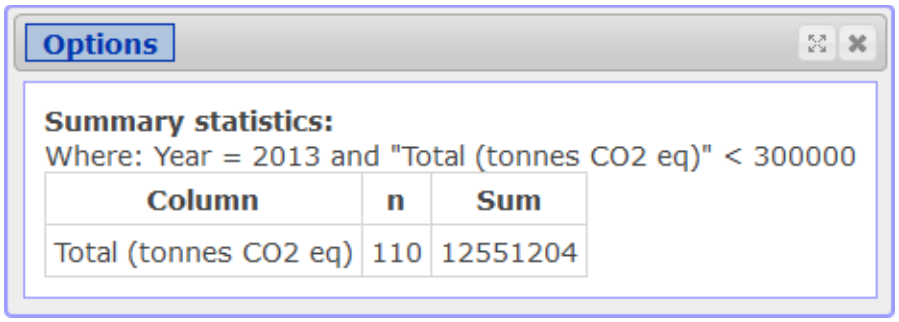
Options

Summary statistics:
Where: Year = 2013 and "Total (tonnes CO2 eq)" >= 50000

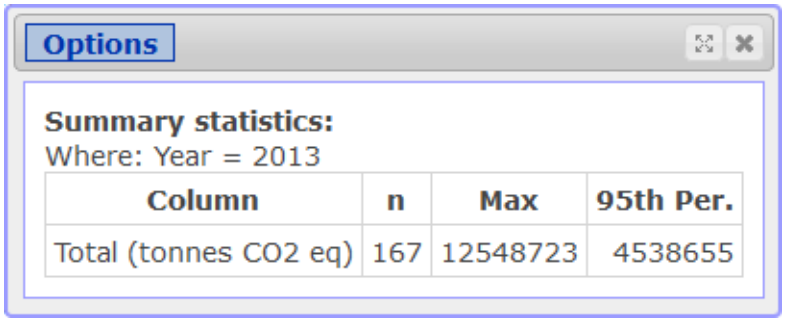
Column	n	Mean	Std. dev.	Median	Q1	Q3	Max	Sum
Total (tonnes CO2 eq)	147	899083.33	2010245.2	220450	93648	546385	12548723	1.3216525e8

(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).

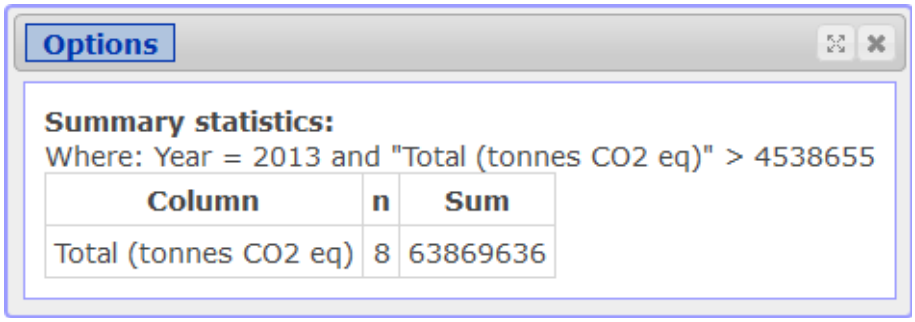
The largest total CO2e emissions in 2013 from a single facility were 12.548723 Mt. The value comes from Syncrude Canada Ltd.'s facility at Mildred Lake and Aurora North Plant Sites. The total CO2e emissions from the facility approximately matched the emissions (12.551204 Mt) produced by the 110 lowest CO2e emitters (or the 90 lowest CO2e emitters with annual emissions exceeding the 50kt threshold).



(d) What is the 95th 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.



The 95th percentile is 4,538,655 tonnes (95% of all 167 facilities have total CO2e emissions at or below 4.539 Mt). The remaining 5% of facilities (the top 8) contribute 63,869,636 tonnes, which constitutes almost 50% of all total CO2e emissions in Alberta.
(Exact value = 63,869,636/132,655,010 = 0.48147 = 48.147%)



(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.

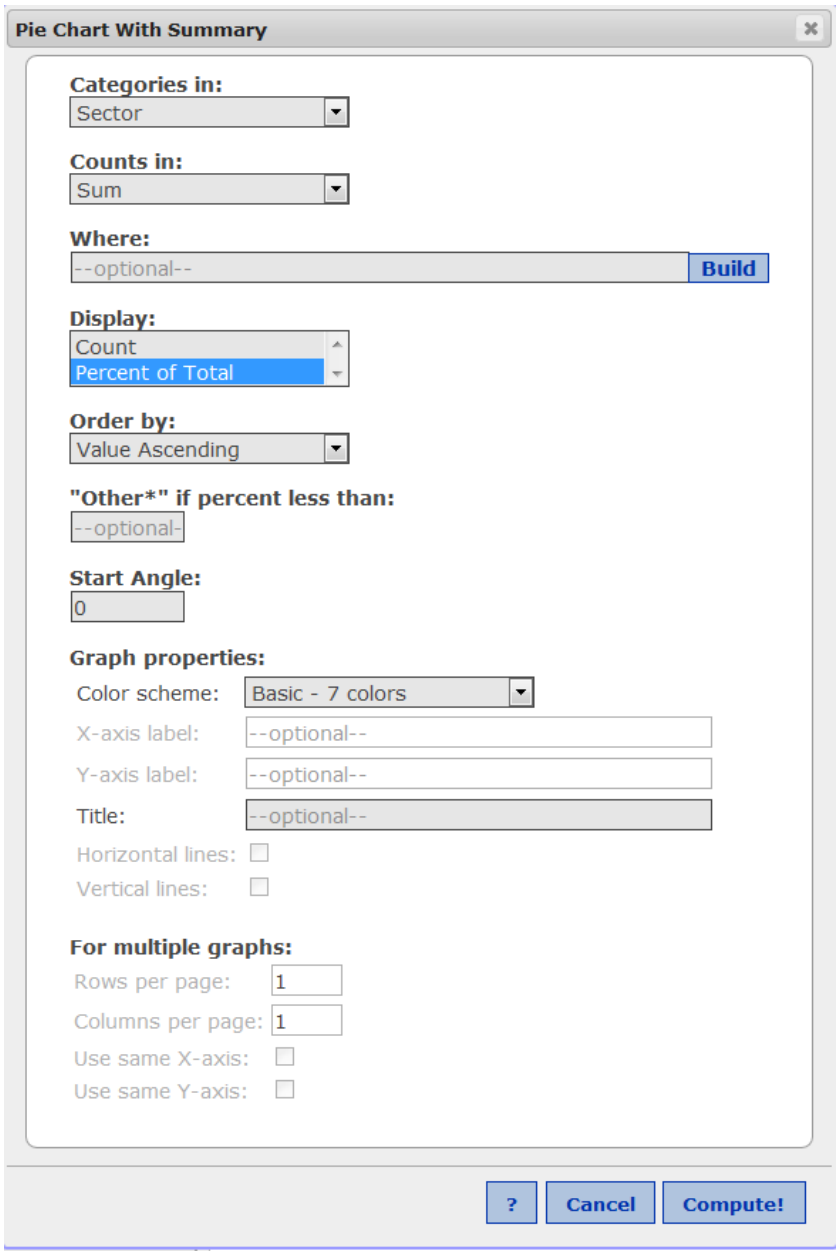
Options

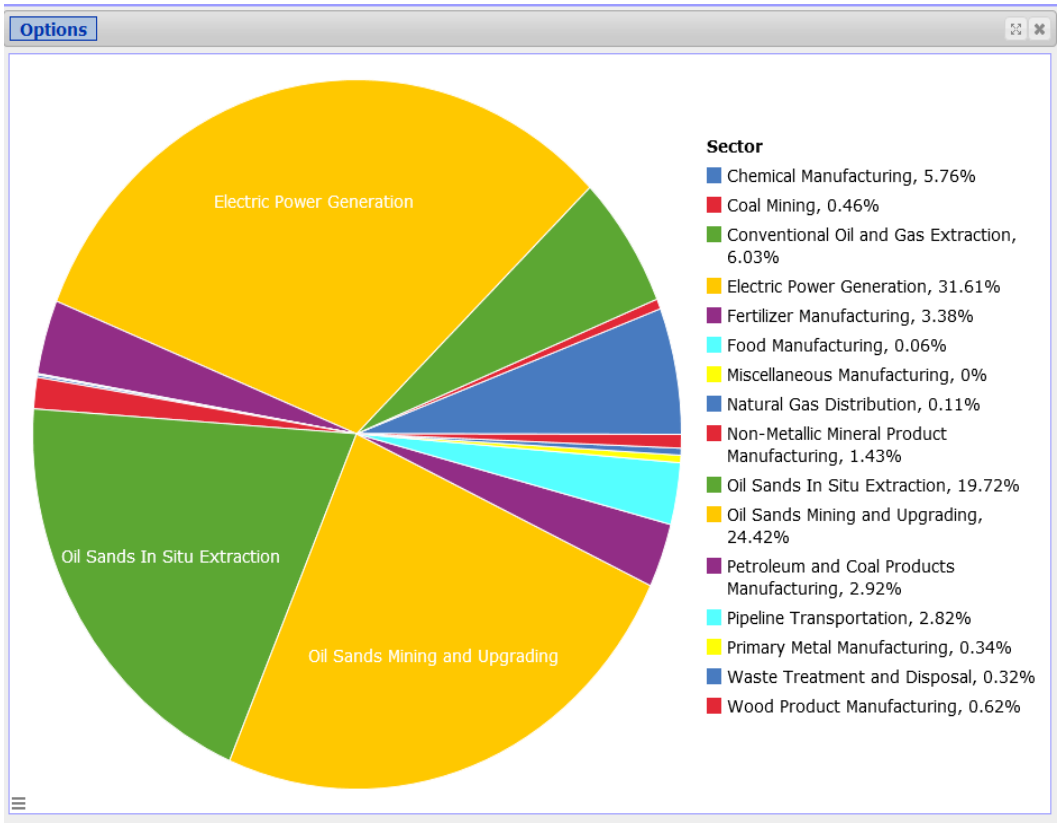
Summary statistics for Total (tonnes CO2 eq):
Where: Year = 2013
Group by: Sector

Sector	n	Sum
Chemical Manufacturing	12	7647359
Coal Mining	4	613974
Conventional Oil and Gas Extraction	58	8005222
Electric Power Generation	23	41932714
Fertilizer Manufacturing	5	4486525
Food Manufacturing	1	82665
Miscellaneous Manufacturing	2	2559
Natural Gas Distribution	1	140260
Non-Metallic Mineral Product Manufacturing	4	1892374
Oil Sands In Situ Extraction	22	26160060
Oil Sands Mining and Upgrading	8	32390594
Petroleum and Coal Products Manufacturing	5	3876739
Pipeline Transportation	4	3735903
Primary Metal Manufacturing	2	449187
Waste Treatment and Disposal	7	418145
Wood Product Manufacturing	9	820731

The total CO2e emissions for each of the 16 sectors is shown in the above figure. Based on the output, the electric power generation sector is the largest source of total CO2e emissions with a value of 41.933 Mt (or 31.610% of all total CO2e emissions). Note that if one combines the two Oil Sands sectors (Oil Sands in situ extraction and Oil Sands mining and upgrading) together, the combined value of 58.551 Mt (or 44.138% of all total CO2e emissions) would be the largest.

After storing the output from part (e) into the data table, the following StatCrunch tool was used to obtain the pie chart:



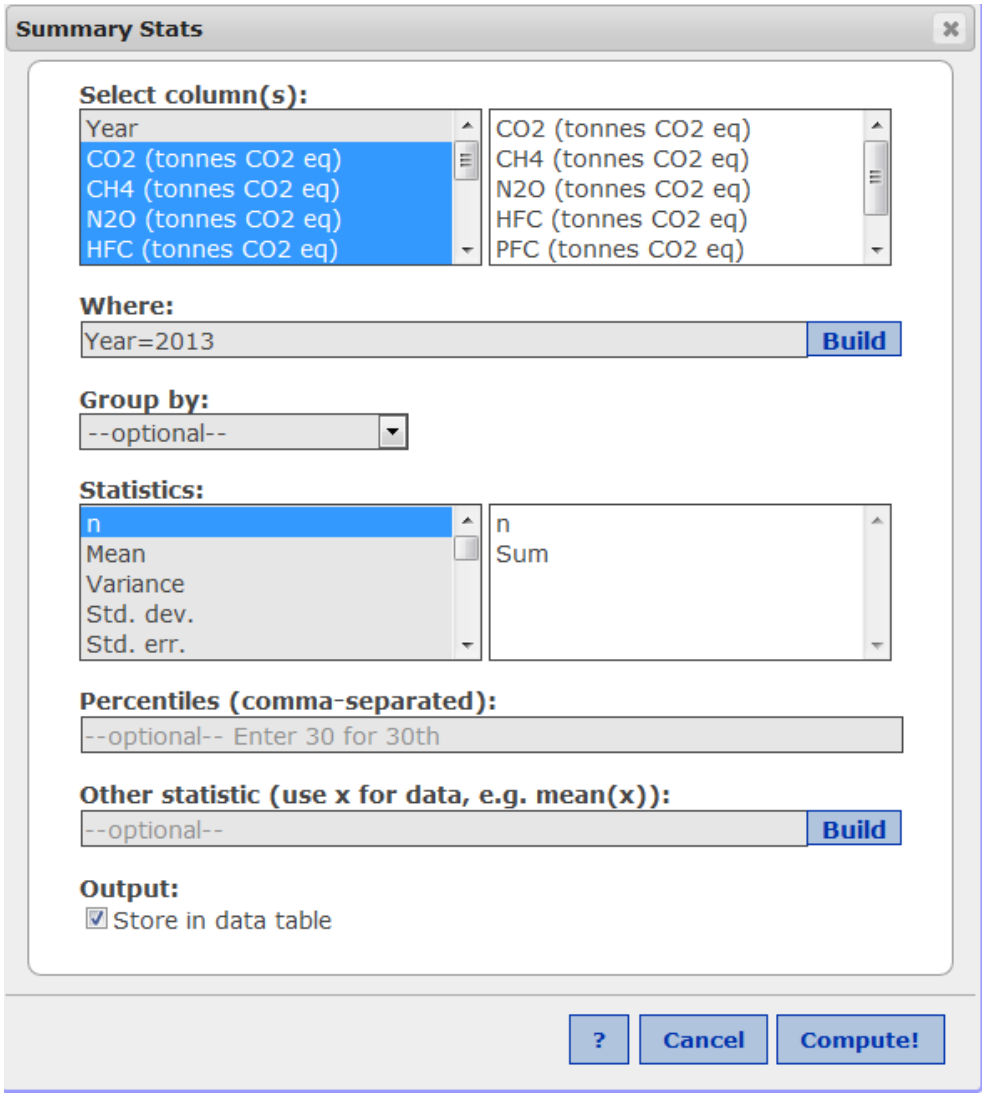


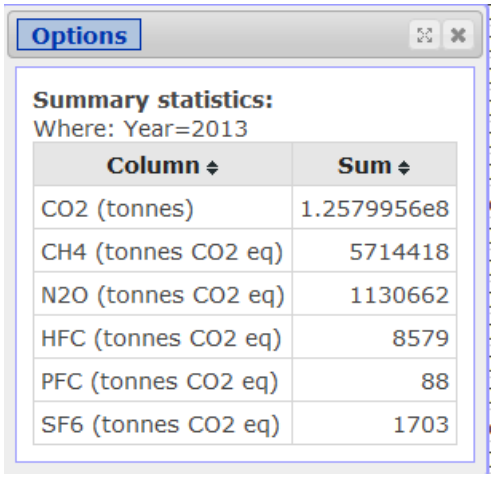
The pie chart confirms that the electric power generation sector is the largest. Also, combining the two Oil Sands sectors would be even larger, even though they are individually smaller than the electric power generation sector. The smallest sector is miscellaneous manufacturing, reported to be 0% from the pie chart, but 2559 tonnes (or 0.00193% of all total CO2e emissions) from the table.

(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.)

The StatCrunch tool, corresponding table output, and pie chart are shown below.

Approximately 95% of the total reported GHG emissions (exact value of 94.832%) were in the form of carbon dioxide (CO_2), noticeably the largest section of the pie chart. Most of the remainder was in the form of methane, nitrous oxide, and hydrofluorocarbons. More precisely, the percentages were: CH_4 (4.308% or 5.714 Mt), N_2O (0.852% or 1.131 Mt), HFCs (0.00647% or 8.579 kt), SF_6 (0.00128% or 1.703 kt), and PFCs (less than 0.001% or 0.088 kt). The last three percentages are so small, that they don't visually appear on the pie chart.

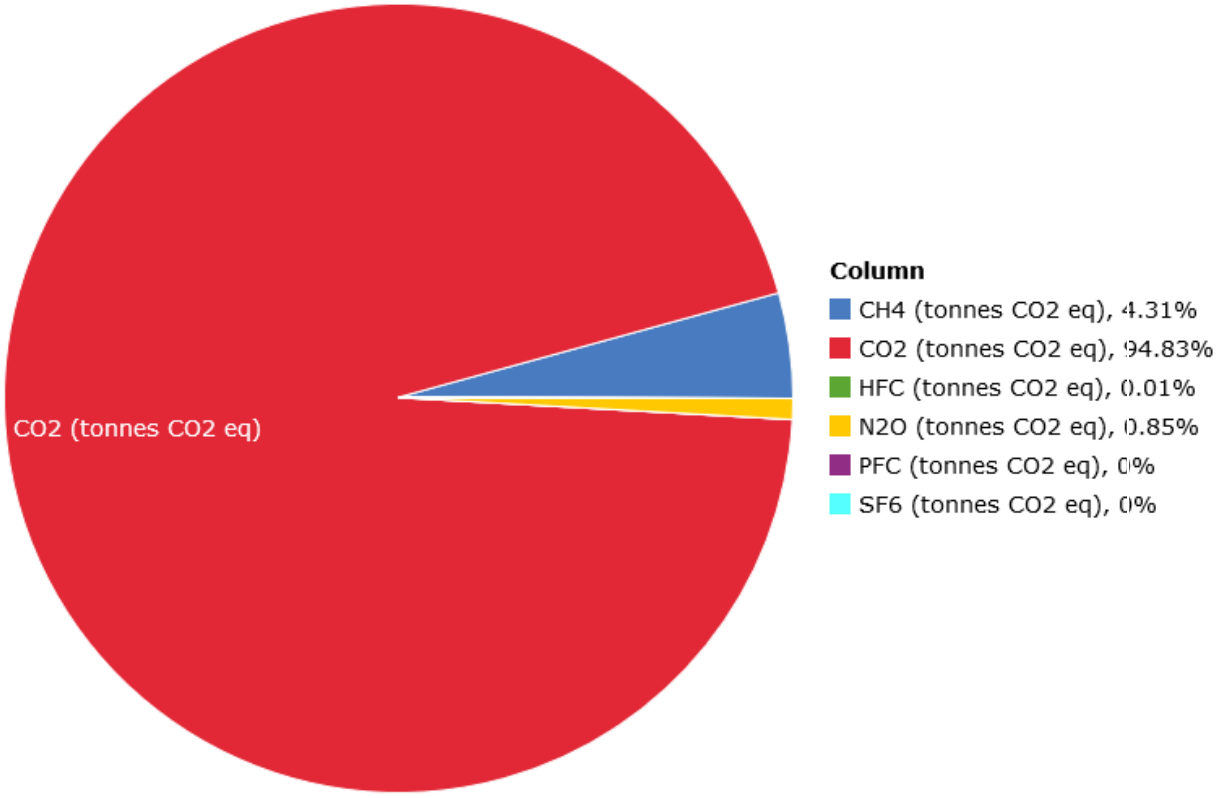




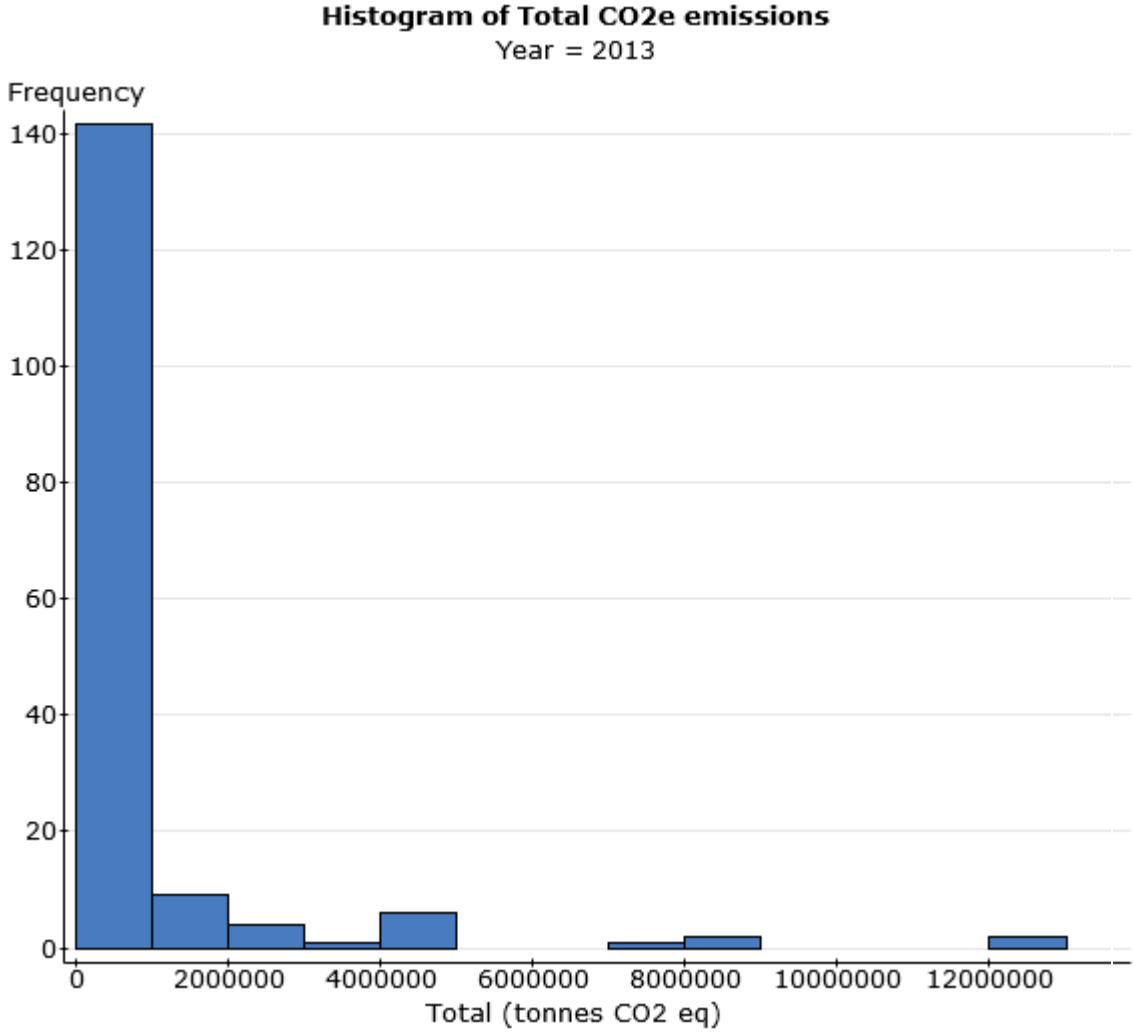
Column	Sum
CO2 (tonnes)	1.2579956e8
CH4 (tonnes CO2 eq)	5714418
N2O (tonnes CO2 eq)	1130662
HFC (tonnes CO2 eq)	8579
PFC (tonnes CO2 eq)	88
SF6 (tonnes CO2 eq)	1703

The summary can also be obtained in the data table:

GHGs	Sum
CO2 (tonnes)	1.2579956e8
CH4 (tonnes CO2 eq)	5714418
N2O (tonnes CO2 eq)	1130662
HFC (tonnes CO2 eq)	8579
PFC (tonnes CO2 eq)	88
SF6 (tonnes CO2 eq)	1703



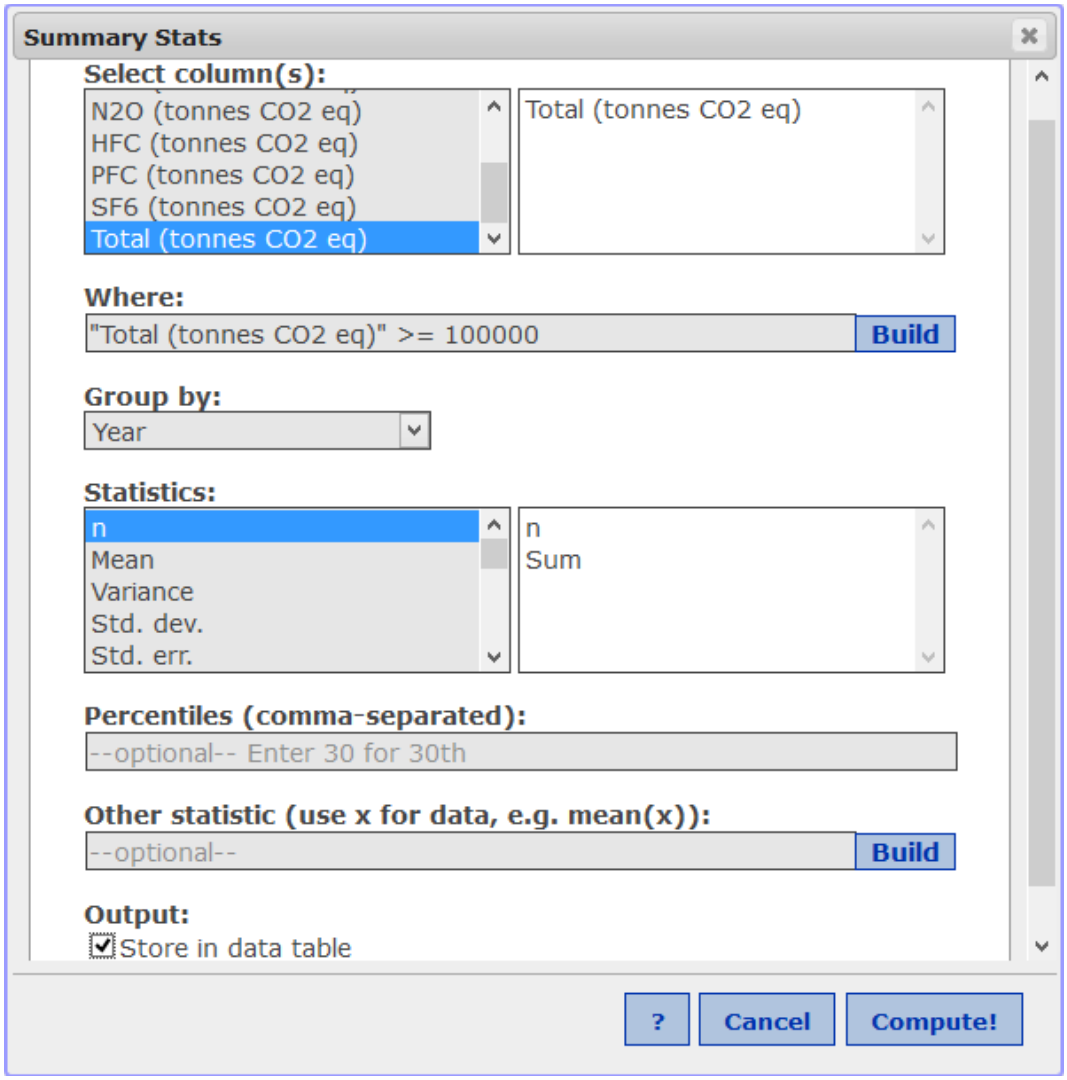
(g) Obtain a frequency histogram of total CO2e emissions in 2013 with bins starting at 0 and a width of 1,000,000. The format of your histogram should be the same as the format of the histogram in the *Lab1Instructions* (title and axis labels). Describe the shape of the histogram. What do you conclude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



The histogram of total CO2e emissions in 2013 is single-peaked (unimodal) and heavily skewed to the right. Most of the facilities (140 out of 167) produced emissions below 1 Mt (1 million tonnes). There are several facilities with the total CO2e emissions exceeding 4 Mt. There are two facilities with the total emissions exceeding 12 Mt. Therefore, only a few facilities seem to be accounting for most of the 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.



Year	Facilities at least 100 kt	Total CO2e
2004	92	1.0712893e8
2005	88	1.0623038e8
2006	91	1.1313808e8
2007	89	1.1140294e8
2008	93	1.099846e8
2009	91	1.1374536e8
2010	100	1.2028061e8
2011	104	1.2148721e8
2012	102	1.2412236e8
2013	106	1.2927642e8

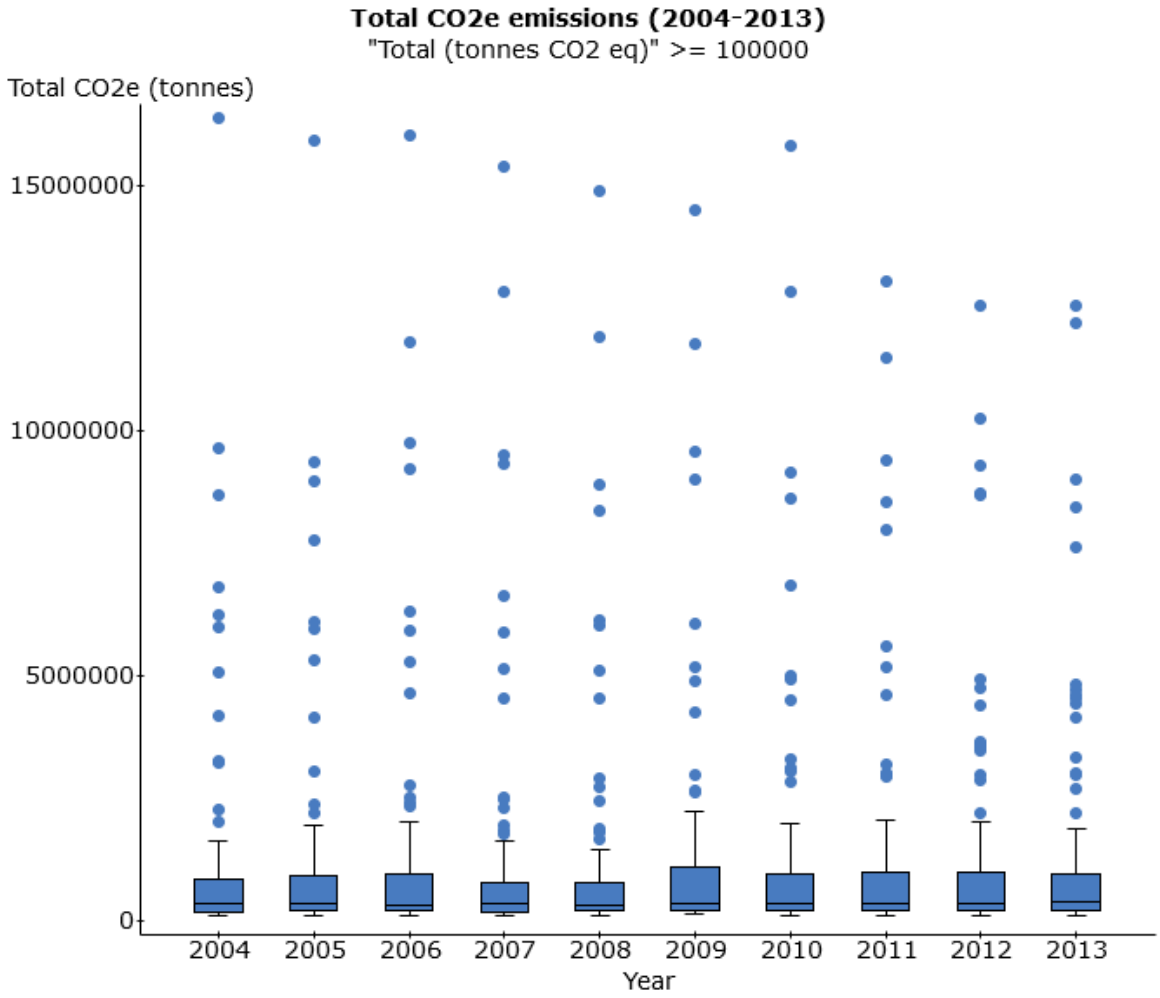
Over the ten-year period, the total emissions increase by 20.674%.

$$\frac{1.2927642 - 1.0712893}{1.0712893} = 0.20674 = 20.674\%$$

(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.

Overall, there is a steep increase in total CO2e emissions over the period. The total emissions fall between 2006 to 2008. The rate of increase seems to be steeper between 2008 to 2013.

(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.)



The distribution of total CO2e for each year in the period (2004 – 2013) is highly skewed to the right with several outliers each. The median values appear to be very similar over the period, in the range of a few hundred thousand tonnes. The spread measured by IQR is also very consistent over the period. Starting in 2011, there is some reduction in emissions for the largest emitters, but given the short time frame (2011 – 2013), it is not clear whether this is an overall trend or temporary variation.

Sundance Thermal Electric Power Generating Plant and Mildred Lake and Aurora North Plant Sites (Oil Sands) facilities are the largest CO2e emitters. Sundance considerably lowered its emissions over the period, from the high of 16.016 Mt in 2006 to a final value of 12.181 Mt in 2013. The emissions from Mildred Lake and Aurora North Plant Sites, however, remain in the same range of 12.5 – 13 Mt from 2010 to 2013. Note that those two plants come from very different sectors with different potentials for emissions reductions.

(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.

The timeplot shows the sum of all total CO2e emissions for all facilities with emissions 100 kt or more for each year in the period, which is important information in terms of understanding global warming. Yet the timeplot provides only one piece of information (total CO2e for all facilities) for each year. The boxplots provide much more information about the distribution of total CO2e emissions in each year like shape, center, spread, and outliers. They do not, however, provide any information about the total CO2e each year.

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.

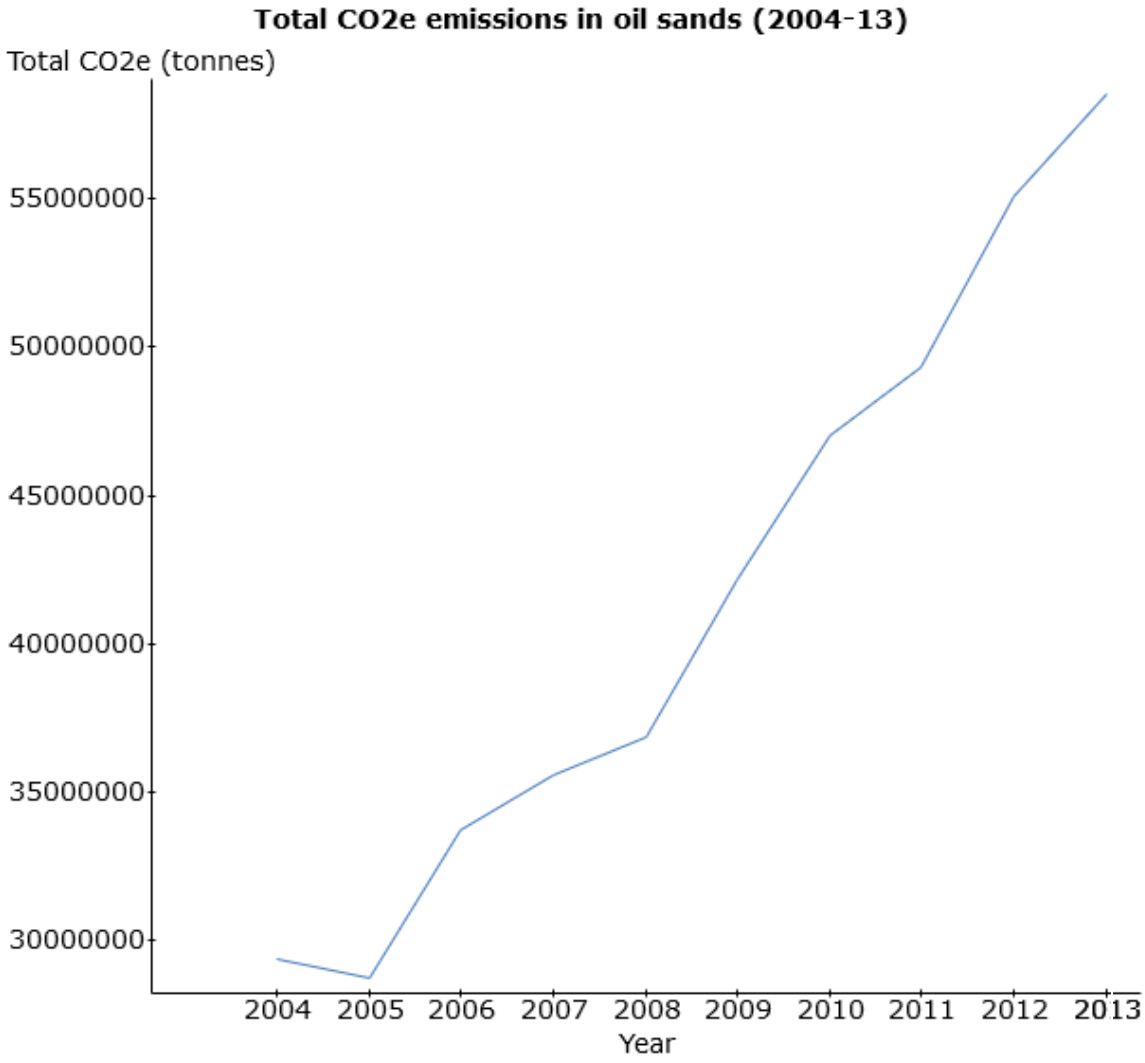
Options

Summary statistics for Total (tonnes CO2 eq):
Where: ("Sector" = "Oil Sands In Situ Extraction" or "Sector" = "Oil Sands Mining and Upgrading") and "Total (tonnes CO2 eq)" >= 100000
Group by: Year

Year	n	Mean	Std. dev.	Median	Sum	IQR
2004	14	2101080.2	3187036.1	442513.5	29415123	1763021
2005	13	2214237.6	3042836.1	713340	28785089	1670021
2006	14	2411807.9	3671916.7	619641	33765311	2216486
2007	16	2222895.7	3699563.2	539033.5	35566331	1923513.5
2008	21	1756130.4	3096792.6	474234	36878739	892523
2009	22	1917210.5	2975177.4	807039.5	42178630	1617497
2010	25	1880558	2982131.4	540128	47013950	1732308
2011	27	1826107.6	2893541.2	700898	49304906	1787483
2012	27	2038635.6	2919615.8	676317	55043160	3202370
2013	29	2018820	2832974.5	695380	58545779	2615063

The number of Oil Sands facilities reporting their emissions increased from 14 in 2004 to 29 in 2013. With this large increase, the sum unsurprisingly increases over the period as well. The average CO2e emissions were about 2 Mt (2.412 to 1.756 Mt) for the entire period while the medians were much lower (0.443 to 0.807 Mt) and have a smaller range. From either measure of center, there does not appear to be a general decrease in total CO2e emissions. Both measures of spread stay approximately the same throughout the period but the IQR may be showing a general increase since 2008.

(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.



The timeplot shows a steep increase in emissions starting in 2005. The growth in emissions accelerates further after 2008 (steeper slope in the period 2008 to 2013).

(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.

If using the "*Store in data table*" option:

Summary statistics for Total (tonnes CO2e):
Where: ("Total (tonnes CO2e)">=100000 and Sector= "Oil Sands In Situ Extraction") or ("Total (tonnes CO2e)">=100000 and Sector="Oil Sands Mining and Upgrading")
Group by: Year

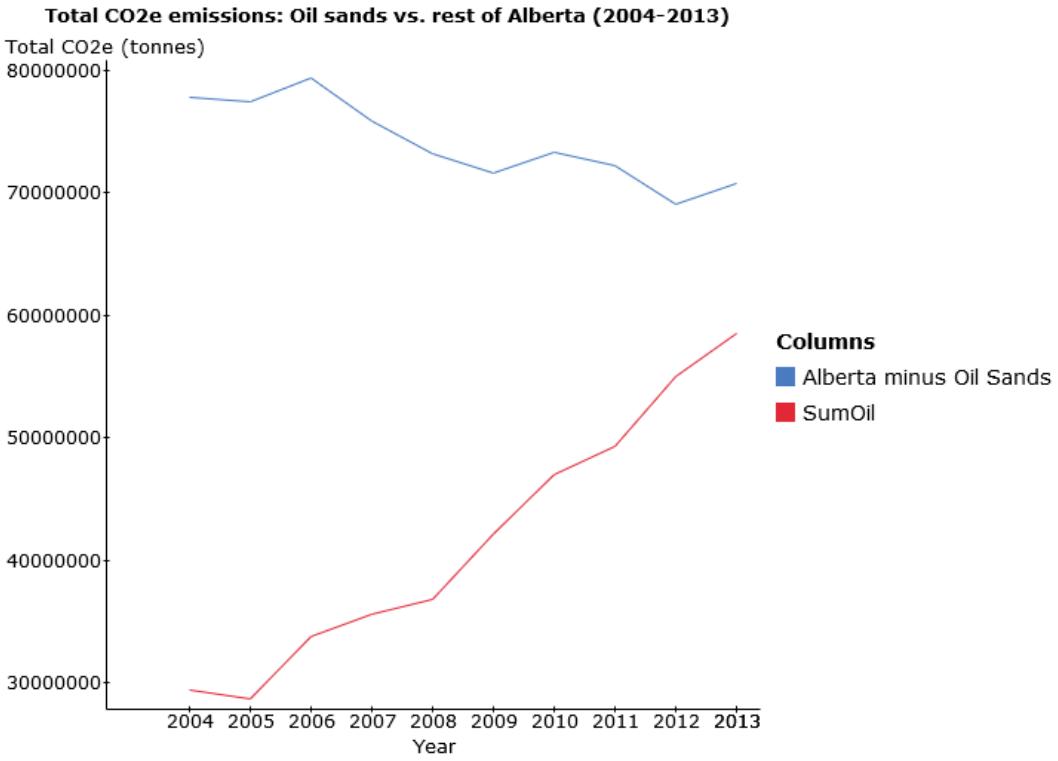
Year	Sum
2004	29415123
2005	28785089
2006	33765311
2007	35566331
2008	36878739
2009	42178630
2010	47013950
2011	49304906
2012	55043160
2013	58545779

Year	All Alberta Facilities >=100 kt	Oil Sands >=100 kt	Alberta minus Oil Sands
2004	1.0712893e8	29415123	77713805
2005	1.0623038e8	28785089	77445293
2006	1.1313808e8	33765311	79372767
2007	1.1140294e8	35566331	75836611
2008	1.099846e8	36878739	73105861
2009	1.1374536e8	42178630	71566725
2010	1.2028061e8	47013950	73266656
2011	1.2148721e8	49304906	72182300
2012	1.2412236e8	55043160	69079197
2013	1.2927642e8	58545779	70730636

Based on the above output, the facilities in Alberta without Oil Sands reduced their total CO2e emissions by approximately 9%. However, the emissions from Oil Sands have almost doubled (increased by 99%) in the same period.

$$\frac{7.0730636 - 7.7713805}{7.7713805} = -0.08986 = -8.986\%$$

$$\frac{5.8545779 - 2.9415123}{2.9415123} = 0.99033 = 99.033\%$$



The graph confirms the conclusions reached above. Total CO2e emissions for "Alberta minus Oil Sands" have slightly decreased over the period. The emissions from Oil Sands facilities, however, increased at a much higher rate than the rate of decrease of emissions from the remaining Alberta facilities. As a consequence, the total CO2e emissions in Alberta (the timeplot from Question 3b) have an overall increase in the period.

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