Electricity Consumption and Natural Gas Consumption Increase the Greenhouse Gas Emissions*

Jia Jia Ji

29 January 2021

Abstract

To analyze the effects of electricity and natural gas consumption on greenhouse gas emissions, I used the dataset containing the information about the attributes of municipal sites in the City of Toronto. I created some plots and tables to visualize the relationships between variables, and then I applied the multiple linear regression model to explore whether these effects were significant or not. Finally, I found that as more amount of electricity and natural gas consumed, greenhouse gas emissions would increase; also, the sites that were classified as some specific operation types tended to have much higher or much lower greenhouse gas emissions than others. These findings were beneficial for people to take more targeted measures to reduce the emissions and protect the environment effectively.

1 Introduction

Accompanying with the rapid advances in industrial and technological development, the environmental issues such as the resource scarcity, global warming and biodiversity destruction have raised people's attention. The large increase in greenhouse gas emissions is among these problems. The emitted greenhouse gas mainly come from the energy resources consumption used for electricity, transportation, industrial production and etc. Having more and more greenhouse gas emitted can cause the global warming, glacial melting and ecological destruction, which are detrimental to the global environment. Fortunately, many countries and people have realized this problem and are jointly committed to saving energy and reducing emissions. Therefore, understanding which factors are correlated with greenhouse gas emissions is crucial for both policymakers and residents. Specifically, policymakers can make more targeted policies to restrict the activities related to the greenhouse gas emissions, meanwhile, residents can take effective actions in daily life to reduce the emissions and protect the environment better.

This report aimed to explore whether the electricity consumption and natural gas consumption had significant effect on greenhouse gas emissions. The data was from Open Data Toronto (Gelfand 2020). The response variable was the annual greenhouse gas emissions for each municipal site in the City of Toronto, and the explanatory variables were the annual consumed amount of electricity and natural gas. Also, I treated each site's total floor area and average weekly operating hours as the confounding variables because the sites that had larger sizes and longer operating hours tended to consume more energy and have more greenhouse gas emitted. Moreover, the data contained 2 other relevant variables: operation type and city classification for each site, which were also considered as covariates. Accordingly, I investigated a secondary research question: whether these factors also significantly affected the greenhouse gas emissions. By answering these questions, people can understand which factors were strongly correlated with greenhouse gas emissions, so that they can take actions to reduce the emissions from these factors' perspectives.

The remaining report was organized into the following 5 sections. In Data section, I discussed the data source, the biases along with the data, the basic statistical characteristics of relevant variables and the data cleaning process. Also, I talked about the exploratory data analysis by showing some plots and tables to visualize the spreads of some variables and the relationships between them. Then, in Model section, I used 2 multiple

*Code and data are available at: https://github.com/jiaj6/greenhouse-gas-emissions-paper.

linear regression models with 2 confounding variables (total floor area and average weekly operating hours) to address the primary and secondary questions respectively. In Model 1, I included the electricity consumption and natural gas consumption as the only 2 explanatory variables to explore whether they had significant effect on greenhouse gas emissions. Besides, in Model 2, I added the operation type and city classification as another covariates to see whether these factors were also correlated with greenhouse gas emissions. In Results section, Model 1 showed that the electricity consumption and natural gas consumption both affected the greenhouse gas emissions significantly, and Model 2 indicated that the site's operation type also affected the emissions. Afterwards, I talked about the main findings in Conclusions section. Finally, I included the graphs relevant to model assumptions in Appendix section.

2 Data

2.1 Data Overview: Source of Data, Data Collection Methods, Biases along with Data, and Statistical Characteristics of Variables

I used `R` (R Core Team 2019), the `tidyverse` package (Wickham et al. 2019), the `ggplot2` package (Wickham 2016) and the `emmeans` package (Lenth 2019) to analyze the data. The report about the annual energy consumption and greenhouse gas emissions prepared by the Environment & Energy Division (Environment & Energy Program Administration 2018) provided information about how the data was collected.

This dataset was about the annual energy consumption and greenhouse gas (GHG) emissions for the municipal sites in the City of Toronto in 2018. All the sites in City of Toronto reported the data on the specified variables to the Environment & Energy Division, then the division collected the data and stored it in the energy management system. This dataset was published on Open Data Toronto (Gelfand 2020) under the Environment topic. 1482 sites in the City of Toronto formed the sample, and the frame was all sites in the City of Toronto because some sites were not required to provide data to the division. The population was all municipal sites in Canada. Since the sites managed by the City of Toronto were all required to report this data to the government, this data collection method was efficient. Also, it provided low-response rate and relatively accurate data. However, some data may not be recorded correctly due to the manual processing errors, or some sites with high GHG emissions may report lower values on purpose because the government encouraged the GHG emissions reduction. These situations could lead to the measurement biases. Another important issue was that the values for fuel oils, propane, coal, wood, district heating and cooling consumption were all missing in the dataset, which may limit the results of the analysis. Moreover, in terms of the external validity, since the sample size was not large (i.e. only 1482 records), the inferences may be inappropriate to generalize to other areas in Canada due to the social or cultural differences among cities.

This dataset included the data on annual GHG emissions, along with the basic attributes of each site, such as the operation type, city classification, address, annual water flow, annual consumption for each type of energy resources and etc. There were 22 variables in total. I dropped 15 irrelevant variables like operation name, address and annual water flow, and I focused on 7 variables that were relevant to the research questions: GHG emissions, operation type, city classification, total floor area, average weekly operating hours, annual electricity consumption and annual natural gas consumption. Specifically, GHG emissions was a discrete variable that was measured in kilograms and it ranged from 0 to 18193619. Similarly, total floor area was a discrete variable measured in square feet. The electricity consumption and natural gas consumption were both continuous variables, with electricity consumption ranging from 0 to 131780414 kWh and natural gas consumption ranging from 0 to 6940806 cubic meters. Besides, operation type and city classification were both categorical nominal variables. Operation type had 26 different levels and city classification had 11 levels, while the average weekly operating hours was a categorical ordinal variable that had 3 different levels.

2.2 Data Cleaning

This dataset contained only few missing values. Row 1463 was an empty record, so I dropped this entire row. Then, in the first row, the values for electricity consumption, natural gas consumption and GHG emissions were all missing, so I dropped these null values as well. Also, the values for city classification were 0 for row 386, 393 and 395, which were inappropriate (i.e. incorrect variable type). In this case, I dropped these values. Then, the remaining data seemed appropriate and I got the clean data.

Table 1: The data summary table for GHG Emissions, Electricity Consumption, Natural Gas Consumption and Total Floor Area

	Min	Q1	Median	Mean	Q3	Max
GHG emissions	0	513.50	12902.5	115114.68	62398.75	18193619
Electricity consumption	0	10339.10	69819.4	1019225.22	388695.24	131780414
Natural gas consumption	0	0.00	0.0	36228.42	17785.47	6940806
Total floor area	0	55.25	5548.5	98694.49	26126.75	23073615

2.3 Exploratory Data Analysis

Firstly, I treated the GHG emissions as the response variable. I made a table to show the statistical summaries of all selected numerical variables. As shown in Table 1, GHG emissions had a mean of 115114.68 and an inter-quantile range (IQR = Q3 - Q1) of 61885.25.

Then, I considered the electricity consumption as an independent variable. It had an average of 1019225.22 and an inter-quantile range (IQR = Q3 - Q1) of 378356.14. I graphed a scatter plot of electricity consumption versus GHG emissions and added a best fit line to visually show their relationship. The pattern in Figure 1 showed that the sites with more amount of electricity consumed tended to have higher GHG emissions generally. Similarly, the natural gas consumption had an average of 36228.42, and as in Figure 2, the scatter plot of natural gas consumption versus GHG emissions illustrated that the sites with more natural gas consumption tended to have higher emissions. Moreover, the linear best fit line in these 2 scatter plots were seemed appropriate, which further indicated that electricity and natural gas consumptions were possible to be linearly correlated with GHG emissions.

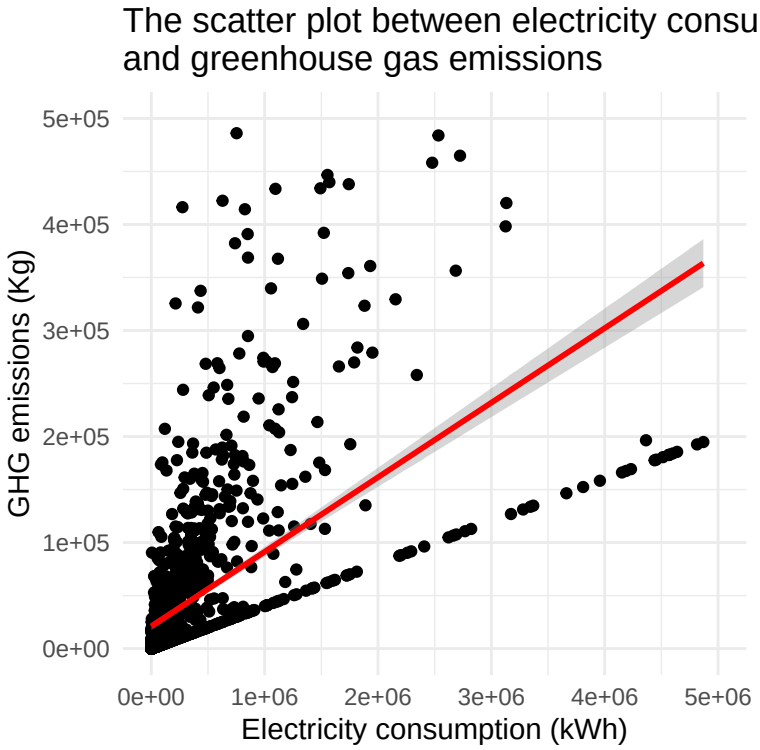

Figure 1: The scatter plot of electricity consumption versus GHG emissions of sites in the City of Toronto in 2018

Then, for the variable operation type, I plotted a bar chart of the average GHG emissions for each of its level. As shown in Figure 3, facilities related to the treatment of sewage, facilities related to the treatment of water

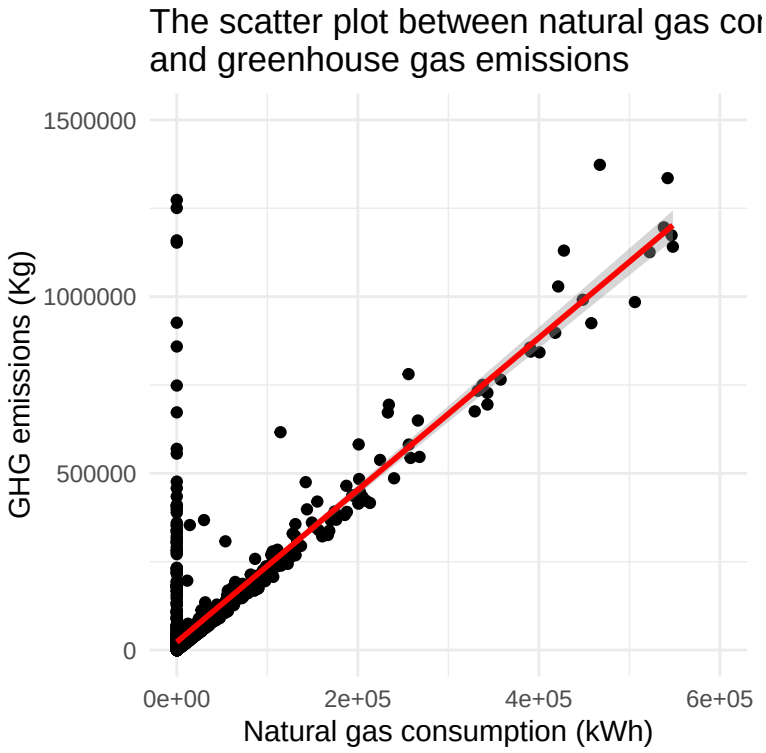

Figure 2: The scatter plot of natural gas consumption versus GHG emissions of sites in the City of Toronto in 2018

and long-term care were the 3 sites that had the highest average GHG emissions in 2018; meanwhile, Parking garages, facilities related to the pumping of sewage and fire stations and associated offices and facilities were the sites that had the lowest average annual GHG emissions. Similarly, for the city classification, I graphed another bar chart of the average GHG emissions for each of its group in Figure 4. This graph illustrated that Markham, Scarborough and Etobicoke were the 3 sites with the highest average GHG emissions; Mississauga, Tornhill and Pickering were the sites that had the lowest average emissions. These 2 bar charts showed that sites with different operation types and city classifications tended to have different GHG emissions, which further revealed that operation type and city classification were possible to affect GHG emissions.

3 Model

3.1 Model 1: The Multiple Linear Regression Model with Effects of Electricity Consumption and Natural Gas Consumption

To address the primary question, I used a multiple linear regression model including electricity consumption and natural gas consumption as the explanatory variables, because even though GHG emissions was a discrete variable, it was not the count. Also, the linear model was appropriate here. Meanwhile, since the sites with more floor area and longer average weekly operating hours tended to have higher GHG emissions generally, I considered the total floor area and average weekly operating hours as the confounding variables. And I included them as the covariates in the regression model to adjust for these differences. Then, I fitted this model and checked the p-values to see whether electricity consumption and natural gas consumption had significant effect on GHG emissions or not. Also, I checked the assumptions of normality and constant variance for errors.

3.2 Model 2: The Linear Regression Model with Effects of Other Relevant Factors

Then, motivated by the exploratory analysis findings (i.e. some sites with specific operation types and

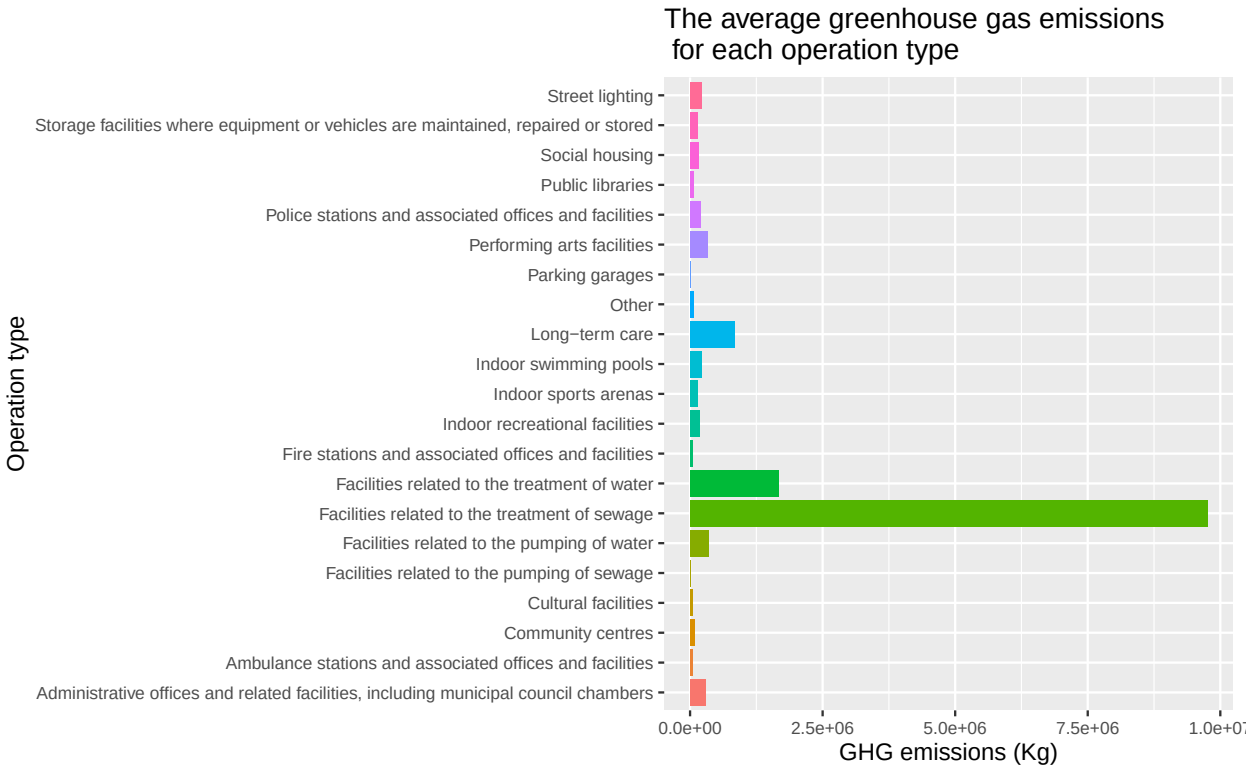

Figure 3: The bar chart of average GHG emissions for each operation type of sites in the City of Toronto in 2018

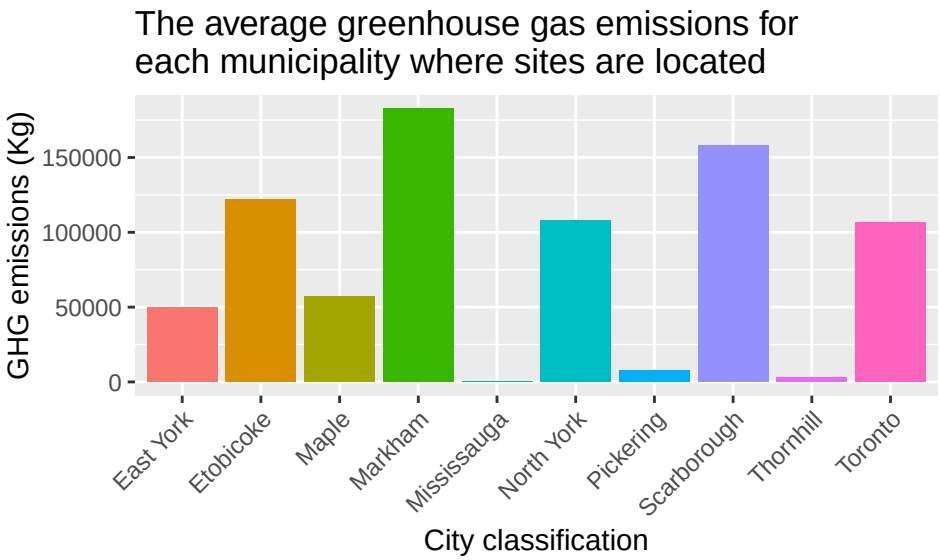

Figure 4: The bar chart of average GHG emissions for each city classification of sites in the City of Toronto in 2018

Table 2: The table of estimates for fitted Model 1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.759988e+02	3769.4336936	0.0997494	0.9205568
electricity	4.290020e-02	0.0006440	66.6172480	0.0000000
natural_gas	1.867660e+00	0.0122224	152.8058514	0.0000000
total_floor_area	5.225200e-03	0.0032880	1.5891717	0.1122364
avg_weekly_hrs168	2.227963e+02	5525.9281378	0.0403183	0.9678448
avg_weekly_hrs70	1.853735e+04	7543.3375283	2.4574462	0.0141075

Table 3: The table of estimates for each operation type

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
electricity	1	4.368963e+14	4.368963e+14	48209.101	0.000
natural_gas	1	2.181192e+14	2.181192e+14	24068.252	0.000
operation_type	20	6.572671e+11	3.286336e+10	3.626	0.000
city_class	9	3.304105e+10	3.671227e+09	0.405	0.933
total_floor_area	1	2.526801e+10	2.526801e+10	2.788	0.095
avg_weekly_hrs	2	2.414756e+09	1.207378e+09	0.133	0.875
Residuals	1443	1.307723e+13	9.062528e+09	NA	NA

city classifications tended to have higher or lower GHG emissions), I included the operation type and city classification as another 2 covariates into the fitted Model 1. I fitted this model and checked the p-values to see whether the effects of these factors on the GHG emissions were significant or not.

4 Results & Discussions

4.1 Model 1: The Multiple Linear Regression Model with Effects of Electricity Consumption and Natural Gas Consumption

For the fitted linear regression model, the p-values (Table 2) for electricity consumption and natural gas consumption were very small (<2e-16), which showed that they affected GHG emissions significantly. Also, the estimate for electricity consumption was 0.043, which revealed that GHG emissions would increase by 0.043 kg as one more unit electricity consumed. Similarly, the estimate for natural gas consumption was 1.867, which indicated that GHG emissions would be increase by 1.867 kg as one more unit natural gas consumed.

Furthermore, I checked that the errors were normal distributed and had constant variance, because most points were near the QQ line in QQ plot and there was no apparent pattern in the scatter plot errors.

4.2 Model 2: The Linear Regression Model with Effects of Other Relevant Factors

In the fitted model that included other relevant explanatory variables, the p-value (Table 3) for operation type was small (0.000). This indicated that the site's operation type significantly affected GHG emissions. Besides, Table 4 was the summary of estimated marginal means for each operation type. The site with facilities related to the treatment of water had an estimated marginal mean of -41807.21 (i.e. the lowest value), which showed that the sites with this operation type tended to have the lowest GHG emissions; while the estimated marginal mean for the performing arts facilities was 313303 (i.e. the highest value), which showed that the sites with this operation type tended to have higher GHG emissions. Moreover, the p-value for city classification was relatively large (0.933), which revealed that the city that site was located was not correlated with GHG emissions.

Table 4: The table of estimated marginal means for each operation type

operation_type	emmean
Administrative offices and related facilities, including municipal council chambers	218170.78
Ambulance stations and associated offices and facilities	111757.52
Community centres	114157.80
Cultural facilities	112653.37
Facilities related to the pumping of sewage	110572.36
Facilities related to the pumping of water	72600.59
Facilities related to the treatment of sewage	26558.07
Facilities related to the treatment of water	-41807.21
Fire stations and associated offices and facilities	112438.94
Indoor recreational facilities	115269.30
Indoor sports arenas	114922.14
Indoor swimming pools	115840.98
Long-term care	113274.39
Other	112885.90
Parking garages	107263.84
Performing arts facilities	318384.75
Police stations and associated offices and facilities	112584.01
Public libraries	119795.62
Social housing	110728.11
Storage facilities where equipment or vehicles are maintained, repaired or stored	123448.59
Street lighting	92145.72

5 Conclusions

Electricity consumption and natural gas consumption had significant effect on GHG emissions, specifically, the sites consuming more amount of electricity or natural gas tended to have higher emissions. Besides, GHG emissions also depended on some other attributes of sites. For example, the sites with some specific operation types tended to have much higher or much lower emissions than others.

Appendix

6.1 Model 1: The Multiple Linear Regression Model with Effects of Electricity Consumption and Natural Gas Consumption

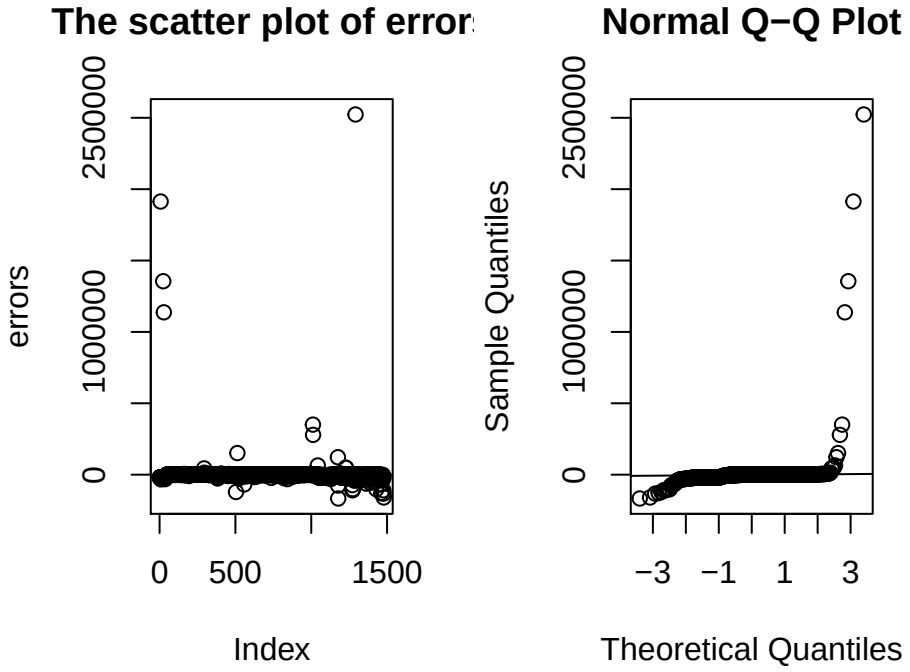

Figure 5: (#fig:model_ass)The scatter plot and the QQ plot of errors for Model 1 to check the normality and constant variance assumptions

References

Environment & Energy Program Administration, City of Toronto, Environment & Energy Division. 2018. "Annual Energy Consumption & Greenhouse Gas (Ghg) Emissions." https://www.toronto.ca/wp-content/uploads/2019/01/958c-2017-Annual-energy-consumption-and-GHG-emissions-report-compressed.pdf.

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

Lenth, Russell. 2019. *Emmeans: Estimated Marginal Means, Aka Least-Squares Means*. https://CRAN.R-project.org/package=emmeans.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain Fran<U+00E7>ois, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. doi:10.21105/joss.01686.