Prediction of Greenhouse Gas Emission

# Abstract

The goal of this project is to predict the quantity of greenhouse gases emitted based on a set of given or known parameters or variables. The theme is data mining and knowledge discovery.

In 2009 the Ontario Government enacted the Green Energy Act where government agencies were mandated to submit their annual energy consumption and greenhouse gas emission data together with their plans for energy conservation and demand **[1]**. These organizations included the broader public sector organizations such as hospitals, school boards, colleges, universities, municipal service boards and municipalities. The data submitted from 2011 to 2015 which is available as an open data set on the Ontario.ca website was used for this study.

The link to this data set is <https://www.ontario.ca/data/energy-use-and-greenhouse-gas-emissions-broader-public-sector> **[2]**.

# Introduction

The data used in this study has many attributes including the type of broader public sector, type of organization, type of operation, location, indoor area and quantity of various forms of energy used. The aim is to find out which of these attributes affect the amount of greenhouse gas produced. Also is it possible to predict how much greenhouse gas will be produced if we know the values of these parameters.

The 5 data files (one for each year) in excel format have 44 attributes and together consist of 88,600 rows (around 17,000 per year).

This data has been entered manually and there is a need to spend time in the initial analysis stage to get the data in a suitable state before proceeding to the next stages of the project. This involves creation of a data dictionary of the variables, correction of typing errors, incorrect spellings, and then initial analysis to treat missing values, outliers and other discrepancies. The cleaned data will be used in the exploratory analysis stage to determine the relationships between the variables. Variables would be reduced to the most important ones by dimensionality reduction using feature selection techniques. In the experimental design phase the data will be split into training and test sets and algorithms like regression analysis will be used to model the data to come up with a prediction formula, results and conclusion.

# Literature Review

Countries around the world are working towards reducing the emission of greenhouse gases (GHG) as it is shown that these gases contribute to global warming. There are many papers on this subject including what are GHG, how can GHG be measured, how are countries fairing in the reduction of GHG, how can we predict the quantity of GHG produced now and in the future. This is a review of some of the literature.

The atmospheric Infrared sounder AIRS is an instrument on the Earth Observing system spacecraft, on NASA’s Aqua satellite launched in 2002 **[3].** It is playing an important role in predicting weather patterns and GHG in the atmosphere. The infrared signals in the infrared spectra generated are used to determine the quantity and distribution of moisture and other gases like ozone, carbon dioxide, carbon monoxide, sulphur dioxide, methane and aerosols. Since clouds are found everywhere and can impede the accuracy of the results, the calculations are done by using a cloud clearing process and a cloud clearing algorithm. AIRS can take measurements for regions around the world. It can even predict the amount of sulfur dioxide emitted during a volcanic event.

Sahebalam et al describe a new method to calculate and forecast the amount of GHG produced using hourly electricity consumption data obtained from the IESO **[4].** They converted the electricity consumption to the carbon dioxide (CO2) equivalent. The general known method to calculate the CO2 equivalent is multiplying the electricity usage in gram by the CO2 intensity in gr/kwh. The values for CO2 intensity are found in the Government of Ontario ECCC standard. These authors modified the calculation by using a different calculation for the CO2 intensity. The Ontario grid is powered by nuclear, solar, wind, hydro, gas, biofuel. They took all these fuel types into account in their calculation. Their method of calculation is useful because it could predict energy consumption in any type of building.

The amount of GHG that would be produced by households and firms in Austin, Texas in 2030 is predicted using simulated data **[5]**. The calculations are based on energy consumption in 2030 taking into account ownership of vehicles, travel, migration of households and firms to the suburbs, expansion of freeways, taxes levied on gas and road tolls. The authors found that the least amount of GHG would be released if there is an urban growth boundary restricting migration to the suburbs and that the most GHG would be released if nothing changes and that life in 2030 would be the same as now. I think this is very hypothetical because new technologies could emerge by 2030 and there could be a huge shift in our way of life including travel, transportation, housing and urban migration.

A non-parametric, supervised algorithm, general regression neural network GRNN was used to predict GHG emitted by 26 countries in Europe based on their energy consumption from 2004 to 2012 **[6].** GHG were calculated as CO2 equivalents. The GHG intensity was calculated as the CO2equivalent divided by Gross Inland energy consumption GIEC. The GHG intensity was computed for all countries and was used in the model. Statistical regression models - linear regression and multiple polynomial regression were also used on the same data and it was found that GRNN had a higher accuracy.

The same authors used general regression neural network (GRNN) to study GHG produced in European countries where the emission sources were transportation, agriculture, industry, energy supply, land use and waste management **[7]**. They found that energy supply and use was the greatest contributor to GHG followed by transportation. They used correlation analysis and principal component analysis for dimensionality reduction to get their final dataset. Their results show that the difference between the actual and forecasted GHG for the different countries is less than 10% except for one outlier. The authors recommend artificial neural network model to predict GHG because there are no strictly defined parameters required in the input and therefore it can be applied in varying situations.

# Dataset

The raw data files for 2011 to 2015 in xls format were used. The number of records in the raw data files for 2011, 2012, 2013, 2014, and 2015 are 18743, 179778, 17010, 17224 and 17663 respectively.

The year attribute was added to identify the year of the data.

The attributes in the data files are shown below.

**Original attributes**

# column headers  
names(all\_data)

## [1] "Year"   
## [2] "Sector"   
## [3] "Organization"   
## [4] "Operation"   
## [5] "OperationType"   
## [6] "Address"   
## [7] "City"   
## [8] "PostalCode"   
## [9] "TotalIndoorSpace"   
## [10] "UnitofMeasure"   
## [11] "TotalIndoorSpace\_sqft"   
## [12] "WeeklyAverageHours"   
## [13] "AnnualFlow\_M"   
## [14] "NumberofPortables"   
## [15] "SwimmingPool"   
## [16] "Electricity\_kwh"   
## [17] "Electricity\_Unit"   
## [18] "NaturalGas\_Quantity"   
## [19] "NaturalGas\_Unit"   
## [20] "NaturalGas\_Cubicmeter"   
## [21] "FuelOil12\_L"   
## [22] "FuelOil12\_Unit"   
## [23] "FuelOil46\_L"   
## [24] "FuelOil46\_Unit"   
## [25] "Propane\_Litre"   
## [26] "Propane\_Unit"   
## [27] "Coal\_Quantity"   
## [28] "Coal\_Unit"   
## [29] "Wood\_Metrictonne"   
## [30] "Wood\_Unit"   
## [31] "DistrictHeating\_Quantity"   
## [32] "DistrictHeating\_Unit"   
## [33] "DistrictHeating\_GJ"   
## [34] "DistrictHeating\_IsRenewable"   
## [35] "DistrictHeating\_RenewableEmissionFactor"  
## [36] "DistrictCooling"   
## [37] "DistrictCooling\_Unit"   
## [38] "DistrictCooling\_GJ"   
## [39] "DistrictCooling\_IsRenewable"   
## [40] "DistrictCooling\_RenewableEmissionFactor"  
## [41] "GHGEmissions\_KG"   
## [42] "EnergyIntensityekWh\_sqft"   
## [43] "EnergyIntensity\_ekWh\_mega\_litre"   
## [44] "EnergyIntensity\_GJ\_m2"   
## [45] "EnergyIntensityGJ\_mega\_litre"

There is a wide spread in the values of the GHG emissions from close to 0 to 75277858 KG. It is the same with the other attributes. Therefore the data was normalized in the data preparation stage.

The sectors are Hospital, Municipality, Post-Secondary Institution and School Board. The Unit columns show the measure used for the previous attribute.

Energy Intensity is in GJ/Mega Litres or ekWh/Mega Litres for operations of type water works or sewage works. In other cases, Energy Intensity is given in ekWh/sqft and GJ/m2. GHG emissions are in Kg. Area is given in square feet and square meters.

There are many cells with missing data as shown below. They were not applicable in many instances and therefore 0 was substituted later for these empty attributes.

# How many missing data in columns  
sapply(all\_data, function(x) sum(is.na(x)))

## Year Sector   
## 0 0   
## Organization Operation   
## 0 0   
## OperationType Address   
## 0 17   
## City PostalCode   
## 0 17   
## TotalIndoorSpace\_sqft WeeklyAverageHours   
## 3054 4379   
## AnnualFlow\_M NumberofPortables   
## 17405 16481   
## SwimmingPool Electricity\_kwh   
## 105 254   
## NaturalGas\_Cubicmeter FuelOil12\_L   
## 16760 13887   
## FuelOil46\_L Propane\_Litre   
## 14599 14027   
## Coal\_Quantity Wood\_Metrictonne   
## 14665 14658   
## DistrictHeating\_GJ DistrictHeating\_IsRenewable   
## 65549 105   
## DistrictCooling\_GJ DistrictCooling\_IsRenewable   
## 65836 105   
## GHGEmissions\_KG EnergyIntensityekWh\_sqft   
## 0 3270   
## EnergyIntensity\_ekWh\_mega\_litre   
## 15624

This is the statistical information for the combined raw data.

summary(all\_data)

## Year Sector   
## 2011:18730 Municipal :56318   
## 2012:17978 Post-Secondary Educational Institution: 4608   
## 2013:17009 Public Hospital : 2272   
## 2014:17092 School Board :25061   
## 2015:17450   
##   
##   
## Organization Operation   
## City of Toronto : 4287 Municipal Office : 334   
## Toronto District School Board : 2866 Water Treatment Plant: 271   
## City of Greater Sudbury : 1455 Fire Hall : 261   
## Peel District School Board : 1280 Public Works Garage : 203   
## City of Ottawa : 1245 Town Hall : 177   
## York Region District School Board: 1098 Library : 162   
## (Other) :76028 (Other) :86851   
## OperationType Address   
## School :23867 Intentionally Omitted : 465   
## Administrative office: 9950 public works related facility: 368   
## Storage : 6772 4700 Keele Street : 330   
## Fire station : 5917 1385 Woodroffe Avenue : 271   
## Community centres : 5711 1001 Fanshawe College Blvd. : 188   
## Sewage pumping : 4946 (Other) :86620   
## (Other) :31096 NA's : 17   
## City PostalCode TotalIndoorSpace\_sqft  
## Toronto : 8751 M3J 1P3: 269 Min. :0.000e+00   
## Ottawa : 2878 N5Y 5R6: 252 1st Qu.:1.768e+03   
## Brampton : 2097 L7B 1B3: 225 Median :9.030e+03   
## Mississauga: 1964 L6P 1K7: 213 Mean :8.532e+04   
## London : 1434 K2G 1V8: 203 3rd Qu.:3.807e+04   
## Hamilton : 1421 (Other):87080 Max. :4.262e+09   
## (Other) :69714 NA's : 17   
## WeeklyAverageHours AnnualFlow\_M NumberofPortables SwimmingPool  
## Min. : 0.00 Min. :0.000e+00 Min. : 0.0000 0:88067   
## 1st Qu.: 40.00 1st Qu.:0.000e+00 1st Qu.: 0.0000 1: 192   
## Median : 60.00 Median :0.000e+00 Median : 0.0000   
## Mean : 82.39 Mean :7.220e+04 Mean : 0.4598   
## 3rd Qu.: 126.00 3rd Qu.:0.000e+00 3rd Qu.: 0.0000   
## Max. :8760.00 Max. :1.171e+09 Max. :33.0000   
##   
## Electricity\_kwh NaturalGas\_Cubicmeter FuelOil12\_L   
## Min. :0.000e+00 Min. :-5.740e+02 Min. :0.000e+00   
## 1st Qu.:2.273e+04 1st Qu.: 0.000e+00 1st Qu.:0.000e+00   
## Median :1.082e+05 Median : 9.492e+03 Median :0.000e+00   
## Mean :4.375e+06 Mean : 7.418e+05 Mean :3.794e+06   
## 3rd Qu.:3.421e+05 3rd Qu.: 4.495e+04 3rd Qu.:0.000e+00   
## Max. :7.709e+10 Max. : 1.131e+10 Max. :7.968e+10   
##   
## FuelOil46\_L Propane\_Litre Coal\_Quantity   
## Min. :0.000e+00 Min. : 0.0 Min. : 0.00   
## 1st Qu.:0.000e+00 1st Qu.: 0.0 1st Qu.: 0.00   
## Median :0.000e+00 Median : 0.0 Median : 0.00   
## Mean :6.779e+05 Mean : 716.6 Mean : 1.98   
## 3rd Qu.:0.000e+00 3rd Qu.: 0.0 3rd Qu.: 0.00   
## Max. :1.163e+10 Max. :1149351.0 Max. :33559.88   
##   
## Wood\_Metrictonne DistrictHeating\_GJ DistrictHeating\_IsRenewable  
## Min. : 0.00 Min. : 0 0:88238   
## 1st Qu.: 0.00 1st Qu.: 0 1: 21   
## Median : 0.00 Median : 0   
## Mean : 3.23 Mean : 161   
## 3rd Qu.: 0.00 3rd Qu.: 0   
## Max. :37347.42 Max. :1463421   
##   
## DistrictCooling\_GJ DistrictCooling\_IsRenewable GHGEmissions\_KG   
## Min. : 0.00 0:88243 Min. :0.000e+00   
## 1st Qu.: 0.00 1: 16 1st Qu.:6.751e+03   
## Median : 0.00 Median :3.511e+04   
## Mean : 21.39 Mean :1.433e+07   
## 3rd Qu.: 0.00 3rd Qu.:1.141e+05   
## Max. :147285.75 Max. :2.800e+11   
##   
## EnergyIntensityekWh\_sqft EnergyIntensity\_ekWh\_mega\_litre  
## Min. :0.000e+00 Min. :0.000e+00   
## 1st Qu.:1.300e+01 1st Qu.:0.000e+00   
## Median :2.100e+01 Median :0.000e+00   
## Mean :3.281e+06 Mean :1.457e+14   
## 3rd Qu.:3.300e+01 3rd Qu.:0.000e+00   
## Max. :2.860e+11 Max. :8.430e+18   
##

# Approach

1. **Prepare and Analyze the data**

Initial analysis

* Create Data dictionary
* Handle incorrect data, missing data, duplicate rows, outliers, discrepancies
* Remove unnecessary attributes and use data imputation if appropriate
* What is the correlation between variables?
* Can any attribute be converted to a categorical variable?
* Dimensionality reduction - preliminary feature selection using decision tree, correlation analysis, principal component analysis.
* Use R programming for correlation analysis, correlation matrix, regression models, scatterplots, boxplots, histograms
* The dependent variable is the quantity of GHG

Exploratory Analysis

* Normalize the data
* Do feature selection using forward elimination
* Do visualization of feature relationships using R
* Determine whether it is better to use a subset of the data
* Do data visualization using a reporting tool like Oracle business intelligence or R
* What are the interesting conclusions or stories that can be drawn from the data
* Split the data into training and testing sets

1. **Design the experiments**

* My research question - forecasting the GHG produced by various attributes
* Algorithm – multivariate regression (generalized linear model glm, also compare with simple linear model lm)

1. **Build, Test, Evaluate Model**

* The response variable quantity of GHG is numerical
* Train the model using the training data to get a good prediction of GHG
* Test the model using the test data
* Evaluate the model
* Modify the model if necessary

1. **Tune the parameters**

If accuracy is too high (overfitting) use new attributes/variables. If accuracy is low use new attributes/variables and modify training data/testing data split

Repeat Steps 1, 2, 3 as necessary – it is an iterative process – till meaningful results for prediction of GHG are obtained.

1. **Prediction of GHG**

* What is the difference between the forecasted GHG using my model and the actual GHG

1. **Conclusion**

* Can I automate this prediction of GHG?
* What is the risk and validity of my model?

**Figure 1: High level Approach**

Iterative process till the best prediction of GHG is obtained

**Figure 2: More detailed approach**

## My GITHUB LINK

<https://github.com/ermafernando/GHG.git>

## Step 1: From Raw data to technically correct data

Added a year column and filled this with the year of the file (2011, 2012, 2013, 2014, 2015).

The title of the columns and order of the columns were not consistent across the years of data. Units of measure were not consistent, there were typing errors and also French versions. Some attributes were given in different units of measure and had to be converted to a uniform unit of measure. For example Total Floor Area was converted to square feet, Natural Gas was Converted to Cubic meter, Cooling and Heating were converted to Giga joules.

Inspection of the data files and attributes showed that the majority of the values in Renewable 1 have value “FALSE” and therefore this column was eliminated.

There are missing values in many columns which could be due to the fact that it is not applicable for that operation type. In these cases it is better to substitute 0 for missing values. The address, city and postal code are related in many cases. However these attributes have dummy data like “various locations”, “Intentionally Omitted” - especially when type of operation is sewage treatment, street lighting. Postal code is also not reliable because it has values like “X1X 1X1”, “unknown”.

Removed units of measure columns and columns that were very sparsely populated to end up with 27 attributes.

## [1] "Year" "Sector"   
## [3] "Organization" "Operation"   
## [5] "OperationType" "Address"   
## [7] "City" "PostalCode"   
## [9] "TotalIndoorSpace\_sqft" "WeeklyAverageHours"   
## [11] "AnnualFlow\_M" "NumberofPortables"   
## [13] "SwimmingPool" "Electricity\_kwh"   
## [15] "NaturalGas\_Cubicmeter" "FuelOil12\_L"   
## [17] "FuelOil46\_L" "Propane\_Litre"   
## [19] "Coal\_Quantity" "Wood\_Metrictonne"   
## [21] "DistrictHeating\_GJ" "DistrictHeating\_IsRenewable"   
## [23] "DistrictCooling\_GJ" "DistrictCooling\_IsRenewable"   
## [25] "GHGEmissions\_KG" "EnergyIntensityekWh\_sqft"   
## [27] "EnergyIntensity\_ekWh\_mega\_litre"

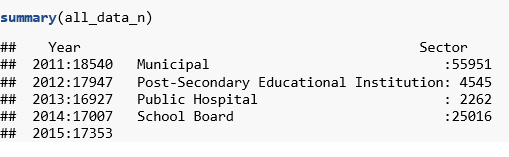
## Step 2: From technically correct data to consistent data

Rearranged columns and created column headers across the 5 files.

Deleted rows which had very low greenhouse gas emission - values less than 1 kg of emission.

Replaced missing values for numerical values with 0 because if it is not applicable it is like saying that the value is 0.

Yes/No attributes like DistrictCooling\_IsRenewable, DistrictHeating\_IsRenewable, Swiming Pool were converted to 0, 1 and converted to factor. Year was converted to factor. Created a new numerical column for type of Sector and type of operation and made it a factor column. Normalized the data (all\_data\_n).



## Step 3: Exploratory data analysis

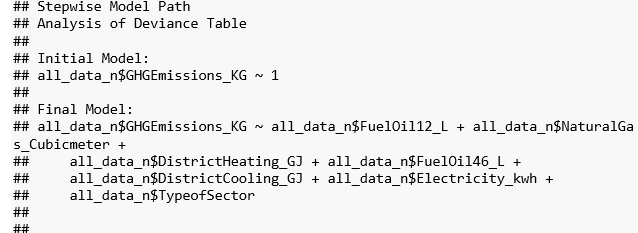
Normalized the data as there was a very big spread between the values for the various attributes. Plots, box plots and correlations are shown in Appendix 1, 2, 3. These were done using the normalized data.

## Step 4: Split into training and testing sets

The normalized data was split in 70:30 ratio into training and testing sets.

## Step 5: Feature selection

See Appendix 6a and 6b. Principal component analysis (Appendix 6a) did not reduce the number of attributes. Feature selection using forward elimination (Appendix 6b) was successful and showed that the important attributes are FuelOil12\_L, NaturalGas\_Cubicmeter, DistrictHeating\_GJ, FuelOil46\_L, DistrictCooling\_GJ, Electricity\_kwh, TypeofSector.



## Step 6: Application of Regression models (Build models for prediction)

# Multivariate Linear Regression model

1. **Simple linear regression (lm) including Type of Sector**

###### Simple multivariate linear regression model  
# build model using train.set  
###### using lm  
set.seed(111)  
model\_mlr1 <- lm(train.set$GHGEmissions\_KG ~ train.set$FuelOil12\_L + train.set$NaturalGas\_Cubicmeter + train.set$DistrictHeating\_GJ + train.set$FuelOil46\_L + train.set$DistrictCooling\_GJ + train.set$Electricity\_kwh + train.set$TypeofSector, data=train.set)   
   
summary(model\_mlr1)

##   
## Call:  
## lm(formula = train.set$GHGEmissions\_KG ~ train.set$FuelOil12\_L +   
## train.set$NaturalGas\_Cubicmeter + train.set$DistrictHeating\_GJ +   
## train.set$FuelOil46\_L + train.set$DistrictCooling\_GJ + train.set$Electricity\_kwh +   
## train.set$TypeofSector, data = train.set)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.0054583 -0.0000002 0.0000002 0.0000002 0.0090358   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.942e-06 9.493e-07 -6.260 3.88e-10 \*\*\*  
## train.set$FuelOil12\_L 7.444e-01 2.247e-03 331.235 < 2e-16 \*\*\*  
## train.set$NaturalGas\_Cubicmeter 2.998e-02 3.004e-03 9.982 < 2e-16 \*\*\*  
## train.set$DistrictHeating\_GJ 1.042e-02 3.882e-05 268.434 < 2e-16 \*\*\*  
## train.set$FuelOil46\_L 1.776e-01 3.018e-03 58.847 < 2e-16 \*\*\*  
## train.set$DistrictCooling\_GJ -3.209e-03 4.385e-05 -73.190 < 2e-16 \*\*\*  
## train.set$Electricity\_kwh 5.502e-02 2.235e-03 24.616 < 2e-16 \*\*\*  
## train.set$TypeofSector2 3.435e-06 1.651e-06 2.081 0.0375 \*   
## train.set$TypeofSector3 5.735e-06 9.866e-07 5.812 6.19e-09 \*\*\*  
## train.set$TypeofSector4 6.272e-06 1.031e-06 6.084 1.18e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.33e-05 on 61431 degrees of freedom  
## Multiple R-squared: 0.9999, Adjusted R-squared: 0.9999   
## F-statistic: 6.722e+07 on 9 and 61431 DF, p-value: < 2.2e-16

coefficients(model\_mlr1)

## (Intercept) train.set$FuelOil12\_L   
## -5.942238e-06 7.443641e-01   
## train.set$NaturalGas\_Cubicmeter train.set$DistrictHeating\_GJ   
## 2.998127e-02 1.041964e-02   
## train.set$FuelOil46\_L train.set$DistrictCooling\_GJ   
## 1.776194e-01 -3.209362e-03   
## train.set$Electricity\_kwh train.set$TypeofSector2   
## 5.501600e-02 3.435260e-06   
## train.set$TypeofSector3 train.set$TypeofSector4   
## 5.734743e-06 6.272004e-06

aov(model\_mlr1)

## Call:  
## aov(formula = model\_mlr1)  
##   
## Terms:  
## train.set$FuelOil12\_L train.set$NaturalGas\_Cubicmeter  
## Sum of Squares 1.7078038 0.0107075  
## Deg. of Freedom 1 1  
## train.set$DistrictHeating\_GJ train.set$FuelOil46\_L  
## Sum of Squares 0.0001846 0.0001188  
## Deg. of Freedom 1 1  
## train.set$DistrictCooling\_GJ train.set$Electricity\_kwh  
## Sum of Squares 0.0000154 0.0000017  
## Deg. of Freedom 1 1  
## train.set$TypeofSector Residuals  
## Sum of Squares 0.0000001 0.0001745  
## Deg. of Freedom 3 61431  
##   
## Residual standard error: 5.330338e-05  
## Estimated effects may be unbalanced

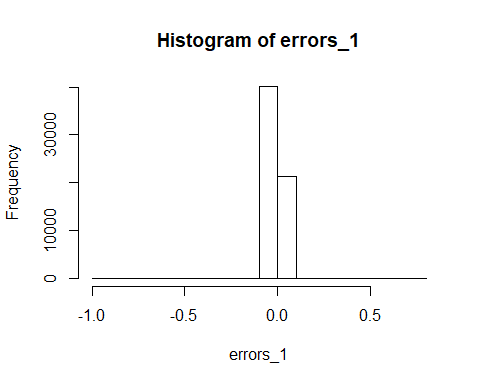
prediction\_1 <- predict(model\_mlr1, interval="prediction", newdata =test.set)

## Warning: 'newdata' had 26333 rows but variables found have 61441 rows

# see errors and plot on histogram  
errors\_1<- prediction\_1[,"fit"] - test.set$GHGEmissions\_KG

## Warning in prediction\_1[, "fit"] - test.set$GHGEmissions\_KG: longer object  
## length is not a multiple of shorter object length

hist(errors\_1)



# Compute the root mean square error and find the percentage of cases with less than 25% error.  
  
rmse\_1 <- sqrt(sum((prediction\_1[,"fit"] - all\_data\_n$GHGEmissions\_KG)^2)/nrow(test.set))

## Warning in prediction\_1[, "fit"] - all\_data\_n$GHGEmissions\_KG: longer  
## object length is not a multiple of shorter object length

rel\_change\_1 <- 1 - ((test.set$GHGEmissions\_KG - abs(errors\_1)) / test.set$GHGEmissions\_KG)

## Warning in test.set$GHGEmissions\_KG - abs(errors\_1): longer object length  
## is not a multiple of shorter object length

## Warning in (test.set$GHGEmissions\_KG - abs(errors\_1))/test.set  
## $GHGEmissions\_KG: longer object length is not a multiple of shorter object  
## length

pred25\_1 <- table(rel\_change\_1<0.25)["TRUE"] / nrow(test.set)   
paste("RMSE\_1:", rmse\_1)

## [1] "RMSE\_1: 0.0136331380956868"

paste("PRED(25):", pred25\_1)

## [1] "PRED(25): 0.0786465651463943"

1. **Simple linear regression (lm) excluding Type of Sector**

###### using lm - dropped variable "TypeofSector"  
set.seed(222)  
model\_mlr2 <- lm(train.set$GHGEmissions\_KG ~ train.set$FuelOil12\_L + train.set$NaturalGas\_Cubicmeter + train.set$DistrictHeating\_GJ + train.set$FuelOil46\_L + train.set$DistrictCooling\_GJ + train.set$Electricity\_kwh, data=train.set)   
   
summary(model\_mlr2)

##   
## Call:  
## lm(formula = train.set$GHGEmissions\_KG ~ train.set$FuelOil12\_L +   
## train.set$NaturalGas\_Cubicmeter + train.set$DistrictHeating\_GJ +   
## train.set$FuelOil46\_L + train.set$DistrictCooling\_GJ + train.set$Electricity\_kwh,   
## data = train.set)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.0054537 0.0000004 0.0000004 0.0000005 0.0090417   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.054e-07 2.153e-07 -1.883 0.0596 .   
## train.set$FuelOil12\_L 7.442e-01 2.246e-03 331.377 <2e-16 \*\*\*  
## train.set$NaturalGas\_Cubicmeter 2.968e-02 3.001e-03 9.890 <2e-16 \*\*\*  
## train.set$DistrictHeating\_GJ 1.041e-02 3.878e-05 268.455 <2e-16 \*\*\*  
## train.set$FuelOil46\_L 1.779e-01 3.016e-03 58.994 <2e-16 \*\*\*  
## train.set$DistrictCooling\_GJ -3.224e-03 4.379e-05 -73.623 <2e-16 \*\*\*  
## train.set$Electricity\_kwh 5.522e-02 2.233e-03 24.724 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.332e-05 on 61434 degrees of freedom  
## Multiple R-squared: 0.9999, Adjusted R-squared: 0.9999   
## F-statistic: 1.008e+08 on 6 and 61434 DF, p-value: < 2.2e-16

coefficients(model\_mlr2)

## (Intercept) train.set$FuelOil12\_L   
## -4.054311e-07 7.441562e-01   
## train.set$NaturalGas\_Cubicmeter train.set$DistrictHeating\_GJ   
## 2.968370e-02 1.041156e-02   
## train.set$FuelOil46\_L train.set$DistrictCooling\_GJ   
## 1.779203e-01 -3.223940e-03   
## train.set$Electricity\_kwh   
## 5.522074e-02

aov(model\_mlr2)

## Call:  
## aov(formula = model\_mlr2)  
##   
## Terms:  
## train.set$FuelOil12\_L train.set$NaturalGas\_Cubicmeter  
## Sum of Squares 1.7078038 0.0107075  
## Deg. of Freedom 1 1  
## train.set$DistrictHeating\_GJ train.set$FuelOil46\_L  
## Sum of Squares 0.0001846 0.0001188  
## Deg. of Freedom 1 1  
## train.set$DistrictCooling\_GJ train.set$Electricity\_kwh  
## Sum of Squares 0.0000154 0.0000017  
## Deg. of Freedom 1 1  
## Residuals  
## Sum of Squares 0.0001747  
## Deg. of Freedom 61434  
##   
## Residual standard error: 5.331956e-05  
## Estimated effects may be unbalanced

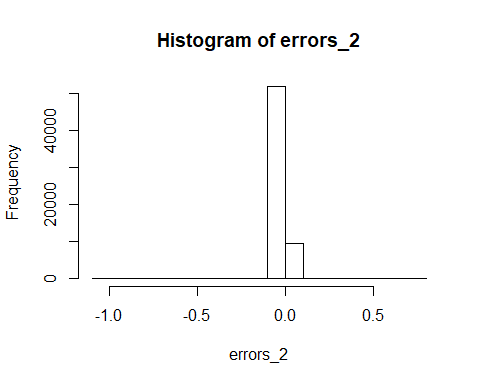
prediction\_2 <- predict(model\_mlr2, interval="prediction", newdata =test.set)

## Warning: 'newdata' had 26333 rows but variables found have 61441 rows

# see errors and plot on histogram  
errors\_2<- prediction\_2[,"fit"] - test.set$GHGEmissions\_KG

## Warning in prediction\_2[, "fit"] - test.set$GHGEmissions\_KG: longer object  
## length is not a multiple of shorter object length

hist(errors\_2)



# Compute the root mean square error and find the percentage of cases with less than 25% error.  
  
rmse\_2 <- sqrt(sum((prediction\_2[,"fit"] - all\_data\_n$GHGEmissions\_KG)^2)/nrow(test.set))

## Warning in prediction\_2[, "fit"] - all\_data\_n$GHGEmissions\_KG: longer  
## object length is not a multiple of shorter object length

rel\_change\_2 <- 1 - ((test.set$GHGEmissions\_KG - abs(errors\_2)) / test.set$GHGEmissions\_KG)

## Warning in test.set$GHGEmissions\_KG - abs(errors\_2): longer object length  
## is not a multiple of shorter object length

## Warning in (test.set$GHGEmissions\_KG - abs(errors\_2))/test.set  
## $GHGEmissions\_KG: longer object length is not a multiple of shorter object  
## length

pred25\_2 <- table(rel\_change\_2<0.25)["TRUE"] / nrow(test.set)   
paste("RMSE\_2:", rmse\_2)

## [1] "RMSE\_2: 0.0136331387825749"

paste("PRED(25):", pred25\_2)

## [1] "PRED(25): 0.0364941328371245"

1. **Generalized linear regression (glm) excluding Type of Sector**

###### General multivariate linear regression model  
###### using glm - dropped variable "TypeofSector"  
set.seed(333)  
model\_mlr3<-glm(train.set$GHGEmissions\_KG ~ train.set$FuelOil12\_L + train.set$NaturalGas\_Cubicmeter + train.set$DistrictHeating\_GJ + train.set$FuelOil46\_L + train.set$DistrictCooling\_GJ + train.set$Electricity\_kwh, data=train.set)  
  
summary(model\_mlr3)

##   
## Call:  
## glm(formula = train.set$GHGEmissions\_KG ~ train.set$FuelOil12\_L +   
## train.set$NaturalGas\_Cubicmeter + train.set$DistrictHeating\_GJ +   
## train.set$FuelOil46\_L + train.set$DistrictCooling\_GJ + train.set$Electricity\_kwh,   
## data = train.set)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.0054537 0.0000004 0.0000004 0.0000005 0.0090417   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.054e-07 2.153e-07 -1.883 0.0596 .   
## train.set$FuelOil12\_L 7.442e-01 2.246e-03 331.377 <2e-16 \*\*\*  
## train.set$NaturalGas\_Cubicmeter 2.968e-02 3.001e-03 9.890 <2e-16 \*\*\*  
## train.set$DistrictHeating\_GJ 1.041e-02 3.878e-05 268.455 <2e-16 \*\*\*  
## train.set$FuelOil46\_L 1.779e-01 3.016e-03 58.994 <2e-16 \*\*\*  
## train.set$DistrictCooling\_GJ -3.224e-03 4.379e-05 -73.623 <2e-16 \*\*\*  
## train.set$Electricity\_kwh 5.522e-02 2.233e-03 24.724 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 2.842975e-09)  
##   
## Null deviance: 1.71900651 on 61440 degrees of freedom  
## Residual deviance: 0.00017466 on 61434 degrees of freedom  
## AIC: -1034690  
##   
## Number of Fisher Scoring iterations: 2

coefficients(model\_mlr3)

## (Intercept) train.set$FuelOil12\_L   
## -4.054311e-07 7.441562e-01   
## train.set$NaturalGas\_Cubicmeter train.set$DistrictHeating\_GJ   
## 2.968370e-02 1.041156e-02   
## train.set$FuelOil46\_L train.set$DistrictCooling\_GJ   
## 1.779203e-01 -3.223940e-03   
## train.set$Electricity\_kwh   
## 5.522074e-02

aov(model\_mlr3)

## Call:  
## aov(formula = model\_mlr3)  
##   
## Terms:  
## train.set$FuelOil12\_L train.set$NaturalGas\_Cubicmeter  
## Sum of Squares 1.7078038 0.0107075  
## Deg. of Freedom 1 1  
## train.set$DistrictHeating\_GJ train.set$FuelOil46\_L  
## Sum of Squares 0.0001846 0.0001188  
## Deg. of Freedom 1 1  
## train.set$DistrictCooling\_GJ train.set$Electricity\_kwh  
## Sum of Squares 0.0000154 0.0000017  
## Deg. of Freedom 1 1  
## Residuals  
## Sum of Squares 0.0001747  
## Deg. of Freedom 61434  
##   
## Residual standard error: 5.331956e-05  
## Estimated effects may be unbalanced

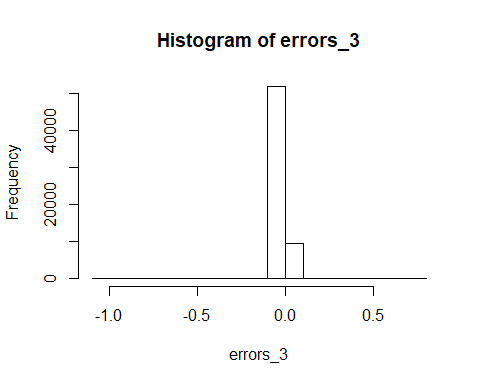
#aov gives intercept and coefficient for each variable  
# From the result, formula to calculate GHG is  
  
  
prediction\_3 <- predict(model\_mlr3, type="response", se.fit=FALSE, newdata =test.set)

## Warning: 'newdata' had 26333 rows but variables found have 61441 rows

# see errors and plot on histogram  
errors\_3<- prediction\_3 - test.set$GHGEmissions\_KG

## Warning in prediction\_3 - test.set$GHGEmissions\_KG: longer object length is  
## not a multiple of shorter object length

hist(errors\_3)



# Compute the root mean square error and find the percentage of cases with less than 25% error.  
  
rmse\_3 <- sqrt(sum((prediction\_3 - all\_data\_n$GHGEmissions\_KG)^2)/nrow(test.set))

## Warning in prediction\_3 - all\_data\_n$GHGEmissions\_KG: longer object length  
## is not a multiple of shorter object length

rel\_change\_3 <- 1 - ((test.set$GHGEmissions\_KG - abs(errors\_3)) / test.set$GHGEmissions\_KG)

## Warning in test.set$GHGEmissions\_KG - abs(errors\_3): longer object length  
## is not a multiple of shorter object length

## Warning in (test.set$GHGEmissions\_KG - abs(errors\_3))/test.set  
## $GHGEmissions\_KG: longer object length is not a multiple of shorter object  
## length

pred25\_3 <- table(rel\_change\_3<0.25)["TRUE"] / nrow(test.set)   
paste("RMSE\_3:", rmse\_3)

## [1] "RMSE\_3: 0.0136331387825749"

paste("PRED(25):", pred25\_3)

## [1] "PRED(25): 0.0364941328371245"

1. **Compare the 3 models**
2. **Anova (for 3 models)**

aov(model\_mlr1)

## Call:  
## aov(formula = model\_mlr1)  
##   
## Terms:  
## train.set$FuelOil12\_L train.set$NaturalGas\_Cubicmeter  
## Sum of Squares 1.7078038 0.0107075  
## Deg. of Freedom 1 1  
## train.set$DistrictHeating\_GJ train.set$FuelOil46\_L  
## Sum of Squares 0.0001846 0.0001188  
## Deg. of Freedom 1 1  
## train.set$DistrictCooling\_GJ train.set$Electricity\_kwh  
## Sum of Squares 0.0000154 0.0000017  
## Deg. of Freedom 1 1  
## train.set$TypeofSector Residuals  
## Sum of Squares 0.0000001 0.0001745  
## Deg. of Freedom 3 61431  
##   
## Residual standard error: 5.330338e-05  
## Estimated effects may be unbalanced

aov(model\_mlr2)

## Call:  
## aov(formula = model\_mlr2)  
##   
## Terms:  
## train.set$FuelOil12\_L train.set$NaturalGas\_Cubicmeter  
## Sum of Squares 1.7078038 0.0107075  
## Deg. of Freedom 1 1  
## train.set$DistrictHeating\_GJ train.set$FuelOil46\_L  
## Sum of Squares 0.0001846 0.0001188  
## Deg. of Freedom 1 1  
## train.set$DistrictCooling\_GJ train.set$Electricity\_kwh  
## Sum of Squares 0.0000154 0.0000017  
## Deg. of Freedom 1 1  
## Residuals  
## Sum of Squares 0.0001747  
## Deg. of Freedom 61434  
##   
## Residual standard error: 5.331956e-05  
## Estimated effects may be unbalanced

aov(model\_mlr3)

## Call:  
## aov(formula = model\_mlr3)  
##   
## Terms:  
## train.set$FuelOil12\_L train.set$NaturalGas\_Cubicmeter  
## Sum of Squares 1.7078038 0.0107075  
## Deg. of Freedom 1 1  
## train.set$DistrictHeating\_GJ train.set$FuelOil46\_L  
## Sum of Squares 0.0001846 0.0001188  
## Deg. of Freedom 1 1  
## train.set$DistrictCooling\_GJ train.set$Electricity\_kwh  
## Sum of Squares 0.0000154 0.0000017  
## Deg. of Freedom 1 1  
## Residuals  
## Sum of Squares 0.0001747  
## Deg. of Freedom 61434  
##   
## Residual standard error: 5.331956e-05  
## Estimated effects may be unbalanced

1. **Root mean square error (for 3 models)**

paste("RMSE\_1:", rmse\_1)

## [1] "RMSE\_1: 0.0136331380956868"

paste("RMSE\_2:", rmse\_2)

## [1] "RMSE\_2: 0.0136331387825749"

paste("RMSE\_3:", rmse\_3)

## [1] "RMSE\_3: 0.0136331387825749"

**© Coefficients (for 3 models)**

coefficients(model\_mlr1)

## (Intercept) train.set$FuelOil12\_L   
## -5.942238e-06 7.443641e-01   
## train.set$NaturalGas\_Cubicmeter train.set$DistrictHeating\_GJ   
## 2.998127e-02 1.041964e-02   
## train.set$FuelOil46\_L train.set$DistrictCooling\_GJ   
## 1.776194e-01 -3.209362e-03   
## train.set$Electricity\_kwh train.set$TypeofSector2   
## 5.501600e-02 3.435260e-06   
## train.set$TypeofSector3 train.set$TypeofSector4   
## 5.734743e-06 6.272004e-06

coefficients(model\_mlr2)

## (Intercept) train.set$FuelOil12\_L   
## -4.054311e-07 7.441562e-01   
## train.set$NaturalGas\_Cubicmeter train.set$DistrictHeating\_GJ   
## 2.968370e-02 1.041156e-02   
## train.set$FuelOil46\_L train.set$DistrictCooling\_GJ   
## 1.779203e-01 -3.223940e-03   
## train.set$Electricity\_kwh   
## 5.522074e-02

coefficients(model\_mlr3)

## (Intercept) train.set$FuelOil12\_L   
## -4.054311e-07 7.441562e-01   
## train.set$NaturalGas\_Cubicmeter train.set$DistrictHeating\_GJ   
## 2.968370e-02 1.041156e-02   
## train.set$FuelOil46\_L train.set$DistrictCooling\_GJ   
## 1.779203e-01 -3.223940e-03   
## train.set$Electricity\_kwh   
## 5.522074e-02

**(d) Percentage of cases with less than 25% error (for 3 models)**

paste("PRED(25):", pred25\_1)

## [1] "PRED(25): 0.0786465651463943"

paste("PRED(25):", pred25\_2)

## [1] "PRED(25): 0.0364941328371245"

paste("PRED(25):", pred25\_3)

## [1] "PRED(25): 0.0364941328371245

**(e) Prediction/calculation of GHG (using coefficients from glm)**

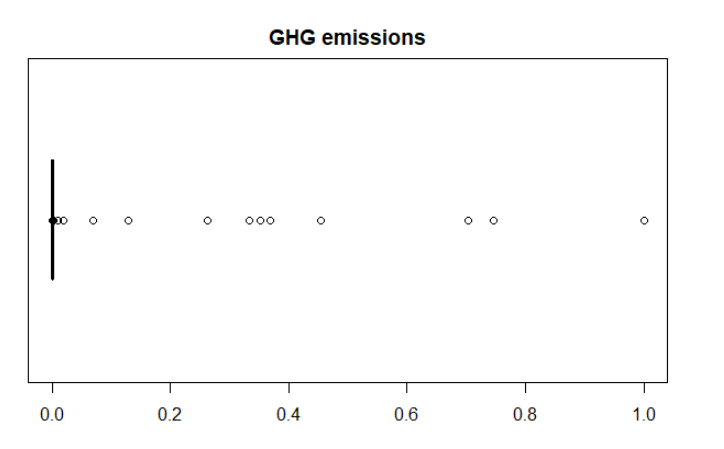
GHGEmissions\_KG = -4.054311e-07 + 7.441562e-01 (FuelOil12\_L) + 2.968370e-02(NaturalGas\_Cubicmeter) + 1.041156e-02 (DistrictHeating\_GJ) + 1.779203e-01 (FuelOil46\_L) + -3.223940e-03 (DistrictCooling\_GJ) + 5.522074e-02(Electricity\_kwh)

# Results

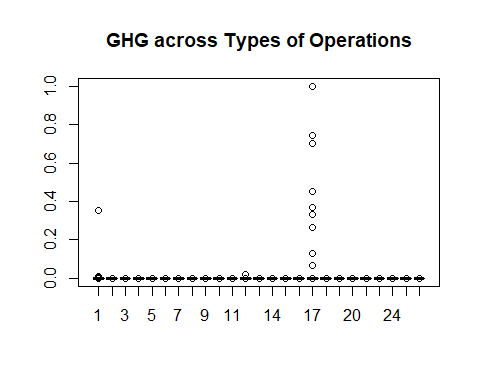
The 4 sectors are Municipal, Post-Secondary Educational Institution, Public Hospital, School Board and the Operation types were reduced to

GHG emission varied from 0 to a very large number. Therefore normalized data was used for the study. See plot below.

**Plot 1: GHG emissions using normalized data**



**Plot2:** **GHG emissions using normalized data across operation types**



|  |  |
| --- | --- |
| **OperationType** | **TypeofOperation** |
| Administrative office | 1 |
| Library | 2 |
| Water treatment | 3 |
| Water pumping | 4 |
| Sewage treatment | 5 |
| Sewage pumping | 6 |
| Police station | 7 |
| Fire station | 8 |
| Storage | 9 |
| Community centres | 10 |
| Classrooms | 11 |
| Hospital | 12 |
| Ambulance station | 13 |
| Laboratories | 14 |
| Student residences | 15 |
| Recreational facilities | 16 |
| School | 17 |
| Parking | 18 |
| Indoor swimming pools | 19 |
| Indoor ice rinks | 20 |
| Multi-use | 21 |
| Art galleries | 22 |
| Performing arts facilities | 23 |
| Auditoriums | 24 |
| Other | 25 |
|  |  |

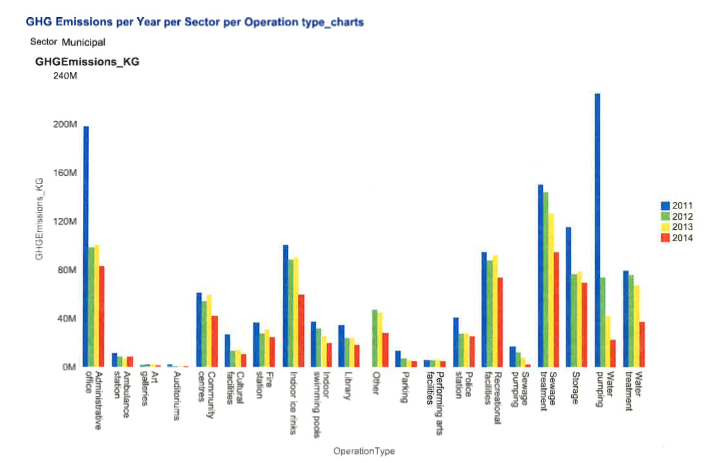
Sector School Board has the highest number of records in the dataset and also the highest GHG. Data below shows the total GHG using the normalized data.

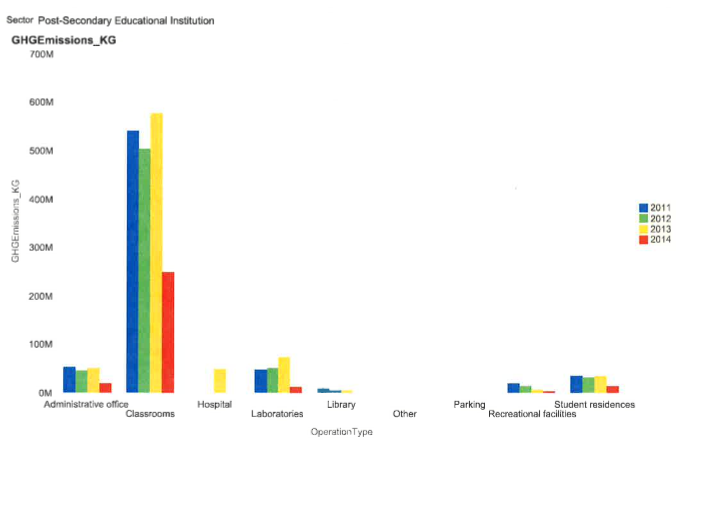
Sector NumofRows totalGHG  
  
1 School Board 25016 4.46   
2 Public Hospital 2262 0.0330  
3 Municipal 55951 0.0159  
4 Post-Secondary Educational Institution 4545 0.0123

Chart 1 and Chart 2 were obtained using Oracle Business Intelligence Enterprise Edition. They show the variation in GHG across operation types through the years 2011-2014.

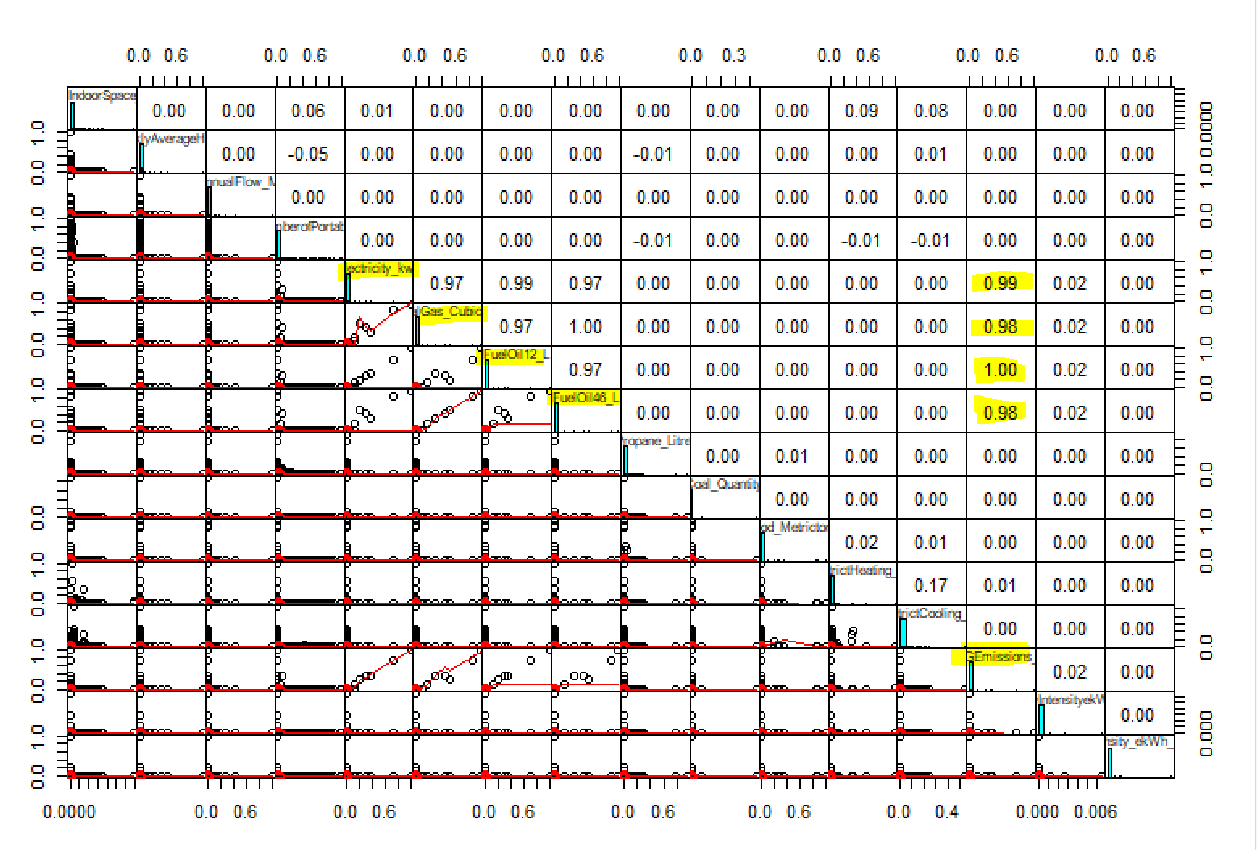
Principal Component Analysis was not an efficient method to reduce the number of dimension as it showed almost the same number of attributes at the conclusion of the analysis. Dimensionality reduction using Forward selection worked well. It showed that GHG can be determined from the attributes Electricity\_kwh, NaturalGas\_Cubicmeter, FuelOil12\_L , FuelOil46\_L, DistrictHeating\_GJ, DistrictCooling\_GJ and TypeofSector.

Correlation analysis showed a strong positive linear relationship between GHG and Electricity\_kwh, NaturalGas\_Cubicmeter, FuelOil12\_L, FuelOil46\_L (Chart 3) with correlations of 0.99, 0.98, 1.00, 0.98 respectively.

**Chart 1: Municipal Sector – Variation of GHG across operation types for 2011-2014**

**Chart 2: Post-Secondary Educational Sector – Variation of GHG across operation types for 2011-2014**.

**Chart 3: Correlation between Attributes**



The simple linear regression model algorithm treated TypeofSector as numerical although it is a factor and therefore this attribute was excluded in models 2 and 3.

train.set$GHGEmissions\_KG ~ train.set$FuelOil12\_L + train.set$NaturalGas\_Cubicmeter + train.set$DistrictHeating\_GJ + train.set$FuelOil46\_L + train.set$DistrictCooling\_GJ + train.set$Electricity\_kwh

In the glm model Akaike Information Criteria AIC is -1034690 which is low (good).

Use of simple linear regression and general linear regression gave the same root mean square error rmse 0.013633. RMSE is the square root of the variance of the residuals. It shows how closely the predicted value matches the actual value. It is a good measure of the fit of the model.

This low value for rmse shows that the model has high accuracy.

Both lm and glm gave the same intercept and coefficients. It can be concluded that GHG can be calculated using the formula equation

GHGEmissions\_KG = -4.054311e-07 + 7.441562e-01 (FuelOil12\_L) + 2.968370e-02(NaturalGas\_Cubicmeter) + 1.041156e-02 (DistrictHeating\_GJ) + 1.779203e-01 (FuelOil46\_L) + -3.223940e-03 (DistrictCooling\_GJ) + 5.522074e-02(Electricity\_kwh)

# Conclusions

Schools in the province as a whole emit the largest amount of GHG. Inspection of the raw data showed a big spread in GHG emission across organizations of the same operation type. Floor area and volume of water do not directly influence the GHG produced.

This study shows that the quantity of GHG emitted is ultimately dependent on the amount of energy consumed. The use of coal, wood and renewables are insignificant in quantity and therefore are not major contributors to GHG. Significant contributors to GHG are the quantity of electricity, gas and fuel oil used for heating and cooling.

The following equation can be used to predict the GHG for any sector and operation type if we insert the values for the independent variables.

GHGEmissions\_KG = -4.054311e-07 + 7.441562e-01 (FuelOil12\_L) + 2.968370e-02(NaturalGas\_Cubicmeter) + 1.041156e-02 (DistrictHeating\_GJ) + 1.779203e-01 (FuelOil46\_L) + -3.223940e-03 (DistrictCooling\_GJ) + 5.522074e-02(Electricity\_kwh)

**References:**

* 1. Green energy Act 2009, Ontario Regulation 397/11

<https://www.ontario.ca/laws/regulation/110397>

* 1. Dataset used

Energy use and greenhouse gas emissions for the Broader Public Sector

<https://www.ontario.ca/data/energy-use-and-greenhouse-gas-emissions-broader-public-sector>

* 1. AIRS – Improving Weather Forecasting and Providing New data on greenhouse gases, Moustafa T. Chahine et al

<https://app.dimensions.ai/details/publication/pub.1011166148>

* 1. Hourly Live GHG Calculation and Forecasting, A.Sahebalam et al

<http://www.screamingpower.ca/ieee-hourly-calculation-forecasting/>

* 1. Forecasting Greenhouse Gas Emissions from Urban Regions: Microsimulation of Land Use and Transport Patterns in Austin Texas, Sumala Tirumalachetty et al

<http://www.caee.utexas.edu/prof/kockelman/public_html/trb10microsimulationco2.pdf>

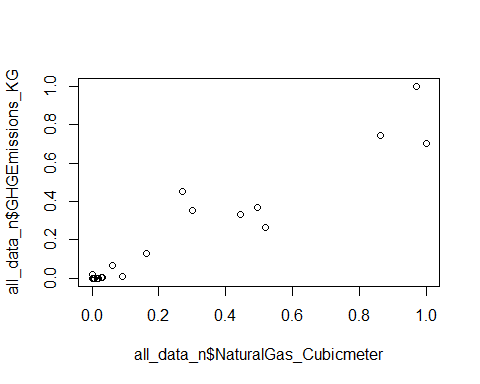
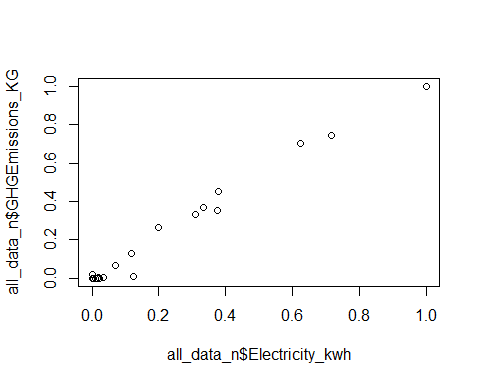
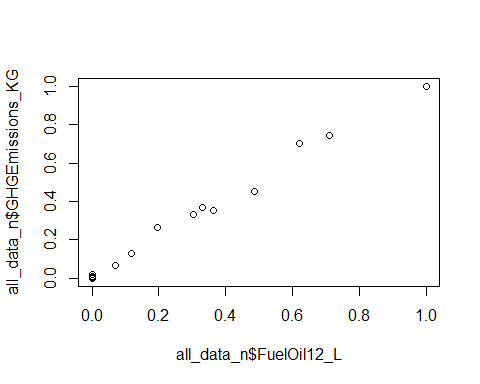
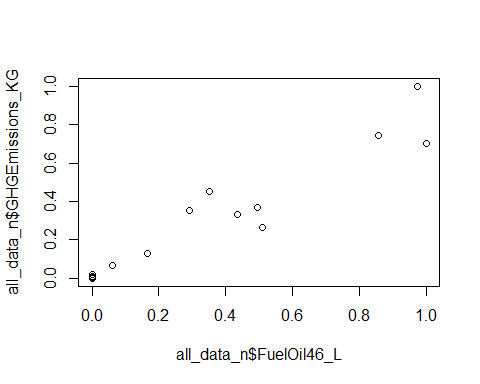
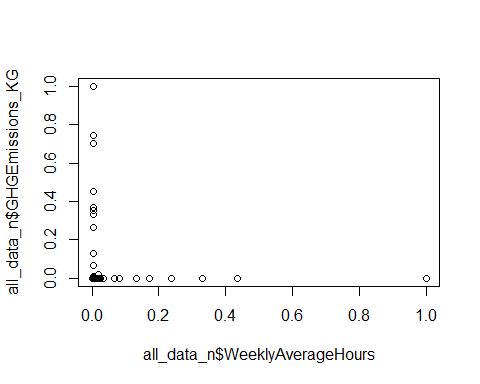
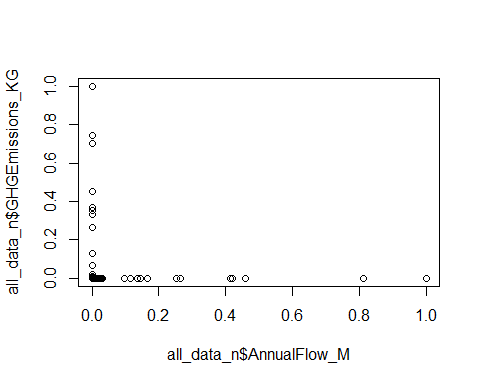
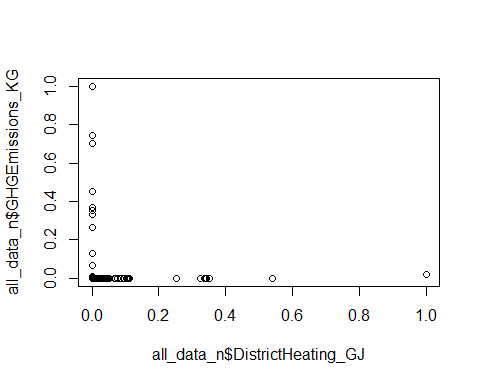
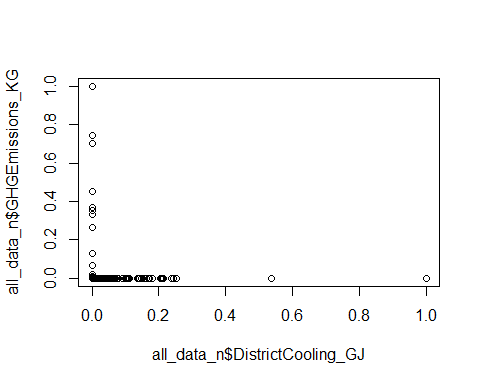
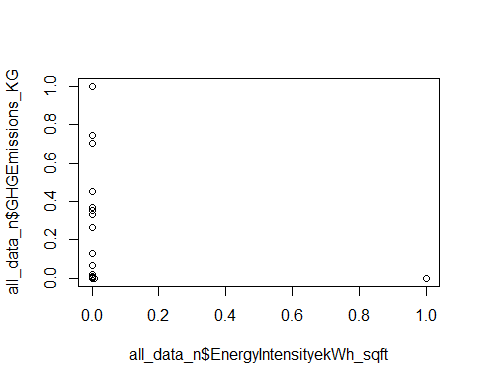
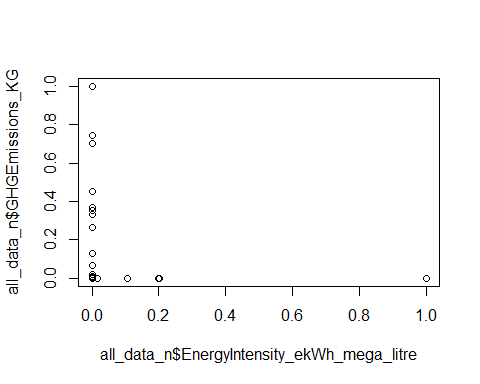
* 1. Modeling of energy consumption and GHG (greenhouse gas) intensity and emissions in Europe using general regression neural networks, Davor Antanasijevic et al

<https://www.tib.eu/en/search/id/BLSE%3ARN369803220/Modeling-of-energy-consumption-and-related-GHG/>

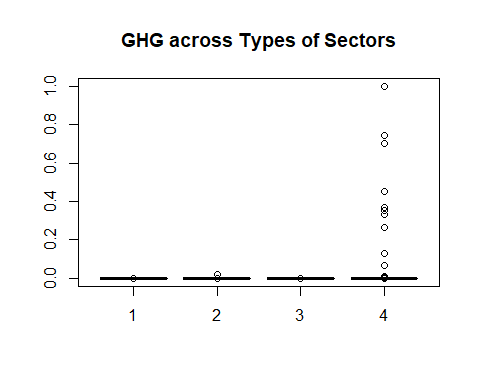
* 1. Forecasting GHG emissions using an optimized artificial neural network model based on correlation and principal component analysis, Davor Antanasijevic et al

<https://www.infona.pl/resource/bwmeta1.element.elsevier-9652f505-e560-3e9e-869d-767844bfdc94>

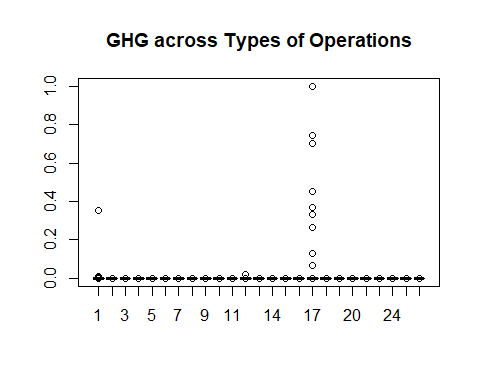
**Appendix 1: Correlation of GHG emission with numeric variables using normalized data**

**Appendix 2: GHG across Sectors and Operation Types**

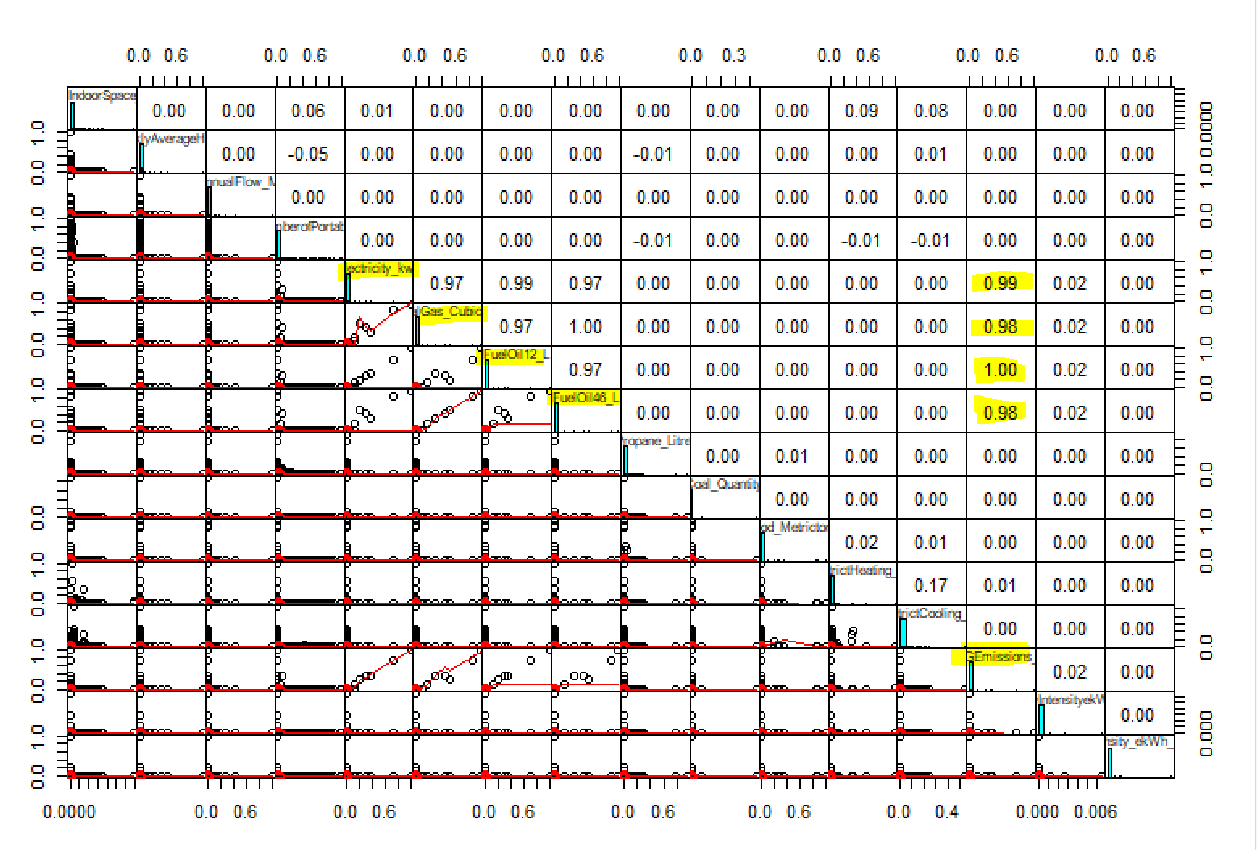


|  |  |
| --- | --- |
| **Sector** | **TypeofSector** |
| Post-Secondary Educational Institution | 1 |
| Public Hospital | 2 |
| Municipal | 3 |
| School Board | 4 |

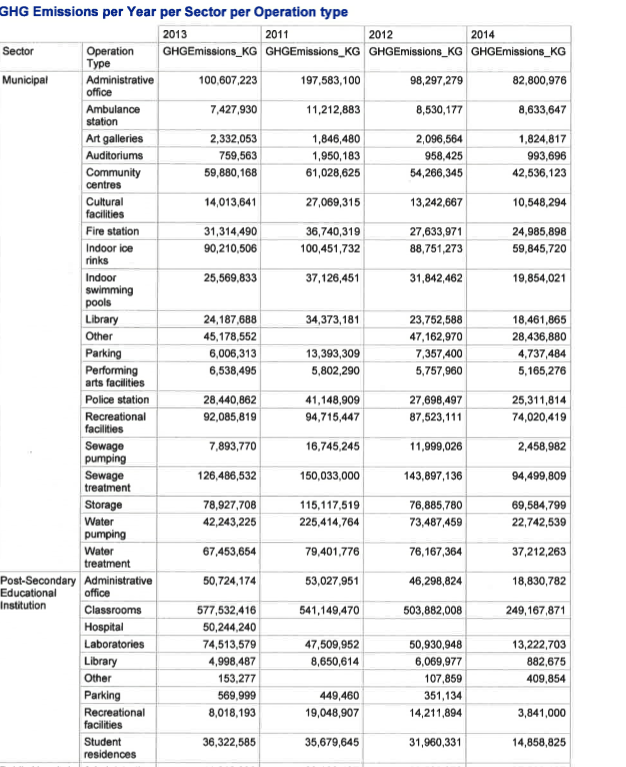


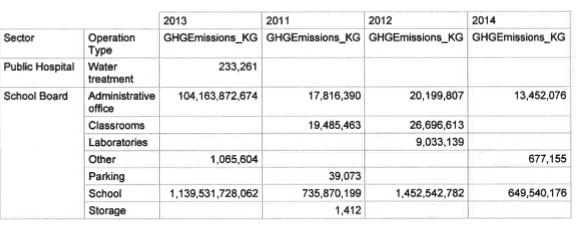
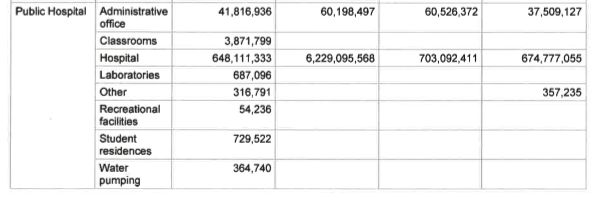
|  |  |
| --- | --- |
| **OperationType** | **TypeofOperation** |
| Administrative office | 1 |
| Library | 2 |
| Water treatment | 3 |
| Water pumping | 4 |
| Sewage treatment | 5 |
| Sewage pumping | 6 |
| Police station | 7 |
| Fire station | 8 |
| Storage | 9 |
| Community centres | 10 |
| Classrooms | 11 |
| Hospital | 12 |
| Ambulance station | 13 |
| Laboratories | 14 |
| Student residences | 15 |
| Recreational facilities | 16 |
| School | 17 |
| Parking | 18 |
| Indoor swimming pools | 19 |
| Indoor ice rinks | 20 |
| Multi-use | 21 |
| Art galleries | 22 |
| Performing arts facilities | 23 |
| Auditoriums | 24 |
| Other | 25 |

**Appendix 3: Correlations between GHG and numeric variables (using normalized data)**

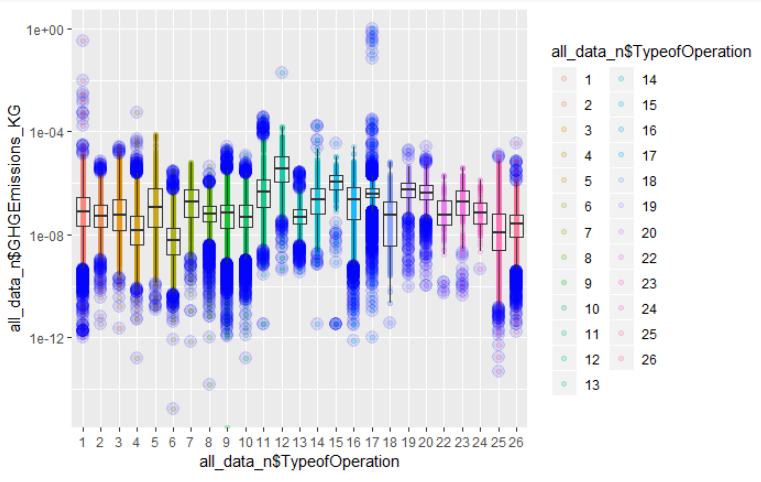
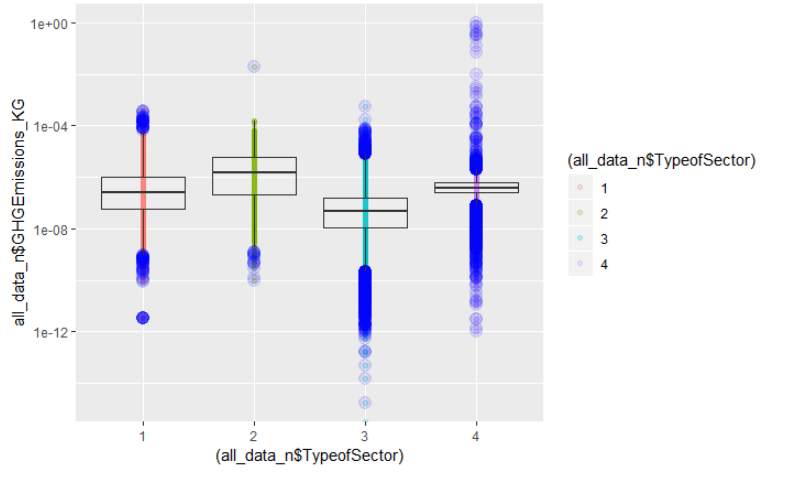


**Appendix 4: Summary of Total GHG (using Oracle Business Intelligence)**





**Appendix 5: Outliers or Extreme values (using normalized data)**

Plot of log GHG versus Type of Operation, (blue shows the outliers)Plot of log GHG versus Sector, (blue shows the outliers)

Appendix 6a: Principal Component Analysis

all\_data\_n\_pca<-prcomp(all\_data\_n[,c(9:12,14:21, 23,25:27)], scale = TRUE)  
summary(all\_data\_n\_pca)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6  
## Standard deviation 2.2155 1.17770 1.07886 1.02180 1.00112 1.00005  
## Proportion of Variance 0.3068 0.08669 0.07275 0.06525 0.06264 0.06251  
## Cumulative Proportion 0.3068 0.39345 0.46620 0.53146 0.59410 0.65660  
## PC7 PC8 PC9 PC10 PC11 PC12  
## Standard deviation 1.0000 1.0000 0.99947 0.97615 0.89941 0.80121  
## Proportion of Variance 0.0625 0.0625 0.06243 0.05955 0.05056 0.04012  
## Cumulative Proportion 0.7191 0.7816 0.84404 0.90359 0.95415 0.99427  
## PC13 PC14 PC15 PC16  
## Standard deviation 0.28707 0.09548 0.009751 0.006503  
## Proportion of Variance 0.00515 0.00057 0.000010 0.000000  
## Cumulative Proportion 0.99942 0.99999 1.000000 1.000000

# PCA shows that the variance of first 13 principalcomponents account for 99.9% of the variance  
#PC1 has proportion of variance 30.68%, PC2 8.66 %, PC3 7.3% ..  
# Principal component analysis may not be the best method for feature selection as it does not narrow down the number of attributes of importance  
  
  
# PCA excluding GHG column  
all\_data\_n\_pca<-prcomp(all\_data\_n[,c(9:12,14:21, 23,26,27)], scale = TRUE)  
summary(all\_data\_n\_pca)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6  
## Standard deviation 1.979 1.17770 1.0789 1.02180 1.00112 1.00005  
## Proportion of Variance 0.261 0.09246 0.0776 0.06961 0.06682 0.06667  
## Cumulative Proportion 0.261 0.35347 0.4311 0.50067 0.56749 0.63416  
## PC7 PC8 PC9 PC10 PC11 PC12  
## Standard deviation 1.00001 0.99999 0.9995 0.97615 0.89940 0.80120  
## Proportion of Variance 0.06667 0.06667 0.0666 0.06352 0.05393 0.04279  
## Cumulative Proportion 0.70083 0.76749 0.8341 0.89761 0.95154 0.99434  
## PC13 PC14 PC15  
## Standard deviation 0.27731 0.08922 0.009506  
## Proportion of Variance 0.00513 0.00053 0.000010  
## Cumulative Proportion 0.99946 0.99999 1.000000

# PCA shows that the variance of first 13 principalcomponents account for 99.9% of the variance]]  
#PC1 has proportion of variance 26.1%, PC2 9.2 %, PC3 7.7% ..  
# Principal component analysis may not be the best method for feature selection as it does not narrow down the number of attributes of importance

**Appendix 6b: Feature optimization – using forward selection**

# Feature optimization (Dimensionality Reduction) - using the forward selection algorithm

library(MASS)

## Warning: package 'MASS' was built under R version 3.5.1

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

# all the independent variables are included  
full <- lm(all\_data\_n$GHGEmissions\_KG ~ all\_data\_n$TypeofSector+ all\_data\_n$TypeofOperation + all\_data\_n$TotalIndoorSpace\_sqft + all\_data\_n$WeeklyAverageHours + all\_data\_n$AnnualFlow\_M + all\_data$NumberofPortables + all\_data\_n$Electricity\_kwh + all\_data\_n$NaturalGas\_Cubicmeter + all\_data\_n$Propane\_Litre + all\_data\_n$DistrictHeating\_GJ + all\_data\_n$DistrictCooling\_GJ + all\_data\_n$FuelOil12\_L + all\_data\_n$FuelOil46\_L, data=all\_data\_n)  
  
# none of the independent variables are selected.  
null <- lm(all\_data\_n$GHGEmissions\_KG~1,data=all\_data\_n)  
  
stepF <- stepAIC(null, scope=list(lower=null, upper=full), direction= "forward", trace=FALSE)  
  
#display results]  
summary(stepF)

##   
## Call:  
## lm(formula = all\_data\_n$GHGEmissions\_KG ~ all\_data\_n$FuelOil12\_L +   
## all\_data\_n$NaturalGas\_Cubicmeter + all\_data\_n$DistrictHeating\_GJ +   
## all\_data\_n$FuelOil46\_L + all\_data\_n$DistrictCooling\_GJ +   
## all\_data\_n$Electricity\_kwh + all\_data\_n$TypeofSector, data = all\_data\_n)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.0051341 -0.0000001 0.0000001 0.0000001 0.0096459   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -5.011e-06 6.865e-07 -7.299 2.92e-13  
## all\_data\_n$FuelOil12\_L 7.678e-01 1.165e-03 658.988 < 2e-16  
## all\_data\_n$NaturalGas\_Cubicmeter 6.186e-02 1.538e-03 40.230 < 2e-16  
## all\_data\_n$DistrictHeating\_GJ 9.821e-03 3.260e-05 301.262 < 2e-16  
## all\_data\_n$FuelOil46\_L 1.466e-01 1.550e-03 94.541 < 2e-16  
## all\_data\_n$DistrictCooling\_GJ -2.355e-03 3.173e-05 -74.202 < 2e-16  
## all\_data\_n$Electricity\_kwh 3.007e-02 1.153e-03 26.071 < 2e-16  
## all\_data\_n$TypeofSector2 -6.057e-07 1.189e-06 -0.510 0.61  
## all\_data\_n$TypeofSector3 4.890e-06 7.136e-07 6.852 7.32e-12  
## all\_data\_n$TypeofSector4 5.170e-06 7.458e-07 6.932 4.18e-12  
##   
## (Intercept) \*\*\*  
## all\_data\_n$FuelOil12\_L \*\*\*  
## all\_data\_n$NaturalGas\_Cubicmeter \*\*\*  
## all\_data\_n$DistrictHeating\_GJ \*\*\*  
## all\_data\_n$FuelOil46\_L \*\*\*  
## all\_data\_n$DistrictCooling\_GJ \*\*\*  
## all\_data\_n$Electricity\_kwh \*\*\*  
## all\_data\_n$TypeofSector2   
## all\_data\_n$TypeofSector3 \*\*\*  
## all\_data\_n$TypeofSector4 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.616e-05 on 87764 degrees of freedom  
## Multiple R-squared: 0.9999, Adjusted R-squared: 0.9999   
## F-statistic: 1.418e+08 on 9 and 87764 DF, p-value: < 2.2e-16

stepF$anova

## Stepwise Model Path   
## Analysis of Deviance Table  
##   
## Initial Model:  
## all\_data\_n$GHGEmissions\_KG ~ 1  
##   
## Final Model:  
## all\_data\_n$GHGEmissions\_KG ~ all\_data\_n$FuelOil12\_L + all\_data\_n$NaturalGas\_Cubicmeter +   
## all\_data\_n$DistrictHeating\_GJ + all\_data\_n$FuelOil46\_L +   
## all\_data\_n$DistrictCooling\_GJ + all\_data\_n$Electricity\_kwh +   
## all\_data\_n$TypeofSector  
##   
##   
## Step Df Deviance Resid. Df  
## 1 87773  
## 2 + all\_data\_n$FuelOil12\_L 1 2.705538e+00 87772  
## 3 + all\_data\_n$NaturalGas\_Cubicmeter 1 1.291412e-02 87771  
## 4 + all\_data\_n$DistrictHeating\_GJ 1 1.770751e-04 87770  
## 5 + all\_data\_n$FuelOil46\_L 1 1.510065e-04 87769  
## 6 + all\_data\_n$DistrictCooling\_GJ 1 1.188343e-05 87768  
## 7 + all\_data\_n$Electricity\_kwh 1 1.463582e-06 87767  
## 8 + all\_data\_n$TypeofSector 3 1.691453e-07 87764  
## Resid. Dev AIC  
## 1 2.7189809755 -911290.8  
## 2 0.0134427430 -1377331.3  
## 3 0.0005286186 -1661359.5  
## 4 0.0003515436 -1697163.5  
## 5 0.0002005371 -1746432.1  
## 6 0.0001886536 -1751791.9  
## 7 0.0001871901 -1752473.5  
## 8 0.0001870209 -1752546.8

###### Results give the following

**From Forward Selection**  
# Final Model:all\_data\_n$GHGEmissions\_KG ~ all\_data\_n$FuelOil12\_L + all\_data\_n$NaturalGas\_Cubicmeter + all\_data\_n$DistrictHeating\_GJ + all\_data\_n$FuelOil46\_L + all\_data\_n$DistrictCooling\_GJ + all\_data\_n$Electricity\_kwh + all\_data\_n$TypeofSector