CKME136 Capstone Project

Erma Fernando (Student Number 500925471)

# Load Data

### 1. Read the csv files for 2011, 2012, 2013, 1014, 2015

data2011<-read.csv(file="C:/RY/CKME136/OntarioGHG data/FinalData/bps\_raw\_2011\_FINAL.csv",header=T,sep=",")  
data2012<-read.csv(file="C:/RY/CKME136/OntarioGHG data/FinalData/bps\_raw\_2012\_FINAL.csv",header=T,sep=",")  
data2013<-read.csv(file="C:/RY/CKME136/OntarioGHG data/FinalData/bps\_raw\_2013\_FINAL.csv",header=T,sep=",")  
data2014<-read.csv(file="C:/RY/CKME136/OntarioGHG data/FinalData/bps\_raw\_2014\_FINAL.csv",header=T,sep=",")  
data2015<-read.csv(file="C:/RY/CKME136/OntarioGHG data/FinalData/bps\_raw\_2015\_FINAL.csv",header=T,sep=",")  
  
  
### 2. Merge the data frames for 2011, 2012, 2013, 1014, 2015  
all\_data <-rbind(data2011, data2012, data2013, data2014, data2015)

# Initial Analysis of the data and Data Preparation

# 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"

# remove columns that are not necessary like Unit of measure etc  
all\_data<-all\_data[,-c(9,10, 17, 18, 19, 22,24,26,28,30,31,32,35,36,37,40,44,45)]  
# column headers  
names(all\_data)

## [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"

#num of rows,num of columns, data types  
str(all\_data)

## 'data.frame': 88259 obs. of 27 variables:  
## $ Year : int 2011 2011 2011 2011 2011 2011 2011 2011 2011 2011 ...  
## $ Sector : Factor w/ 4 levels "Municipal","Post-Secondary Educational Institution",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Organization : Factor w/ 825 levels "Alexandra Hospital",..: 401 401 401 401 401 401 401 401 401 401 ...  
## $ Operation : Factor w/ 22856 levels "'A' Block","'B' Block",..: 8756 5424 14557 4040 10818 14509 11638 5425 4562 591 ...  
## $ OperationType : Factor w/ 25 levels "Administrative office",..: 1 13 24 23 23 20 19 16 8 21 ...  
## $ Address : Factor w/ 19780 levels "#1 Hwy 527, c/o Box 98, 112 Queen St.",..: 11332 10812 428 5543 8772 5285 14948 10528 6196 8193 ...  
## $ City : Factor w/ 1429 levels "568, rue Lessard",..: 429 429 429 429 429 429 429 429 429 429 ...  
## $ PostalCode : Factor w/ 19483 levels "H0H 0H0","K0A 1A0",..: 7747 7747 7747 7747 7747 7747 7747 7747 7747 7747 ...  
## $ TotalIndoorSpace\_sqft : num 21510 11840 7890 678 1981 ...  
## $ WeeklyAverageHours : num 40 62 168 168 168 168 168 40 15 20 ...  
## $ AnnualFlow\_M : num NA NA 1406 1 1 ...  
## $ NumberofPortables : num NA NA NA NA NA NA NA NA NA NA ...  
## $ SwimmingPool : chr "FALSE" "FALSE" "FALSE" "FALSE" ...  
## $ Electricity\_kwh : num 248444 100147 994783 15519 123972 ...  
## $ NaturalGas\_Cubicmeter : num 203 18695 NA NA NA ...  
## $ FuelOil12\_L : num NA NA NA NA NA ...  
## $ FuelOil46\_L : num NA NA NA NA NA NA NA NA NA NA ...  
## $ Propane\_Litre : num NA NA NA NA NA NA NA NA NA NA ...  
## $ Coal\_Quantity : num NA NA NA NA NA NA NA NA NA NA ...  
## $ Wood\_Metrictonne : num NA NA NA NA NA NA NA NA NA NA ...  
## $ DistrictHeating\_GJ : num NA NA NA NA NA NA NA NA NA NA ...  
## $ DistrictHeating\_IsRenewable : chr "FALSE" "FALSE" "FALSE" "FALSE" ...  
## $ DistrictCooling\_GJ : num NA NA NA NA NA NA NA NA NA NA ...  
## $ DistrictCooling\_IsRenewable : chr "FALSE" "FALSE" "FALSE" "FALSE" ...  
## $ GHGEmissions\_KG : num 24740 45163 97529 1521 12154 ...  
## $ EnergyIntensityekWh\_sqft : num 11.7 25.2 126.1 22.9 62.6 ...  
## $ EnergyIntensity\_ekWh\_mega\_litre: num NA NA 708 15519 123972 ...

# 88259 rows 27 columns  
#convert Year to a factor  
all\_data$Year<- as.factor(all\_data$Year)  
  
# Convert TRUE to 1 and FALSE and NA to 0. Then convert these variables to factors  
all\_data$SwimmingPool<-ifelse(all\_data$SwimmingPool=="TRUE",1,0)  
  
all\_data$DistrictHeating\_IsRenewable<-ifelse(all\_data$DistrictHeating\_IsRenewable=="TRUE",1,0)  
  
all\_data$DistrictCooling\_IsRenewable<-ifelse(all\_data$DistrictCooling\_IsRenewable=="TRUE",1,0)  
  
all\_data$SwimmingPool<- as.factor(all\_data$SwimmingPool)  
  
all\_data$DistrictHeating\_IsRenewable<- as.factor(all\_data$DistrictHeating\_IsRenewable)  
  
all\_data$DistrictCooling\_IsRenewable<- as.factor(all\_data$DistrictCooling\_IsRenewable)  
  
str(all\_data)

## 'data.frame': 88259 obs. of 27 variables:  
## $ Year : Factor w/ 5 levels "2011","2012",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Sector : Factor w/ 4 levels "Municipal","Post-Secondary Educational Institution",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Organization : Factor w/ 825 levels "Alexandra Hospital",..: 401 401 401 401 401 401 401 401 401 401 ...  
## $ Operation : Factor w/ 22856 levels "'A' Block","'B' Block",..: 8756 5424 14557 4040 10818 14509 11638 5425 4562 591 ...  
## $ OperationType : Factor w/ 25 levels "Administrative office",..: 1 13 24 23 23 20 19 16 8 21 ...  
## $ Address : Factor w/ 19780 levels "#1 Hwy 527, c/o Box 98, 112 Queen St.",..: 11332 10812 428 5543 8772 5285 14948 10528 6196 8193 ...  
## $ City : Factor w/ 1429 levels "568, rue Lessard",..: 429 429 429 429 429 429 429 429 429 429 ...  
## $ PostalCode : Factor w/ 19483 levels "H0H 0H0","K0A 1A0",..: 7747 7747 7747 7747 7747 7747 7747 7747 7747 7747 ...  
## $ TotalIndoorSpace\_sqft : num 21510 11840 7890 678 1981 ...  
## $ WeeklyAverageHours : num 40 62 168 168 168 168 168 40 15 20 ...  
## $ AnnualFlow\_M : num NA NA 1406 1 1 ...  
## $ NumberofPortables : num NA NA NA NA NA NA NA NA NA NA ...  
## $ SwimmingPool : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Electricity\_kwh : num 248444 100147 994783 15519 123972 ...  
## $ NaturalGas\_Cubicmeter : num 203 18695 NA NA NA ...  
## $ FuelOil12\_L : num NA NA NA NA NA ...  
## $ FuelOil46\_L : num NA NA NA NA NA NA NA NA NA NA ...  
## $ Propane\_Litre : num NA NA NA NA NA NA NA NA NA NA ...  
## $ Coal\_Quantity : num NA NA NA NA NA NA NA NA NA NA ...  
## $ Wood\_Metrictonne : num NA NA NA NA NA NA NA NA NA NA ...  
## $ DistrictHeating\_GJ : num NA NA NA NA NA NA NA NA NA NA ...  
## $ DistrictHeating\_IsRenewable : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ DistrictCooling\_GJ : num NA NA NA NA NA NA NA NA NA NA ...  
## $ DistrictCooling\_IsRenewable : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ GHGEmissions\_KG : num 24740 45163 97529 1521 12154 ...  
## $ EnergyIntensityekWh\_sqft : num 11.7 25.2 126.1 22.9 62.6 ...  
## $ EnergyIntensity\_ekWh\_mega\_litre: num NA NA 708 15519 123972 ...

# 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

# replace null values in numerical columns with 0   
all\_data[is.na(all\_data)] <- 0

## Warning in `[<-.factor`(`\*tmp\*`, thisvar, value = 0): invalid factor level,  
## NA generated  
  
## Warning in `[<-.factor`(`\*tmp\*`, thisvar, value = 0): invalid factor level,  
## NA generated

# Check rows with missing data - none - because all replaced with 0 in last step  
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   
## 0 0   
## AnnualFlow\_M NumberofPortables   
## 0 0   
## SwimmingPool Electricity\_kwh   
## 0 0   
## NaturalGas\_Cubicmeter FuelOil12\_L   
## 0 0   
## FuelOil46\_L Propane\_Litre   
## 0 0   
## Coal\_Quantity Wood\_Metrictonne   
## 0 0   
## DistrictHeating\_GJ DistrictHeating\_IsRenewable   
## 0 0   
## DistrictCooling\_GJ DistrictCooling\_IsRenewable   
## 0 0   
## GHGEmissions\_KG EnergyIntensityekWh\_sqft   
## 0 0   
## EnergyIntensity\_ekWh\_mega\_litre   
## 0

#num of rows,num of columns, data types 88615 obs. of 27 variables  
str(all\_data)

## 'data.frame': 88259 obs. of 27 variables:  
## $ Year : Factor w/ 5 levels "2011","2012",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Sector : Factor w/ 4 levels "Municipal","Post-Secondary Educational Institution",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Organization : Factor w/ 825 levels "Alexandra Hospital",..: 401 401 401 401 401 401 401 401 401 401 ...  
## $ Operation : Factor w/ 22856 levels "'A' Block","'B' Block",..: 8756 5424 14557 4040 10818 14509 11638 5425 4562 591 ...  
## $ OperationType : Factor w/ 25 levels "Administrative office",..: 1 13 24 23 23 20 19 16 8 21 ...  
## $ Address : Factor w/ 19780 levels "#1 Hwy 527, c/o Box 98, 112 Queen St.",..: 11332 10812 428 5543 8772 5285 14948 10528 6196 8193 ...  
## $ City : Factor w/ 1429 levels "568, rue Lessard",..: 429 429 429 429 429 429 429 429 429 429 ...  
## $ PostalCode : Factor w/ 19483 levels "H0H 0H0","K0A 1A0",..: 7747 7747 7747 7747 7747 7747 7747 7747 7747 7747 ...  
## $ TotalIndoorSpace\_sqft : num 21510 11840 7890 678 1981 ...  
## $ WeeklyAverageHours : num 40 62 168 168 168 168 168 40 15 20 ...  
## $ AnnualFlow\_M : num 0 0 1406 1 1 ...  
## $ NumberofPortables : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ SwimmingPool : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Electricity\_kwh : num 248444 100147 994783 15519 123972 ...  
## $ NaturalGas\_Cubicmeter : num 203 18695 0 0 0 ...  
## $ FuelOil12\_L : num 0 0 0 0 0 ...  
## $ FuelOil46\_L : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Propane\_Litre : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Coal\_Quantity : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Wood\_Metrictonne : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DistrictHeating\_GJ : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DistrictHeating\_IsRenewable : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ DistrictCooling\_GJ : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DistrictCooling\_IsRenewable : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ GHGEmissions\_KG : num 24740 45163 97529 1521 12154 ...  
## $ EnergyIntensityekWh\_sqft : num 11.7 25.2 126.1 22.9 62.6 ...  
## $ EnergyIntensity\_ekWh\_mega\_litre: num 0 0 708 15519 123972 ...

# min, max, median, mean 1st quartile, 3rd quartile  
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   
##

# count of various data types -factor 8 numeric 19  
table(unlist(lapply(all\_data,class)))

##   
## factor numeric   
## 11 16

#list the unique values in variable "OperationType""  
sapply(all\_data, function(x) unique(all\_data$OperationType))

## Year Sector   
## [1,] "Administrative office" "Administrative office"   
## [2,] "Library" "Library"   
## [3,] "Water treatment" "Water treatment"   
## [4,] "Water pumping" "Water pumping"   
## [5,] "Sewage treatment" "Sewage treatment"   
## [6,] "Sewage pumping" "Sewage pumping"   
## [7,] "Police station" "Police station"   
## [8,] "Fire station" "Fire station"   
## [9,] "Storage" "Storage"   
## [10,] "Community centres" "Community centres"   
## [11,] "Classrooms" "Classrooms"   
## [12,] "Hospital" "Hospital"   
## [13,] "Ambulance station" "Ambulance station"   
## [14,] "Laboratories" "Laboratories"   
## [15,] "Student residences" "Student residences"   
## [16,] "Recreational facilities" "Recreational facilities"   
## [17,] "Indoor swimming pools" "Indoor swimming pools"   
## [18,] "Indoor ice rinks" "Indoor ice rinks"   
## [19,] "Cultural facilities" "Cultural facilities"   
## [20,] "Parking" "Parking"   
## [21,] "School" "School"   
## [22,] "Art galleries" "Art galleries"   
## [23,] "Performing arts facilities" "Performing arts facilities"  
## [24,] "Auditoriums" "Auditoriums"   
## [25,] "Other" "Other"   
## Organization Operation   
## [1,] "Administrative office" "Administrative office"   
## [2,] "Library" "Library"   
## [3,] "Water treatment" "Water treatment"   
## [4,] "Water pumping" "Water pumping"   
## [5,] "Sewage treatment" "Sewage treatment"   
## [6,] "Sewage pumping" "Sewage pumping"   
## [7,] "Police station" "Police station"   
## [8,] "Fire station" "Fire station"   
## [9,] "Storage" "Storage"   
## [10,] "Community centres" "Community centres"   
## [11,] "Classrooms" "Classrooms"   
## [12,] "Hospital" "Hospital"   
## [13,] "Ambulance station" "Ambulance station"   
## [14,] "Laboratories" "Laboratories"   
## [15,] "Student residences" "Student residences"   
## [16,] "Recreational facilities" "Recreational facilities"   
## [17,] "Indoor swimming pools" "Indoor swimming pools"   
## [18,] "Indoor ice rinks" "Indoor ice rinks"   
## [19,] "Cultural facilities" "Cultural facilities"   
## [20,] "Parking" "Parking"   
## [21,] "School" "School"   
## [22,] "Art galleries" "Art galleries"   
## [23,] "Performing arts facilities" "Performing arts facilities"  
## [24,] "Auditoriums" "Auditoriums"   
## [25,] "Other" "Other"   
## OperationType Address   
## [1,] "Administrative office" "Administrative office"   
## [2,] "Library" "Library"   
## [3,] "Water treatment" "Water treatment"   
## [4,] "Water pumping" "Water pumping"   
## [5,] "Sewage treatment" "Sewage treatment"   
## [6,] "Sewage pumping" "Sewage pumping"   
## [7,] "Police station" "Police station"   
## [8,] "Fire station" "Fire station"   
## [9,] "Storage" "Storage"   
## [10,] "Community centres" "Community centres"   
## [11,] "Classrooms" "Classrooms"   
## [12,] "Hospital" "Hospital"   
## [13,] "Ambulance station" "Ambulance station"   
## [14,] "Laboratories" "Laboratories"   
## [15,] "Student residences" "Student residences"   
## [16,] "Recreational facilities" "Recreational facilities"   
## [17,] "Indoor swimming pools" "Indoor swimming pools"   
## [18,] "Indoor ice rinks" "Indoor ice rinks"   
## [19,] "Cultural facilities" "Cultural facilities"   
## [20,] "Parking" "Parking"   
## [21,] "School" "School"   
## [22,] "Art galleries" "Art galleries"   
## [23,] "Performing arts facilities" "Performing arts facilities"  
## [24,] "Auditoriums" "Auditoriums"   
## [25,] "Other" "Other"   
## City PostalCode   
## [1,] "Administrative office" "Administrative office"   
## [2,] "Library" "Library"   
## [3,] "Water treatment" "Water treatment"   
## [4,] "Water pumping" "Water pumping"   
## [5,] "Sewage treatment" "Sewage treatment"   
## [6,] "Sewage pumping" "Sewage pumping"   
## [7,] "Police station" "Police station"   
## [8,] "Fire station" "Fire station"   
## [9,] "Storage" "Storage"   
## [10,] "Community centres" "Community centres"   
## [11,] "Classrooms" "Classrooms"   
## [12,] "Hospital" "Hospital"   
## [13,] "Ambulance station" "Ambulance station"   
## [14,] "Laboratories" "Laboratories"   
## [15,] "Student residences" "Student residences"   
## [16,] "Recreational facilities" "Recreational facilities"   
## [17,] "Indoor swimming pools" "Indoor swimming pools"   
## [18,] "Indoor ice rinks" "Indoor ice rinks"   
## [19,] "Cultural facilities" "Cultural facilities"   
## [20,] "Parking" "Parking"   
## [21,] "School" "School"   
## [22,] "Art galleries" "Art galleries"   
## [23,] "Performing arts facilities" "Performing arts facilities"  
## [24,] "Auditoriums" "Auditoriums"   
## [25,] "Other" "Other"   
## TotalIndoorSpace\_sqft WeeklyAverageHours   
## [1,] "Administrative office" "Administrative office"   
## [2,] "Library" "Library"   
## [3,] "Water treatment" "Water treatment"   
## [4,] "Water pumping" "Water pumping"   
## [5,] "Sewage treatment" "Sewage treatment"   
## [6,] "Sewage pumping" "Sewage pumping"   
## [7,] "Police station" "Police station"   
## [8,] "Fire station" "Fire station"   
## [9,] "Storage" "Storage"   
## [10,] "Community centres" "Community centres"   
## [11,] "Classrooms" "Classrooms"   
## [12,] "Hospital" "Hospital"   
## [13,] "Ambulance station" "Ambulance station"   
## [14,] "Laboratories" "Laboratories"   
## [15,] "Student residences" "Student residences"   
## [16,] "Recreational facilities" "Recreational facilities"   
## [17,] "Indoor swimming pools" "Indoor swimming pools"   
## [18,] "Indoor ice rinks" "Indoor ice rinks"   
## [19,] "Cultural facilities" "Cultural facilities"   
## [20,] "Parking" "Parking"   
## [21,] "School" "School"   
## [22,] "Art galleries" "Art galleries"   
## [23,] "Performing arts facilities" "Performing arts facilities"  
## [24,] "Auditoriums" "Auditoriums"   
## [25,] "Other" "Other"   
## AnnualFlow\_M NumberofPortables   
## [1,] "Administrative office" "Administrative office"   
## [2,] "Library" "Library"   
## [3,] "Water treatment" "Water treatment"   
## [4,] "Water pumping" "Water pumping"   
## [5,] "Sewage treatment" "Sewage treatment"   
## [6,] "Sewage pumping" "Sewage pumping"   
## [7,] "Police station" "Police station"   
## [8,] "Fire station" "Fire station"   
## [9,] "Storage" "Storage"   
## [10,] "Community centres" "Community centres"   
## [11,] "Classrooms" "Classrooms"   
## [12,] "Hospital" "Hospital"   
## [13,] "Ambulance station" "Ambulance station"   
## [14,] "Laboratories" "Laboratories"   
## [15,] "Student residences" "Student residences"   
## [16,] "Recreational facilities" "Recreational facilities"   
## [17,] "Indoor swimming pools" "Indoor swimming pools"   
## [18,] "Indoor ice rinks" "Indoor ice rinks"   
## [19,] "Cultural facilities" "Cultural facilities"   
## [20,] "Parking" "Parking"   
## [21,] "School" "School"   
## [22,] "Art galleries" "Art galleries"   
## [23,] "Performing arts facilities" "Performing arts facilities"  
## [24,] "Auditoriums" "Auditoriums"   
## [25,] "Other" "Other"   
## SwimmingPool Electricity\_kwh   
## [1,] "Administrative office" "Administrative office"   
## [2,] "Library" "Library"   
## [3,] "Water treatment" "Water treatment"   
## [4,] "Water pumping" "Water pumping"   
## [5,] "Sewage treatment" "Sewage treatment"   
## [6,] "Sewage pumping" "Sewage pumping"   
## [7,] "Police station" "Police station"   
## [8,] "Fire station" "Fire station"   
## [9,] "Storage" "Storage"   
## [10,] "Community centres" "Community centres"   
## [11,] "Classrooms" "Classrooms"   
## [12,] "Hospital" "Hospital"   
## [13,] "Ambulance station" "Ambulance station"   
## [14,] "Laboratories" "Laboratories"   
## [15,] "Student residences" "Student residences"   
## [16,] "Recreational facilities" "Recreational facilities"   
## [17,] "Indoor swimming pools" "Indoor swimming pools"   
## [18,] "Indoor ice rinks" "Indoor ice rinks"   
## [19,] "Cultural facilities" "Cultural facilities"   
## [20,] "Parking" "Parking"   
## [21,] "School" "School"   
## [22,] "Art galleries" "Art galleries"   
## [23,] "Performing arts facilities" "Performing arts facilities"  
## [24,] "Auditoriums" "Auditoriums"   
## [25,] "Other" "Other"   
## NaturalGas\_Cubicmeter FuelOil12\_L   
## [1,] "Administrative office" "Administrative office"   
## [2,] "Library" "Library"   
## [3,] "Water treatment" "Water treatment"   
## [4,] "Water pumping" "Water pumping"   
## [5,] "Sewage treatment" "Sewage treatment"   
## [6,] "Sewage pumping" "Sewage pumping"   
## [7,] "Police station" "Police station"   
## [8,] "Fire station" "Fire station"   
## [9,] "Storage" "Storage"   
## [10,] "Community centres" "Community centres"   
## [11,] "Classrooms" "Classrooms"   
## [12,] "Hospital" "Hospital"   
## [13,] "Ambulance station" "Ambulance station"   
## [14,] "Laboratories" "Laboratories"   
## [15,] "Student residences" "Student residences"   
## [16,] "Recreational facilities" "Recreational facilities"   
## [17,] "Indoor swimming pools" "Indoor swimming pools"   
## [18,] "Indoor ice rinks" "Indoor ice rinks"   
## [19,] "Cultural facilities" "Cultural facilities"   
## [20,] "Parking" "Parking"   
## [21,] "School" "School"   
## [22,] "Art galleries" "Art galleries"   
## [23,] "Performing arts facilities" "Performing arts facilities"  
## [24,] "Auditoriums" "Auditoriums"   
## [25,] "Other" "Other"   
## FuelOil46\_L Propane\_Litre   
## [1,] "Administrative office" "Administrative office"   
## [2,] "Library" "Library"   
## [3,] "Water treatment" "Water treatment"   
## [4,] "Water pumping" "Water pumping"   
## [5,] "Sewage treatment" "Sewage treatment"   
## [6,] "Sewage pumping" "Sewage pumping"   
## [7,] "Police station" "Police station"   
## [8,] "Fire station" "Fire station"   
## [9,] "Storage" "Storage"   
## [10,] "Community centres" "Community centres"   
## [11,] "Classrooms" "Classrooms"   
## [12,] "Hospital" "Hospital"   
## [13,] "Ambulance station" "Ambulance station"   
## [14,] "Laboratories" "Laboratories"   
## [15,] "Student residences" "Student residences"   
## [16,] "Recreational facilities" "Recreational facilities"   
## [17,] "Indoor swimming pools" "Indoor swimming pools"   
## [18,] "Indoor ice rinks" "Indoor ice rinks"   
## [19,] "Cultural facilities" "Cultural facilities"   
## [20,] "Parking" "Parking"   
## [21,] "School" "School"   
## [22,] "Art galleries" "Art galleries"   
## [23,] "Performing arts facilities" "Performing arts facilities"  
## [24,] "Auditoriums" "Auditoriums"   
## [25,] "Other" "Other"   
## Coal\_Quantity Wood\_Metrictonne   
## [1,] "Administrative office" "Administrative office"   
## [2,] "Library" "Library"   
## [3,] "Water treatment" "Water treatment"   
## [4,] "Water pumping" "Water pumping"   
## [5,] "Sewage treatment" "Sewage treatment"   
## [6,] "Sewage pumping" "Sewage pumping"   
## [7,] "Police station" "Police station"   
## [8,] "Fire station" "Fire station"   
## [9,] "Storage" "Storage"   
## [10,] "Community centres" "Community centres"   
## [11,] "Classrooms" "Classrooms"   
## [12,] "Hospital" "Hospital"   
## [13,] "Ambulance station" "Ambulance station"   
## [14,] "Laboratories" "Laboratories"   
## [15,] "Student residences" "Student residences"   
## [16,] "Recreational facilities" "Recreational facilities"   
## [17,] "Indoor swimming pools" "Indoor swimming pools"   
## [18,] "Indoor ice rinks" "Indoor ice rinks"   
## [19,] "Cultural facilities" "Cultural facilities"   
## [20,] "Parking" "Parking"   
## [21,] "School" "School"   
## [22,] "Art galleries" "Art galleries"   
## [23,] "Performing arts facilities" "Performing arts facilities"  
## [24,] "Auditoriums" "Auditoriums"   
## [25,] "Other" "Other"   
## DistrictHeating\_GJ DistrictHeating\_IsRenewable   
## [1,] "Administrative office" "Administrative office"   
## [2,] "Library" "Library"   
## [3,] "Water treatment" "Water treatment"   
## [4,] "Water pumping" "Water pumping"   
## [5,] "Sewage treatment" "Sewage treatment"   
## [6,] "Sewage pumping" "Sewage pumping"   
## [7,] "Police station" "Police station"   
## [8,] "Fire station" "Fire station"   
## [9,] "Storage" "Storage"   
## [10,] "Community centres" "Community centres"   
## [11,] "Classrooms" "Classrooms"   
## [12,] "Hospital" "Hospital"   
## [13,] "Ambulance station" "Ambulance station"   
## [14,] "Laboratories" "Laboratories"   
## [15,] "Student residences" "Student residences"   
## [16,] "Recreational facilities" "Recreational facilities"   
## [17,] "Indoor swimming pools" "Indoor swimming pools"   
## [18,] "Indoor ice rinks" "Indoor ice rinks"   
## [19,] "Cultural facilities" "Cultural facilities"   
## [20,] "Parking" "Parking"   
## [21,] "School" "School"   
## [22,] "Art galleries" "Art galleries"   
## [23,] "Performing arts facilities" "Performing arts facilities"  
## [24,] "Auditoriums" "Auditoriums"   
## [25,] "Other" "Other"   
## DistrictCooling\_GJ DistrictCooling\_IsRenewable   
## [1,] "Administrative office" "Administrative office"   
## [2,] "Library" "Library"   
## [3,] "Water treatment" "Water treatment"   
## [4,] "Water pumping" "Water pumping"   
## [5,] "Sewage treatment" "Sewage treatment"   
## [6,] "Sewage pumping" "Sewage pumping"   
## [7,] "Police station" "Police station"   
## [8,] "Fire station" "Fire station"   
## [9,] "Storage" "Storage"   
## [10,] "Community centres" "Community centres"   
## [11,] "Classrooms" "Classrooms"   
## [12,] "Hospital" "Hospital"   
## [13,] "Ambulance station" "Ambulance station"   
## [14,] "Laboratories" "Laboratories"   
## [15,] "Student residences" "Student residences"   
## [16,] "Recreational facilities" "Recreational facilities"   
## [17,] "Indoor swimming pools" "Indoor swimming pools"   
## [18,] "Indoor ice rinks" "Indoor ice rinks"   
## [19,] "Cultural facilities" "Cultural facilities"   
## [20,] "Parking" "Parking"   
## [21,] "School" "School"   
## [22,] "Art galleries" "Art galleries"   
## [23,] "Performing arts facilities" "Performing arts facilities"  
## [24,] "Auditoriums" "Auditoriums"   
## [25,] "Other" "Other"   
## GHGEmissions\_KG EnergyIntensityekWh\_sqft   
## [1,] "Administrative office" "Administrative office"   
## [2,] "Library" "Library"   
## [3,] "Water treatment" "Water treatment"   
## [4,] "Water pumping" "Water pumping"   
## [5,] "Sewage treatment" "Sewage treatment"   
## [6,] "Sewage pumping" "Sewage pumping"   
## [7,] "Police station" "Police station"   
## [8,] "Fire station" "Fire station"   
## [9,] "Storage" "Storage"   
## [10,] "Community centres" "Community centres"   
## [11,] "Classrooms" "Classrooms"   
## [12,] "Hospital" "Hospital"   
## [13,] "Ambulance station" "Ambulance station"   
## [14,] "Laboratories" "Laboratories"   
## [15,] "Student residences" "Student residences"   
## [16,] "Recreational facilities" "Recreational facilities"   
## [17,] "Indoor swimming pools" "Indoor swimming pools"   
## [18,] "Indoor ice rinks" "Indoor ice rinks"   
## [19,] "Cultural facilities" "Cultural facilities"   
## [20,] "Parking" "Parking"   
## [21,] "School" "School"   
## [22,] "Art galleries" "Art galleries"   
## [23,] "Performing arts facilities" "Performing arts facilities"  
## [24,] "Auditoriums" "Auditoriums"   
## [25,] "Other" "Other"   
## EnergyIntensity\_ekWh\_mega\_litre  
## [1,] "Administrative office"   
## [2,] "Library"   
## [3,] "Water treatment"   
## [4,] "Water pumping"   
## [5,] "Sewage treatment"   
## [6,] "Sewage pumping"   
## [7,] "Police station"   
## [8,] "Fire station"   
## [9,] "Storage"   
## [10,] "Community centres"   
## [11,] "Classrooms"   
## [12,] "Hospital"   
## [13,] "Ambulance station"   
## [14,] "Laboratories"   
## [15,] "Student residences"   
## [16,] "Recreational facilities"   
## [17,] "Indoor swimming pools"   
## [18,] "Indoor ice rinks"   
## [19,] "Cultural facilities"   
## [20,] "Parking"   
## [21,] "School"   
## [22,] "Art galleries"   
## [23,] "Performing arts facilities"   
## [24,] "Auditoriums"   
## [25,] "Other"

# Convert operation types to numerical values in new column TypeofOperation  
  
TypeofOperation<-ifelse(all\_data$OperationType=='Administrative office',1,  
ifelse(all\_data$OperationType=='Library',2,   
ifelse(all\_data$OperationType=='Water treatment',3,   
ifelse(all\_data$OperationType=='Water pumping',4,  
ifelse(all\_data$OperationType== 'Sewage treatment',5,  
ifelse(all\_data$OperationType== 'Sewage pumping', 6,  
ifelse(all\_data$OperationType== 'Police station',7,  
ifelse(all\_data$OperationType== 'Fire station',8,  
ifelse(all\_data$OperationType== 'Storage',9,  
ifelse(all\_data$OperationType== 'Community centres',10,  
ifelse(all\_data$OperationType== 'Classrooms', 11,  
ifelse(all\_data$OperationType== 'Hospital', 12,  
ifelse(all\_data$OperationType== 'Ambulance station', 13,  
ifelse(all\_data$OperationType== 'Laboratories', 14,  
ifelse(all\_data$OperationType== 'Student residences', 15,   
ifelse(all\_data$OperationType== 'Recreational facilities', 16,   
ifelse(all\_data$OperationType== 'School', 17,  
ifelse(all\_data$OperationType== 'Parking', 18,   
ifelse(all\_data$OperationType== 'Indoor swimming pools', 19,  
ifelse(all\_data$OperationType== 'Indoor ice rinks', 20,  
ifelse(all\_data$OperationType== 'Multi-use', 21,  
ifelse(all\_data$OperationType== 'Art galleries', 22,  
ifelse(all\_data$OperationType== 'Performing arts facilities', 23,   
ifelse(all\_data$OperationType== 'Auditoriums', 24,  
 ifelse(all\_data$OperationType== 'Other', 25 ,26  
 )))))))))))))))))))))))))  
  
#convert Operation to a factor  
TypeofOperation<- as.factor(TypeofOperation)  
  
#combine the Operation with the dataframe  
all\_data<-cbind(all\_data,TypeofOperation)  
  
#list the unique values in variable Sector  
sapply(all\_data, function(x) unique(all\_data$Sector))

## Year   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## Sector   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## Organization   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## Operation   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## OperationType   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## Address   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## City   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## PostalCode   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## TotalIndoorSpace\_sqft   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## WeeklyAverageHours   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## AnnualFlow\_M   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## NumberofPortables   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## SwimmingPool   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## Electricity\_kwh   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## NaturalGas\_Cubicmeter   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## FuelOil12\_L   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## FuelOil46\_L   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## Propane\_Litre   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## Coal\_Quantity   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## Wood\_Metrictonne   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## DistrictHeating\_GJ   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## DistrictHeating\_IsRenewable   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## DistrictCooling\_GJ   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## DistrictCooling\_IsRenewable   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## GHGEmissions\_KG   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## EnergyIntensityekWh\_sqft   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## EnergyIntensity\_ekWh\_mega\_litre   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"   
## TypeofOperation   
## [1,] "Municipal"   
## [2,] "Post-Secondary Educational Institution"  
## [3,] "Public Hospital"   
## [4,] "School Board"

# Convert Sector to numerical values in new column TypeofSector  
TypeofSector<-ifelse(all\_data$Sector=='Post-Secondary Educational Institution',1,  
ifelse(all\_data$Sector=='Public Hospital',2,   
 ifelse(all\_data$Sector=='Municipal',3,   
 ifelse(all\_data$Sector=='School Board',4,5))))  
  
#convert Operation to a factor  
TypeofSector<- as.factor(TypeofSector)  
  
  
#combine the Operation with the dataframe  
all\_data<-cbind(all\_data,TypeofSector)  
  
# 88259 obs. of 29 variables   
str(all\_data)

## 'data.frame': 88259 obs. of 29 variables:  
## $ Year : Factor w/ 5 levels "2011","2012",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Sector : Factor w/ 4 levels "Municipal","Post-Secondary Educational Institution",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Organization : Factor w/ 825 levels "Alexandra Hospital",..: 401 401 401 401 401 401 401 401 401 401 ...  
## $ Operation : Factor w/ 22856 levels "'A' Block","'B' Block",..: 8756 5424 14557 4040 10818 14509 11638 5425 4562 591 ...  
## $ OperationType : Factor w/ 25 levels "Administrative office",..: 1 13 24 23 23 20 19 16 8 21 ...  
## $ Address : Factor w/ 19780 levels "#1 Hwy 527, c/o Box 98, 112 Queen St.",..: 11332 10812 428 5543 8772 5285 14948 10528 6196 8193 ...  
## $ City : Factor w/ 1429 levels "568, rue Lessard",..: 429 429 429 429 429 429 429 429 429 429 ...  
## $ PostalCode : Factor w/ 19483 levels "H0H 0H0","K0A 1A0",..: 7747 7747 7747 7747 7747 7747 7747 7747 7747 7747 ...  
## $ TotalIndoorSpace\_sqft : num 21510 11840 7890 678 1981 ...  
## $ WeeklyAverageHours : num 40 62 168 168 168 168 168 40 15 20 ...  
## $ AnnualFlow\_M : num 0 0 1406 1 1 ...  
## $ NumberofPortables : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ SwimmingPool : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Electricity\_kwh : num 248444 100147 994783 15519 123972 ...  
## $ NaturalGas\_Cubicmeter : num 203 18695 0 0 0 ...  
## $ FuelOil12\_L : num 0 0 0 0 0 ...  
## $ FuelOil46\_L : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Propane\_Litre : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Coal\_Quantity : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Wood\_Metrictonne : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DistrictHeating\_GJ : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DistrictHeating\_IsRenewable : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ DistrictCooling\_GJ : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DistrictCooling\_IsRenewable : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ GHGEmissions\_KG : num 24740 45163 97529 1521 12154 ...  
## $ EnergyIntensityekWh\_sqft : num 11.7 25.2 126.1 22.9 62.6 ...  
## $ EnergyIntensity\_ekWh\_mega\_litre: num 0 0 708 15519 123972 ...  
## $ TypeofOperation : Factor w/ 25 levels "1","2","3","4",..: 1 2 3 4 4 5 6 7 8 9 ...  
## $ TypeofSector : Factor w/ 4 levels "1","2","3","4": 3 3 3 3 3 3 3 3 3 3 ...

#write back to disk file will contain all 5 years of data if want to verify data  
# commented out code below  
#write.csv(all\_data, file = "C:/RY/CKME136/OntarioGHG data/FinalData/All\_Data.csv")

# Load the libraries

# When Load packages - packages will be stored in C:\software\packages because set env variable R\_LIBS\_USER  
  
#install.packages("dplyr")  
library(plyr)  
library(dplyr)

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

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

#install.packages("ggplot2")  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.5.1

#install.packages("corrplot")  
library(corrplot)

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

## corrplot 0.84 loaded

# Exploratory phase

# Normalize the data (normalize numerical columns)

#first remove rows that have value less than 1 KG for GHG emisssions  
#now have 87774 obs. of 29 variables  
all\_data<-all\_data[all\_data$GHGEmissions\_KG>1,]  
  
str(all\_data)

## 'data.frame': 87774 obs. of 29 variables:  
## $ Year : Factor w/ 5 levels "2011","2012",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Sector : Factor w/ 4 levels "Municipal","Post-Secondary Educational Institution",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Organization : Factor w/ 825 levels "Alexandra Hospital",..: 401 401 401 401 401 401 401 401 401 401 ...  
## $ Operation : Factor w/ 22856 levels "'A' Block","'B' Block",..: 8756 5424 14557 4040 10818 14509 11638 5425 4562 591 ...  
## $ OperationType : Factor w/ 25 levels "Administrative office",..: 1 13 24 23 23 20 19 16 8 21 ...  
## $ Address : Factor w/ 19780 levels "#1 Hwy 527, c/o Box 98, 112 Queen St.",..: 11332 10812 428 5543 8772 5285 14948 10528 6196 8193 ...  
## $ City : Factor w/ 1429 levels "568, rue Lessard",..: 429 429 429 429 429 429 429 429 429 429 ...  
## $ PostalCode : Factor w/ 19483 levels "H0H 0H0","K0A 1A0",..: 7747 7747 7747 7747 7747 7747 7747 7747 7747 7747 ...  
## $ TotalIndoorSpace\_sqft : num 21510 11840 7890 678 1981 ...  
## $ WeeklyAverageHours : num 40 62 168 168 168 168 168 40 15 20 ...  
## $ AnnualFlow\_M : num 0 0 1406 1 1 ...  
## $ NumberofPortables : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ SwimmingPool : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Electricity\_kwh : num 248444 100147 994783 15519 123972 ...  
## $ NaturalGas\_Cubicmeter : num 203 18695 0 0 0 ...  
## $ FuelOil12\_L : num 0 0 0 0 0 ...  
## $ FuelOil46\_L : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Propane\_Litre : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Coal\_Quantity : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Wood\_Metrictonne : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DistrictHeating\_GJ : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DistrictHeating\_IsRenewable : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ DistrictCooling\_GJ : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DistrictCooling\_IsRenewable : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ GHGEmissions\_KG : num 24740 45163 97529 1521 12154 ...  
## $ EnergyIntensityekWh\_sqft : num 11.7 25.2 126.1 22.9 62.6 ...  
## $ EnergyIntensity\_ekWh\_mega\_litre: num 0 0 708 15519 123972 ...  
## $ TypeofOperation : Factor w/ 25 levels "1","2","3","4",..: 1 2 3 4 4 5 6 7 8 9 ...  
## $ TypeofSector : Factor w/ 4 levels "1","2","3","4": 3 3 3 3 3 3 3 3 3 3 ...

summary(all\_data)

## Year Sector   
## 2011:18540 Municipal :55951   
## 2012:17947 Post-Secondary Educational Institution: 4545   
## 2013:16927 Public Hospital : 2262   
## 2014:17007 School Board :25016   
## 2015:17353   
##   
##   
## Organization Operation   
## City of Toronto : 4276 Municipal Office : 334   
## Toronto District School Board : 2866 Water Treatment Plant: 270   
## City of Greater Sudbury : 1454 Fire Hall : 260   
## Peel District School Board : 1280 Public Works Garage : 203   
## City of Ottawa : 1243 Town Hall : 177   
## York Region District School Board: 1098 Library : 161   
## (Other) :75557 (Other) :86369   
## OperationType Address   
## School :23834 Intentionally Omitted : 465   
## Administrative office: 9860 public works related facility: 368   
## Storage : 6733 4700 Keele Street : 330   
## Fire station : 5903 1385 Woodroffe Avenue : 271   
## Community centres : 5675 1001 Fanshawe College Blvd. : 188   
## Sewage pumping : 4904 (Other) :86135   
## (Other) :30865 NA's : 17   
## City PostalCode TotalIndoorSpace\_sqft  
## Toronto : 8738 M3J 1P3: 269 Min. :0.000e+00   
## Ottawa : 2876 N5Y 5R6: 252 1st Qu.:1.800e+03   
## Brampton : 2097 L7B 1B3: 225 Median :9.200e+03   
## Mississauga: 1964 L6P 1K7: 213 Mean :8.577e+04   
## London : 1434 K2G 1V8: 203 3rd Qu.:3.828e+04   
## Hamilton : 1420 (Other):86595 Max. :4.262e+09   
## (Other) :69245 NA's : 17   
## WeeklyAverageHours AnnualFlow\_M NumberofPortables SwimmingPool  
## Min. : 0.00 Min. :0.000e+00 Min. : 0.0000 0:87582   
## 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 : 82.61 Mean :7.260e+04 Mean : 0.4624   
## 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.:2.336e+04 1st Qu.: 0.000e+00 1st Qu.:0.000e+00   
## Median :1.098e+05 Median : 9.718e+03 Median :0.000e+00   
## Mean :4.399e+06 Mean : 7.459e+05 Mean :3.815e+06   
## 3rd Qu.:3.444e+05 3rd Qu.: 4.521e+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 :6.816e+05 Mean : 720.6 Mean : 1.99   
## 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 0:87753   
## 1st Qu.: 0.00 1st Qu.: 0.0 1: 21   
## Median : 0.00 Median : 0.0   
## Mean : 3.25 Mean : 161.9   
## 3rd Qu.: 0.00 3rd Qu.: 0.0   
## Max. :37347.42 Max. :1463421.2   
##   
## DistrictCooling\_GJ DistrictCooling\_IsRenewable GHGEmissions\_KG   
## Min. : 0.00 0:87758 Min. :1.000e+00   
## 1st Qu.: 0.00 1: 16 1st Qu.:6.965e+03   
## Median : 0.00 Median :3.575e+04   
## Mean : 21.51 Mean :1.441e+07   
## 3rd Qu.: 0.00 3rd Qu.:1.146e+05   
## Max. :147285.75 Max. :2.800e+11   
##   
## EnergyIntensityekWh\_sqft EnergyIntensity\_ekWh\_mega\_litre TypeofOperation  
## Min. :0.000e+00 Min. :0.000e+00 17 :23834   
## 1st Qu.:1.300e+01 1st Qu.:0.000e+00 1 : 9860   
## Median :2.100e+01 Median :0.000e+00 9 : 6733   
## Mean :3.299e+06 Mean :1.465e+14 8 : 5903   
## 3rd Qu.:3.300e+01 3rd Qu.:0.000e+00 10 : 5675   
## Max. :2.860e+11 Max. :8.430e+18 6 : 4904   
## (Other):30865   
## TypeofSector  
## 1: 4545   
## 2: 2262   
## 3:55951   
## 4:25016   
##   
##   
##

# normalize the data  
normalize <- function(x){  
 return ((x-min(x))/(max(x) - min(x)))  
}  
  
all\_data\_n<- all\_data %>%  
 mutate\_if(is.numeric, normalize)

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

str(all\_data\_n)

## 'data.frame': 87774 obs. of 29 variables:  
## $ Year : Factor w/ 5 levels "2011","2012",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Sector : Factor w/ 4 levels "Municipal","Post-Secondary Educational Institution",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Organization : Factor w/ 825 levels "Alexandra Hospital",..: 401 401 401 401 401 401 401 401 401 401 ...  
## $ Operation : Factor w/ 22856 levels "'A' Block","'B' Block",..: 8756 5424 14557 4040 10818 14509 11638 5425 4562 591 ...  
## $ OperationType : Factor w/ 25 levels "Administrative office",..: 1 13 24 23 23 20 19 16 8 21 ...  
## $ Address : Factor w/ 19780 levels "#1 Hwy 527, c/o Box 98, 112 Queen St.",..: 11332 10812 428 5543 8772 5285 14948 10528 6196 8193 ...  
## $ City : Factor w/ 1429 levels "568, rue Lessard",..: 429 429 429 429 429 429 429 429 429 429 ...  
## $ PostalCode : Factor w/ 19483 levels "H0H 0H0","K0A 1A0",..: 7747 7747 7747 7747 7747 7747 7747 7747 7747 7747 ...  
## $ TotalIndoorSpace\_sqft : num 5.05e-06 2.78e-06 1.85e-06 1.59e-07 4.65e-07 ...  
## $ WeeklyAverageHours : num 0.00457 0.00708 0.01918 0.01918 0.01918 ...  
## $ AnnualFlow\_M : num 0.00 0.00 1.20e-06 8.54e-10 8.54e-10 ...  
## $ NumberofPortables : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ SwimmingPool : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Electricity\_kwh : num 3.22e-06 1.30e-06 1.29e-05 2.01e-07 1.61e-06 ...  
## $ NaturalGas\_Cubicmeter : num 6.87e-08 1.70e-06 5.08e-08 5.08e-08 5.08e-08 ...  
## $ FuelOil12\_L : num 0 0 0 0 0 ...  
## $ FuelOil46\_L : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Propane\_Litre : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Coal\_Quantity : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Wood\_Metrictonne : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DistrictHeating\_GJ : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DistrictHeating\_IsRenewable : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ DistrictCooling\_GJ : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DistrictCooling\_IsRenewable : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ GHGEmissions\_KG : num 8.84e-08 1.61e-07 3.48e-07 5.43e-09 4.34e-08 ...  
## $ EnergyIntensityekWh\_sqft : num 4.07e-11 8.82e-11 4.41e-10 8.00e-11 2.19e-10 ...  
## $ EnergyIntensity\_ekWh\_mega\_litre: num 0.00 0.00 8.40e-17 1.84e-15 1.47e-14 ...  
## $ TypeofOperation : Factor w/ 25 levels "1","2","3","4",..: 1 2 3 4 4 5 6 7 8 9 ...  
## $ TypeofSector : Factor w/ 4 levels "1","2","3","4": 3 3 3 3 3 3 3 3 3 3 ...

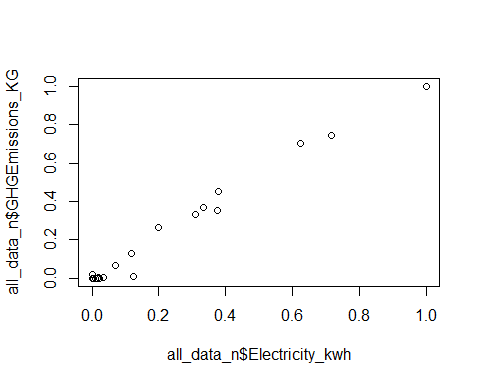
summary(all\_data\_n)

## Year Sector   
## 2011:18540 Municipal :55951   
## 2012:17947 Post-Secondary Educational Institution: 4545   
## 2013:16927 Public Hospital : 2262   
## 2014:17007 School Board :25016   
## 2015:17353   
##   
##   
## Organization Operation   
## City of Toronto : 4276 Municipal Office : 334   
## Toronto District School Board : 2866 Water Treatment Plant: 270   
## City of Greater Sudbury : 1454 Fire Hall : 260   
## Peel District School Board : 1280 Public Works Garage : 203   
## City of Ottawa : 1243 Town Hall : 177   
## York Region District School Board: 1098 Library : 161   
## (Other) :75557 (Other) :86369   
## OperationType Address   
## School :23834 Intentionally Omitted : 465   
## Administrative office: 9860 public works related facility: 368   
## Storage : 6733 4700 Keele Street : 330   
## Fire station : 5903 1385 Woodroffe Avenue : 271   
## Community centres : 5675 1001 Fanshawe College Blvd. : 188   
## Sewage pumping : 4904 (Other) :86135   
## (Other) :30865 NA's : 17   
## City PostalCode TotalIndoorSpace\_sqft  
## Toronto : 8738 M3J 1P3: 269 Min. :0.00e+00   
## Ottawa : 2876 N5Y 5R6: 252 1st Qu.:4.00e-07   
## Brampton : 2097 L7B 1B3: 225 Median :2.20e-06   
## Mississauga: 1964 L6P 1K7: 213 Mean :2.01e-05   
## London : 1434 K2G 1V8: 203 3rd Qu.:9.00e-06   
## Hamilton : 1420 (Other):86595 Max. :1.00e+00   
## (Other) :69245 NA's : 17   
## WeeklyAverageHours AnnualFlow\_M NumberofPortables SwimmingPool  
## Min. :0.000000 Min. :0.0e+00 Min. :0.00000 0:87582   
## 1st Qu.:0.004566 1st Qu.:0.0e+00 1st Qu.:0.00000 1: 192   
## Median :0.006849 Median :0.0e+00 Median :0.00000   
## Mean :0.009431 Mean :6.2e-05 Mean :0.01401   
## 3rd Qu.:0.014384 3rd Qu.:0.0e+00 3rd Qu.:0.00000   
## Max. :1.000000 Max. :1.0e+00 Max. :1.00000   
##   
## Electricity\_kwh NaturalGas\_Cubicmeter FuelOil12\_L   
## Min. :0.00e+00 Min. :0.0e+00 Min. :0.00e+00   
## 1st Qu.:3.00e-07 1st Qu.:1.0e-07 1st Qu.:0.00e+00   
## Median :1.40e-06 Median :9.0e-07 Median :0.00e+00   
## Mean :5.71e-05 Mean :6.6e-05 Mean :4.79e-05   
## 3rd Qu.:4.50e-06 3rd Qu.:4.0e-06 3rd Qu.:0.00e+00   
## Max. :1.00e+00 Max. :1.0e+00 Max. :1.00e+00   
##   
## FuelOil46\_L Propane\_Litre Coal\_Quantity   
## Min. :0.00e+00 Min. :0.0000000 Min. :0.00e+00   
## 1st Qu.:0.00e+00 1st Qu.:0.0000000 1st Qu.:0.00e+00   
## Median :0.00e+00 Median :0.0000000 Median :0.00e+00   
## Mean :5.86e-05 Mean :0.0006269 Mean :5.93e-05   
## 3rd Qu.:0.00e+00 3rd Qu.:0.0000000 3rd Qu.:0.00e+00   
## Max. :1.00e+00 Max. :1.0000000 Max. :1.00e+00   
##   
## Wood\_Metrictonne DistrictHeating\_GJ DistrictHeating\_IsRenewable  
## Min. :0.0e+00 Min. :0.0000000 0:87753   
## 1st Qu.:0.0e+00 1st Qu.:0.0000000 1: 21   
## Median :0.0e+00 Median :0.0000000   
## Mean :8.7e-05 Mean :0.0001106   
## 3rd Qu.:0.0e+00 3rd Qu.:0.0000000   
## Max. :1.0e+00 Max. :1.0000000   
##   
## DistrictCooling\_GJ DistrictCooling\_IsRenewable GHGEmissions\_KG   
## Min. :0.000000 0:87758 Min. :0.00e+00   
## 1st Qu.:0.000000 1: 16 1st Qu.:0.00e+00   
## Median :0.000000 Median :1.00e-07   
## Mean :0.000146 Mean :5.15e-05   
## 3rd Qu.:0.000000 3rd Qu.:4.00e-07   
## Max. :1.000000 Max. :1.00e+00   
##   
## EnergyIntensityekWh\_sqft EnergyIntensity\_ekWh\_mega\_litre TypeofOperation  
## Min. :0.00e+00 Min. :0.00e+00 17 :23834   
## 1st Qu.:0.00e+00 1st Qu.:0.00e+00 1 : 9860   
## Median :0.00e+00 Median :0.00e+00 9 : 6733   
## Mean :1.15e-05 Mean :1.74e-05 8 : 5903   
## 3rd Qu.:0.00e+00 3rd Qu.:0.00e+00 10 : 5675   
## Max. :1.00e+00 Max. :1.00e+00 6 : 4904   
## (Other):30865   
## TypeofSector  
## 1: 4545   
## 2: 2262   
## 3:55951   
## 4:25016   
##   
##   
##

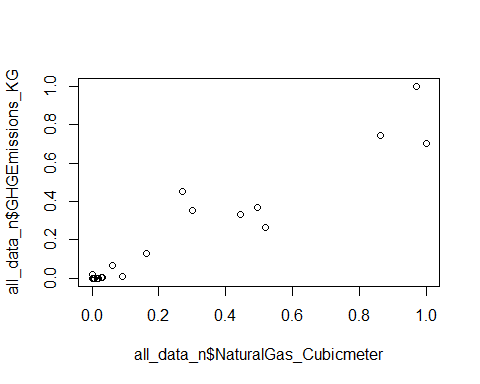
# Exploratory phase

# Plotting the normalized data to look at correlations and outliers

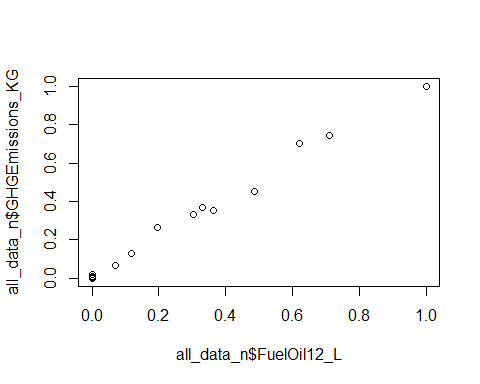
####### Plots  
# plots below show attributes which are correlated to GHG  
plot(x=all\_data\_n$Electricity\_kwh, y=all\_data\_n$GHGEmissions\_KG)



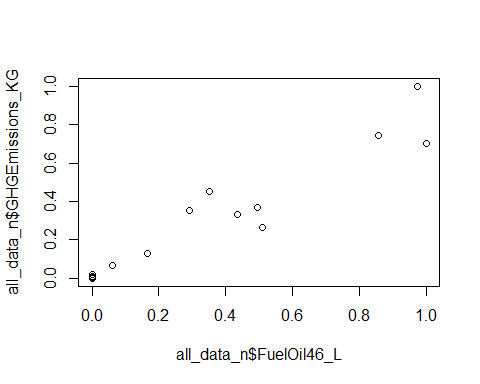
plot(x=all\_data\_n$NaturalGas\_Cubicmeter, y=all\_data\_n$GHGEmissions\_KG)



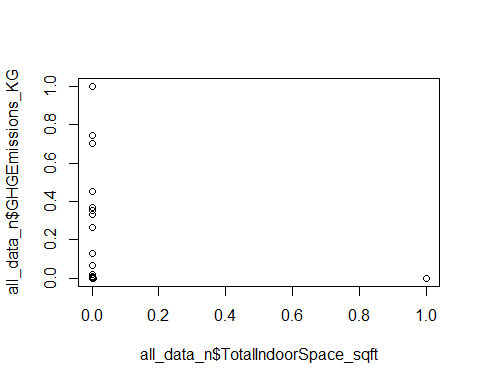
plot(x=all\_data\_n$FuelOil12\_L, y=all\_data\_n$GHGEmissions\_KG)



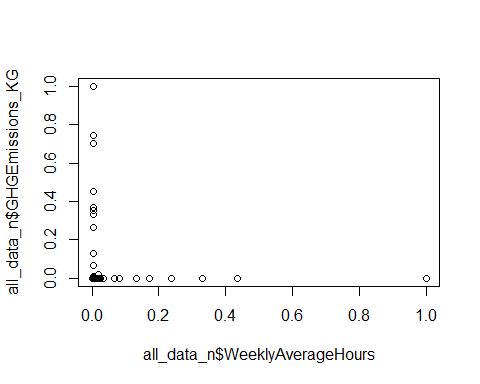
plot(x=all\_data\_n$FuelOil46\_L, y=all\_data\_n$GHGEmissions\_KG)



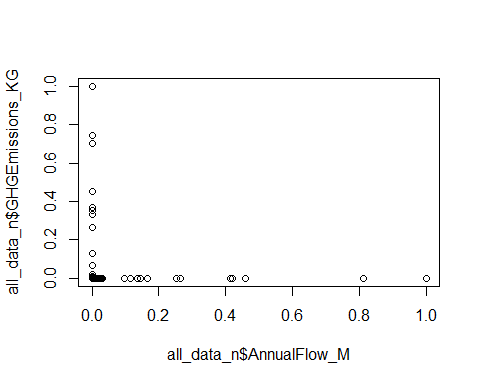
# plots below show attributes which are not correlated to GHG  
plot(x=all\_data\_n$TotalIndoorSpace\_sqft, y=all\_data\_n$GHGEmissions\_KG)



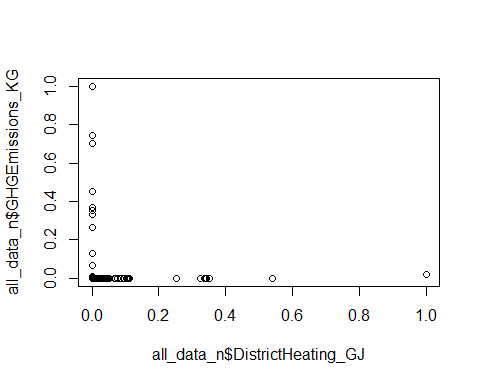
plot(x=all\_data\_n$WeeklyAverageHours, y=all\_data\_n$GHGEmissions\_KG)



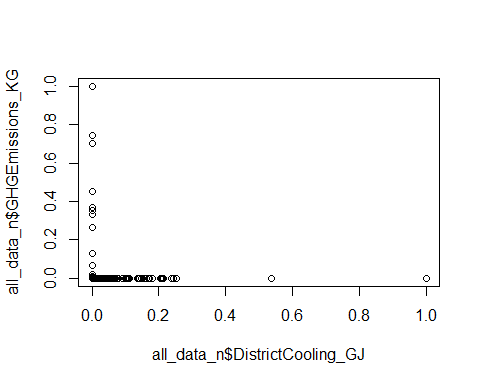
plot(x=all\_data\_n$AnnualFlow\_M , y=all\_data\_n$GHGEmissions\_KG)



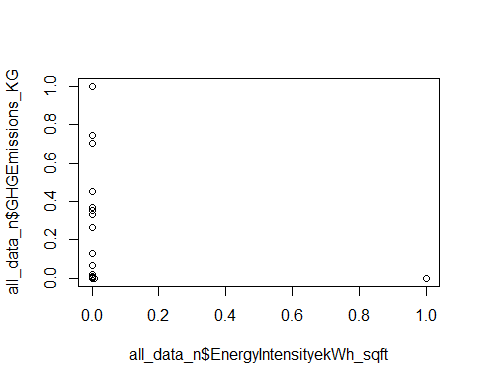
plot(x=all\_data\_n$DistrictHeating\_GJ , y=all\_data\_n$GHGEmissions\_KG)



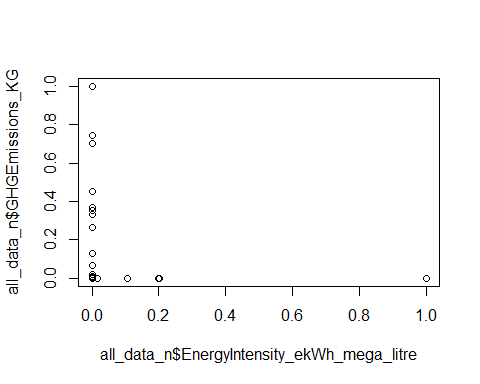
plot(x=all\_data\_n$DistrictCooling\_GJ, y=all\_data\_n$GHGEmissions\_KG)



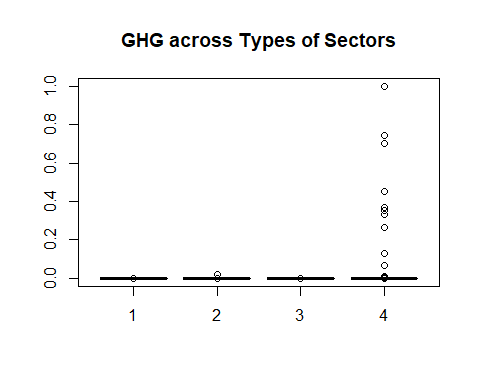
plot(x=all\_data\_n$EnergyIntensityekWh\_sqft, y=all\_data\_n$GHGEmissions\_KG)



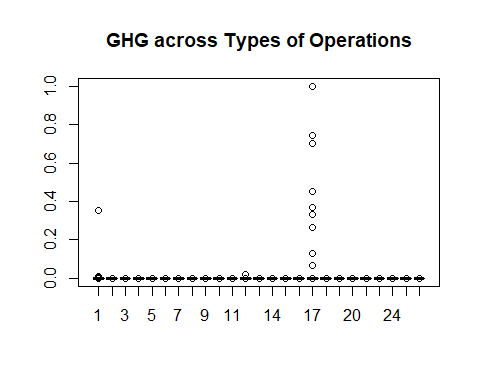
plot(x=all\_data\_n$EnergyIntensity\_ekWh\_mega\_litre, y=all\_data\_n$GHGEmissions\_KG)



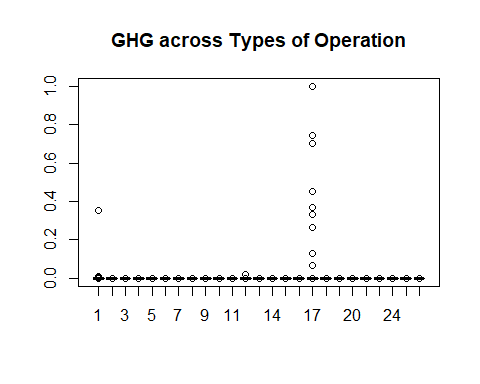
# plot of Sector and OPeration type with GHG  
plot(x=all\_data\_n$TypeofSector, y=all\_data\_n$GHGEmissions\_KG, main="GHG across Types of Sectors")



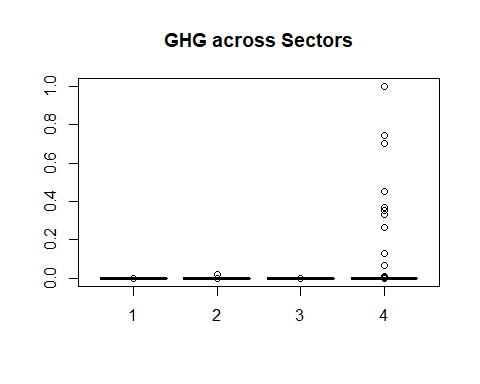
# Sector 4 Schoolboard has a big spread while other 3 sectors are around 0  
plot(x=all\_data\_n$TypeofOperation, y=all\_data\_n$GHGEmissions\_KG,main="GHG across Types of Operations")



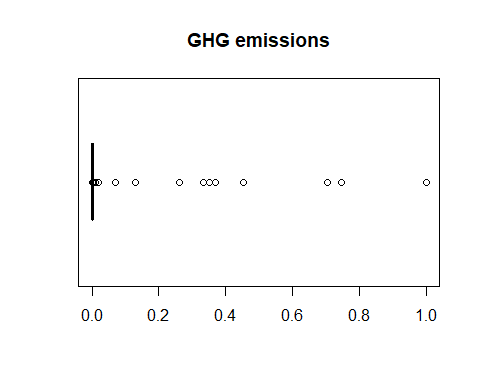
#TypeofOperation 17 (School) big spread, has greatest GHG followed by 1 (Administrative Office)  
  
######## Boxplots  
boxplot(all\_data\_n$GHGEmissions\_KG ~ all\_data\_n$TypeofOperation, data=all\_data\_n, main="GHG across Types of Operation")



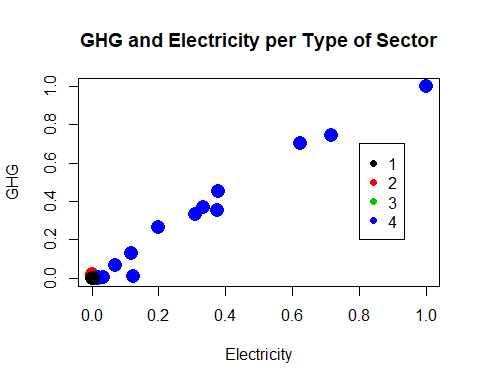
boxplot(all\_data\_n$GHGEmissions\_KG ~ all\_data\_n$TypeofSector, data=all\_data\_n, main="GHG across Sectors")



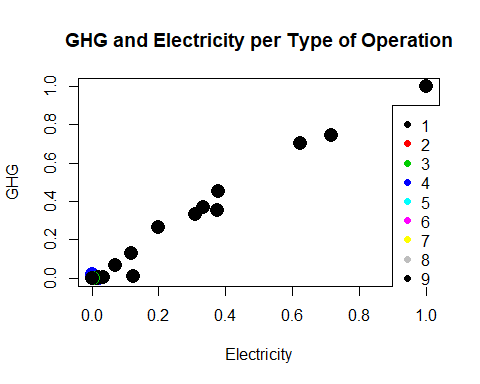
boxplot(all\_data\_n$GHGEmissions\_KG, horizontal=TRUE,main="GHG emissions")



# GHG and type of sector  
plot(all\_data\_n$Electricity\_kwh, all\_data\_n$GHGEmissions\_KG, # x variable, y variable  
 col = all\_data\_n$TypeofSector, # colour by TypeofSector  
 pch = 16, # type of point to use  
 cex = 2, # size of point to use  
 xlab = "Electricity", # x axis label  
 ylab = "GHG", # y axis label  
 main = "GHG and Electricity per Type of Sector") # plot title  
legend (x = .8, y = .7, legend = levels(all\_data\_n$TypeofSector), col = c(1:4), pch = 16)



# GHG and type of Operation  
plot(all\_data\_n$Electricity\_kwh, all\_data\_n$GHGEmissions\_KG, # x variable, y variable  
 col = all\_data\_n$TypeofOperation, # colour by TypeofOperation  
 pch = 16, # type of point to use  
 cex = 2, # size of point to use  
 xlab = "Electricity", # x axis label  
 ylab = "GHG", # y axis label  
 main = "GHG and Electricity per Type of Operation") # plot title  
legend (x = .9, y = .9, legend = levels(all\_data\_n$TypeofOperation), col = c(1:26), pch = 16)

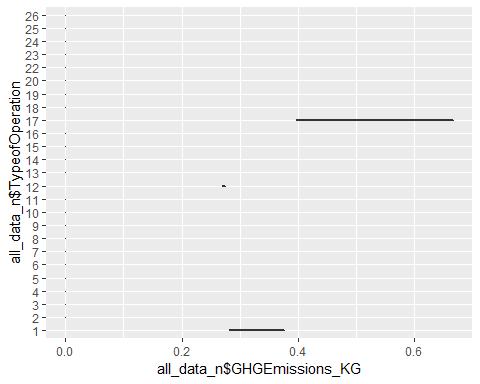


# following shows Type of Operation 17 (School) has highest GHG followed by 1 (Administrative Office)  
p<-ggplot(all\_data\_n, aes( all\_data\_n$GHGEmissions\_KG, all\_data\_n$TypeofOperation))  
  
geom\_boxplot(mapping = NULL, data = NULL, stat = "boxplot",  
 position = "dodge2",all\_data\_n, outlier.colour = NULL,  
 outlier.color = NULL, outlier.fill = NULL, outlier.shape = 19,  
 outlier.size = 1.5, outlier.stroke = 0.5, outlier.alpha = NULL,  
 notch = FALSE, notchwidth = 0.5, varwidth = FALSE, na.rm =TRUE,  
 show.legend = NA, inherit.aes = TRUE)

## Warning: Ignoring unknown parameters:

## geom\_boxplot: outlier.colour = NULL, outlier.fill = NULL, outlier.shape = 19, outlier.size = 1.5, outlier.stroke = 0.5, outlier.alpha = NULL, notch = FALSE, notchwidth = 0.5, varwidth = FALSE, na.rm = TRUE  
## stat\_boxplot: na.rm = TRUE  
## position\_dodge2

p + geom\_boxplot()

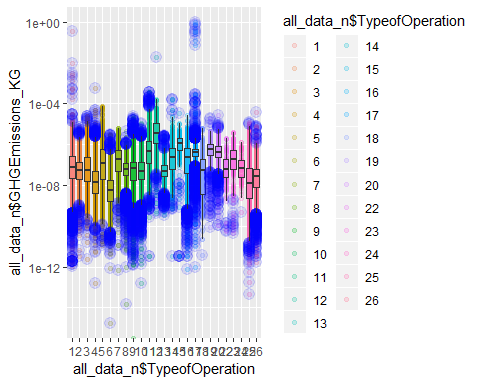


# using log scale for GHG - plot GHG versus Type of Operation,blue shows the outliers  
ggplot(data = all\_data\_n, aes(x=all\_data\_n$TypeofOperation, y=all\_data\_n$GHGEmissions\_KG)) +   
 scale\_y\_log10() +  
 geom\_point(aes(color=all\_data\_n$TypeofOperation), alpha=0.2) +  
 geom\_boxplot(outlier.size=4, outlier.colour='blue', alpha=0.1)

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 1 rows containing non-finite values (stat\_boxplot).

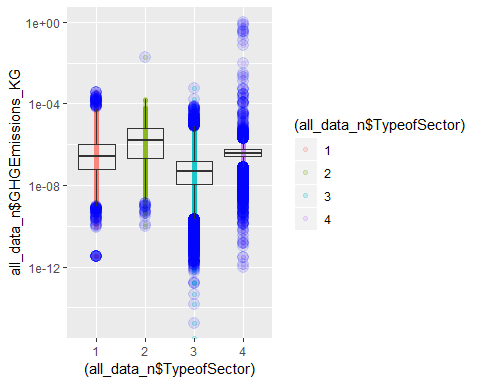


# using log scale for GHG - plot GHG versus Sector,blue shows the outliers  
ggplot(data = all\_data\_n, aes(x=(all\_data\_n$TypeofSector), y=all\_data\_n$GHGEmissions\_KG)) +   
 scale\_y\_log10() +  
 geom\_point(aes(color=(all\_data\_n$TypeofSector)), alpha=0.2) +  
 geom\_boxplot(outlier.size=4, outlier.colour='blue', alpha=0.1)

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 1 rows containing non-finite values (stat\_boxplot).



# Exploratory phase

# Summaries for GHG by Sector and Type of operation

# Feature optimization - using Principal component analysis (using numerical columns of normalized data excluding GHG)

#normalized data - look at spread of values in the variables  
#library(dplyr)  
  
#create table of counts of records for each OperationType arrange by descending counts  
# Results show that operation type school has the highest numbeof rows  
all\_data\_n %>%  
group\_by(OperationType) %>%  
summarise(freq=n(), totalGHG=sum(GHGEmissions\_KG)) %>%  
arrange(desc(freq))

## # A tibble: 25 x 3  
## OperationType freq totalGHG  
## <fct> <int> <dbl>  
## 1 School 23834 4.08   
## 2 Administrative office 9860 0.376   
## 3 Storage 6733 0.00154   
## 4 Fire station 5903 0.000558  
## 5 Community centres 5675 0.00101   
## 6 Sewage pumping 4904 0.000153  
## 7 Water pumping 4161 0.00137   
## 8 Library 3324 0.000538  
## 9 Other 3309 0.000571  
## 10 Recreational facilities 3086 0.00182   
## # ... with 15 more rows

#create table of counts of records for each OperationType arrange by descending GHGEmissions.Kg  
# Results show that operation type school has the highest TotalGHG  
all\_data\_n %>%  
group\_by(OperationType) %>%  
summarise(freq=n(), totalGHG=sum(GHGEmissions\_KG)) %>%  
arrange(desc(totalGHG))

## # A tibble: 25 x 3  
## OperationType freq totalGHG  
## <fct> <int> <dbl>  
## 1 School 23834 4.08   
## 2 Administrative office 9860 0.376   
## 3 Hospital 1426 0.0323   
## 4 Classrooms 1503 0.00945  
## 5 Sewage treatment 1857 0.00223  
## 6 Recreational facilities 3086 0.00182  
## 7 Storage 6733 0.00154  
## 8 Indoor ice rinks 1964 0.00150  
## 9 Water pumping 4161 0.00137  
## 10 Laboratories 802 0.00115  
## # ... with 15 more rows

#create table of counts of records for each Sector arrange by descending GHGEmissions.Kg  
# Results show that sector School Board has the highest numbeof rows  
all\_data\_n %>%  
group\_by(Sector) %>%  
summarise(freq=n(), totalGHG=sum(GHGEmissions\_KG)) %>%  
arrange(desc(totalGHG))

## # A tibble: 4 x 3  
## Sector freq totalGHG  
## <fct> <int> <dbl>  
## 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

####### Principal Component Analysis for feature selection  
  
# PCA including GHG  
str(all\_data\_n[,c(9:12,14:21, 23,25:27)])

## 'data.frame': 87774 obs. of 16 variables:  
## $ TotalIndoorSpace\_sqft : num 5.05e-06 2.78e-06 1.85e-06 1.59e-07 4.65e-07 ...  
## $ WeeklyAverageHours : num 0.00457 0.00708 0.01918 0.01918 0.01918 ...  
## $ AnnualFlow\_M : num 0.00 0.00 1.20e-06 8.54e-10 8.54e-10 ...  
## $ NumberofPortables : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Electricity\_kwh : num 3.22e-06 1.30e-06 1.29e-05 2.01e-07 1.61e-06 ...  
## $ NaturalGas\_Cubicmeter : num 6.87e-08 1.70e-06 5.08e-08 5.08e-08 5.08e-08 ...  
## $ FuelOil12\_L : num 0 0 0 0 0 ...  
## $ FuelOil46\_L : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Propane\_Litre : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Coal\_Quantity : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Wood\_Metrictonne : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DistrictHeating\_GJ : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DistrictCooling\_GJ : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ GHGEmissions\_KG : num 8.84e-08 1.61e-07 3.48e-07 5.43e-09 4.34e-08 ...  
## $ EnergyIntensityekWh\_sqft : num 4.07e-11 8.82e-11 4.41e-10 8.00e-11 2.19e-10 ...  
## $ EnergyIntensity\_ekWh\_mega\_litre: num 0.00 0.00 8.40e-17 1.84e-15 1.47e-14 ...

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

# 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  
# 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

``` # Prediction Stage

# Divide the normalized data into training and testing sets 70:30

train\_index <-sample (1:nrow(all\_data\_n), 0.7\*nrow(all\_data\_n))  
train.set <- all\_data\_n[train\_index,]  
test.set <- all\_data\_n[-train\_index,]  
  
str(train.set)

## 'data.frame': 61441 obs. of 29 variables:  
## $ Year : Factor w/ 5 levels "2011","2012",..: 5 1 2 1 5 4 4 2 5 1 ...  
## $ Sector : Factor w/ 4 levels "Municipal","Post-Secondary Educational Institution",..: 3 1 4 1 1 4 4 1 1 1 ...  
## $ Organization : Factor w/ 825 levels "Alexandra Hospital",..: 375 231 680 412 56 94 292 580 435 45 ...  
## $ Operation : Factor w/ 22856 levels "'A' Block","'B' Block",..: 20112 8740 16416 9398 6971 16350 11347 3798 14762 2095 ...  
## $ OperationType : Factor w/ 25 levels "Administrative office",..: 25 8 18 1 6 18 18 6 23 8 ...  
## $ Address : Factor w/ 19780 levels "#1 Hwy 527, c/o Box 98, 112 Queen St.",..: 13530 12425 15406 5350 6276 1169 6676 12809 1719 12086 ...  
## $ City : Factor w/ 1429 levels "568, rue Lessard",..: 1112 595 650 527 801 940 125 804 800 185 ...  
## $ PostalCode : Factor w/ 19483 levels "H0H 0H0","K0A 1A0",..: 5733 7946 10673 4487 2554 7843 3585 10303 4577 8344 ...  
## $ TotalIndoorSpace\_sqft : num 7.51e-07 1.12e-06 8.77e-06 2.50e-06 4.70e-07 ...  
## $ WeeklyAverageHours : num 5.71e-03 1.92e-02 4.57e-03 2.28e-05 3.68e-03 ...  
## $ AnnualFlow\_M : num 0 0 0 0 0 ...  
## $ NumberofPortables : num 0 0 0.152 0 0 ...  
## $ SwimmingPool : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Electricity\_kwh : num 7.12e-07 1.09e-06 3.78e-06 1.06e-06 6.57e-07 ...  
## $ NaturalGas\_Cubicmeter : num 7.34e-07 8.60e-07 4.07e-06 5.50e-07 5.08e-08 ...  
## $ FuelOil12\_L : num 0 0 0 0 0 ...  
## $ FuelOil46\_L : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Propane\_Litre : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Coal\_Quantity : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Wood\_Metrictonne : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DistrictHeating\_GJ : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DistrictHeating\_IsRenewable : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ DistrictCooling\_GJ : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DistrictCooling\_IsRenewable : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ GHGEmissions\_KG : num 6.01e-08 9.12e-08 4.07e-07 6.68e-08 7.33e-09 ...  
## $ EnergyIntensityekWh\_sqft : num 1.50e-10 1.32e-10 7.24e-11 4.67e-11 8.85e-11 ...  
## $ EnergyIntensity\_ekWh\_mega\_litre: num 0 0 0 0 0 ...  
## $ TypeofOperation : Factor w/ 25 levels "1","2","3","4",..: 24 8 17 1 10 17 17 10 4 8 ...  
## $ TypeofSector : Factor w/ 4 levels "1","2","3","4": 2 3 4 3 3 4 4 3 3 3 ...

# train.set 61441 obs. of 29 variables  
str(test.set)

## 'data.frame': 26333 obs. of 29 variables:  
## $ Year : Factor w/ 5 levels "2011","2012",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Sector : Factor w/ 4 levels "Municipal","Post-Secondary Educational Institution",..: 1 1 1 1 1 1 1 1 3 1 ...  
## $ Organization : Factor w/ 825 levels "Alexandra Hospital",..: 401 401 401 401 401 401 401 401 272 307 ...  
## $ Operation : Factor w/ 22856 levels "'A' Block","'B' Block",..: 5424 4040 5425 4562 9757 10393 2305 3343 9024 3695 ...  
## $ OperationType : Factor w/ 25 levels "Administrative office",..: 13 23 16 8 21 21 21 6 9 20 ...  
## $ Address : Factor w/ 19780 levels "#1 Hwy 527, c/o Box 98, 112 Queen St.",..: 10812 5543 10528 6196 8772 8645 8536 8823 1868 14058 ...  
## $ City : Factor w/ 1429 levels "568, rue Lessard",..: 429 429 429 429 429 429 429 429 765 845 ...  
## $ PostalCode : Factor w/ 19483 levels "H0H 0H0","K0A 1A0",..: 7747 7747 7747 7747 7747 7747 7747 7747 8443 2050 ...  
## $ TotalIndoorSpace\_sqft : num 2.78e-06 1.59e-07 5.30e-07 1.63e-06 9.30e-07 ...  
## $ WeeklyAverageHours : num 0.00708 0.01918 0.00457 0.00171 0.00571 ...  
## $ AnnualFlow\_M : num 0.00 8.54e-10 0.00 0.00 0.00 ...  
## $ NumberofPortables : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ SwimmingPool : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Electricity\_kwh : num 1.30e-06 2.01e-07 6.89e-07 2.73e-07 1.17e-07 ...  
## $ NaturalGas\_Cubicmeter : num 1.70e-06 5.08e-08 3.54e-07 1.63e-06 3.55e-07 ...  
## $ FuelOil12\_L : num 0 0 0 0 0 ...  
## $ FuelOil46\_L : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Propane\_Litre : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Coal\_Quantity : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Wood\_Metrictonne : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DistrictHeating\_GJ : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DistrictHeating\_IsRenewable : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ DistrictCooling\_GJ : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DistrictCooling\_IsRenewable : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ GHGEmissions\_KG : num 1.61e-07 5.43e-09 4.18e-08 1.28e-07 2.64e-08 ...  
## $ EnergyIntensityekWh\_sqft : num 8.82e-11 8.00e-11 1.39e-10 1.06e-10 4.02e-11 ...  
## $ EnergyIntensity\_ekWh\_mega\_litre: num 0.00 1.84e-15 0.00 0.00 0.00 ...  
## $ TypeofOperation : Factor w/ 25 levels "1","2","3","4",..: 2 4 7 8 9 9 9 10 12 5 ...  
## $ TypeofSector : Factor w/ 4 levels "1","2","3","4": 3 3 3 3 3 3 3 3 2 3 ...

# test.set 26333 obs. of 29 variables  
  
  
  
# Very Good plot - using the training set numerical columns show Correlations between pairs of attributes  
# takes very long to run (3 hours)  
# shows that GHG is determined by Electricity, Gas, Fuel Oil12 and Fuel Oil45 only  
# commented out line below because it takes hours to run  
  
# pairs.panels(train.set[,c(9:12,14:21, 23,25:27)], gap=0, pch=21)

# Prediction Stage - build model

# Multivariate Linear Regression model

###### 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

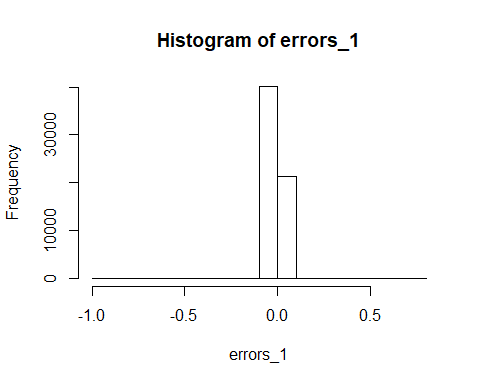
prediction\_1 <- predict(model\_mlr1, interval="prediction", newdata =test.set)

## 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"

###### 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

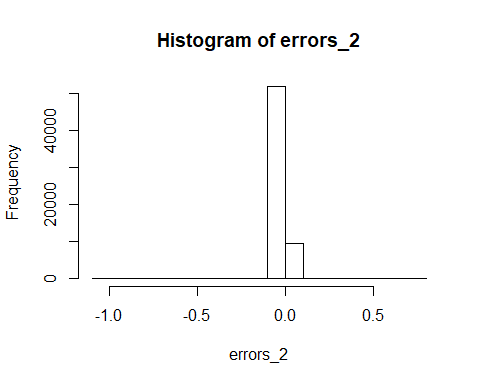
prediction\_2 <- predict(model\_mlr2, interval="prediction", newdata =test.set)

## 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"

###### 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

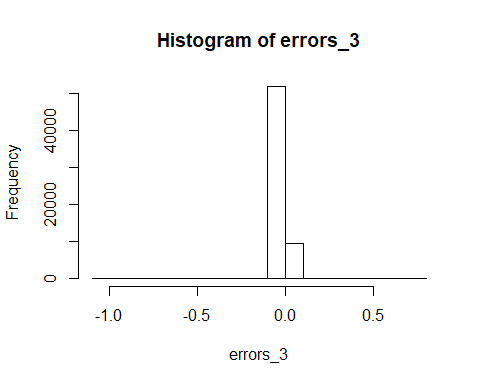
#aov gives intercept and coefficient for each variable  
# From the result, formula to calculate GHG is  
  
  
prediction\_3 <- predict(model\_mlr3, type="response", se.fit=FALSE, newdata =test.set)

## 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"

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

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(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

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"