**Project Report**

**Group 24**

**Palash Chandrakar**

**Milan Desai**

**Danish Bhat**

**617-816-8759**

**857-472-2827**

**617-368-0902**

[**chandrakar.p@husky.neu.edu**](mailto:chandrakar.p@husky.neu.edu)

[**desai.mil@husky.neu.edu**](mailto:desai.mil@husky.neu.edu)

[**bhat.da@husky.neu.edu**](mailto:bhat.da@husky.neu.edu)

Percentage of Effort Contributed by Student 1: 33.33%

Percentage of Effort Contributed by Student 2: 33.33%

Percentage of Effort Contributed by Student 3: 33.33%

Date: 04-15-2020

**Table of Content:**

**BUILDING ENERGY REPORTING AND DISCLOSURE ORDINANCE (BERDO) 3**

**Problem definition 3**

**Data source 4**

**Data description 5**

**Data exploration 8**

**Data mining models 36**

**Model Conclusion 50**

**Insights for decision making 51**

**Problem outcomes 51**

# BUILDING ENERGY REPORTING AND DISCLOSURE ORDINANCE (BERDO)

Boston enacted the Building Energy Reporting and Disclosure Ordinance (BERDO) in 2013, requiring large buildings to report their annual energy and water use to the City. Beginning in 2017 and continuing every year thereafter, the requirements apply to all non-residential buildings greater than 35,000 square feet, residential buildings that are 35,000 square feet or larger or have 35 or more units, and any parcel with multiple buildings that sum to 100,000 square feet or 100 units.

Data from previous years includes different universes of covered buildings as required by BERDO. These differences are noted in their respective descriptions. The ordinance also requires the City to make the information public.

By providing better information on building energy use, reporting and disclosure is enabling owners and tenants to become more aware of energy use, energy costs, and greenhouse gas emissions and opportunities to reduce all three.

##### Here are some of the challenges that we HAVE:

* Usage figures reported in these datasets are provided by building owners and have not been independently verified by the City of Boston.
* This dataset will be updated periodically with usage figures which are submitted or revised after the official reporting deadline so there are many NAs.
* The compliance map reflects buildings' status as of the reporting deadline in the most recent year, and late submissions will remain marked as out of compliance even if usage figures have subsequently been added to the dataset.
* Reports for buildings which show "Not Available" values in one or more fields are pending revision and will be updated when additional data becomes available so many values are not available in the dataset.

***Problem definition***

Since the data here is given in very crude form, the owners of the building do not know what’s the compliance of building they have and how much energy they use ,what should be their water intensity ,their GHG emission after the owner knows all this he should know that why his property was covered in the five year energy assessment plan .So we will create models to classify different facilities whether they are covered in the assessment or not .Thus making department more ready if buildings are required to report for the 5-year Energy Action and Assessment requirement this year and if the department knows this they can make better decisions of the facilities of their compliance level.

**Data Sources :**

SOURCE ENERGY STAR Portfolio Manager®

PUBLISHER Department of Innovation and Technology

UPDATE FREQUENCY Annual

TEMPORAL FROM 2015-01-01

TEMPORAL TO 2018-12-31

TEMPORAL NOTES Data coverage for each file is the calendar prior to reporting year. Updates may be made between reporting periods to include late or revised submissions.

THEME Archaeology, Environment, Facilities

LOCATION Boston (all)

CONTACT POINT Climate and Buildings Program Manager

CONTACT POINT EMAIL benjamin.silverman@boston.gov

CONTACT POINT PHONE NUMBER (617) 635-3850

LICENSE Open Data Commons Public Domain Dedication and License (PDDL)

Link : <https://data.boston.gov/dataset/building-energy-reporting-and-disclosure-ordinance>

**Data Description**

|  |  |
| --- | --- |
| **Field Name** | **Definition** |
| Property Name | Name of building(s) given by submitter. |
| Reported | Indicates if the building(s) is a municipal property, if it reported as required to, or if it voluntarily reported in 2017. |
| Property Type | The property type indicates the single, primary use of the  building(s). Energy usage varies between buildings based on their primary usage. This metric is useful in determining which buildings are comparable with each other to determine relative energy performance. |
| Address | Street address of building(s). |
| ZIP | Zip code of building(s). |
| Gross Area (sq ft) | The Gross Floor Area (GFA) is the total square footage of the property,  measured as the area between the outside exterior walls of the building(s). This includes all areas inside the building(s) including supporting areas. GFA is not the same as rentable space, but rather includes all area inside the building(s). |
| Site EUI (kBTU/sf) | Site Energy is the annual amount of all the energy a building(s) consumes onsite. Site Energy Use Intensity (EUI) is measured in Thousands of British Thermal Units, divided by the Gross Floor Area. This metric is an effective  means to compare the energy efficiency between two buildings. The lower the EUI the more efficient the building(s). |
| Energy Star Score | The ENERGY STAR Score is a measure of how well a property is performing relative to similar properties, when normalized for climate and operational characteristics. The ENERGY STAR scores are based on data from national building energy consumption surveys. The 1-100 scale is set so that 1 represents the worst performing buildings and 100 represents the best performing buildings. A score of 50 indicates that a building is performing at the national median, taking into account its size, location, and operating parameters.   Not all building types are able to receive a ENERGY STAR Score. Only one of the following building types can receive a ENERGY STAR Score: Data center, Hospital, Hotel, K-12 school, Medical office, Multifamily housing, Office, Parking, Residence hall/ dormitory, Retail store, Senior care community, Supermarket/grocery store, Swimming pool, Warehouse, Wastewater treatment plant, and Worship facility. |
| Energy Star Certified | ENERGY STAR Certification is awarded to buildings that earn a 75 or higher on EPA's 1-100 energy performance scale, indicating that the facility performs better than at least 75% of similar buildings nationwide. The ENERGY STAR energy performance scale accounts for differences in operating conditions, regional weather data, and other important considerations. Eligible buildings must submit an application that has been verified by an accredited energy engineer in order to be ENERGY STAR Certified. This column indicates which years a building has been ENERGY STAR Certified. |
| Property Uses | Many buildings have multiple uses beyond just the primary Property Type. This column indicates all uses for the building that were reported. Energy usage can vary among the same primary Property Type depending on the other uses for the building. This column can help to better compare energy performance between different buildings. |
| Year Built | Year building(s) was constructed. |
| GHG Emissions (MTCO2e) | Greenhouse Gas (GHG) Emissions are the carbon dioxide (CO2), methane (CH4), and nitrous oxide (N2O) gases released into the atmosphere as a result of energy consumption at the property, and is a measure of the building's climate impact. GHG emissions are expressed in carbon dioxide equivalent (CO2e), a universal unit of measure that combines the quantity and global warming potential of each greenhouse gas. This metric is measured in Metric Tons of CO2e and are derived using the ENERGY STAR GHG emissions coefficients. These coefficients do not take into account the specific fuel mix of the Boston electrical grid and district steam system, and will differ from the City of Boston's official GHG inventory. |
| GHG Intensity (kgCO2/sf) | The Greenhouse Gas (GHG) Emissions from the building divided by the Gross Area, as measured in kilograms of CO2e. This metric is an effective means to compare the carbon efficiency between two buildings. The lower the GHG Intensity the more efficient the building(s). |
| Total Site Energy (kBTU) | Site Energy is the annual amount of all the energy a building(s) consumes onsite. Total Site Energy is measured in Thousands of British Thermal Units. |
| % Electricity | The percentage of the annual Total Site Energy originating from Electricity consumption. |
| % Gas | The percentage of the annual Total Site Energy originating from Natural Gas consumption. |
| % Steam | The percentage of the annual Total Site Energy originating from District Steam consumption. |
| Water Intensity (gal/sf) | All annual water source consumption divided by the building(s) Gross Area. |
| Onsite Renewable (kWh) | The total amount energy produced from onsite solar panels or wind turbines. |
| User Submitted Info | Additional information the user submitted as part of their energy and water report for this building(s). |
| User Submitted Link | Additional links the user submitted as part of their energy and water report for this building(s). |
| Tax Parcel | The Boston Tax Parcel for this building(s) as either reported by the submitter or determined by the City of Boston. |
| Years Reported | Years in which this building(s) reported for the Building Energy Reporting and Disclosure Ordinance. |
| Covered by 5 Year Energy Action and Assessment Required in (Year) | Indicates if buildings are required to report for the 5-year Energy Action and Assessment requirement this year. |
| 5 Year Energy Action and Assessment Compliance Status | If covered by Energy Action and Assessment this year, column indicates compliance status and pathway for compliance. Definitions of field content below. |

**Data Exploration:**

**Here we deleted many columns which have data of less than 50%, because it is difficult to fill the data for these many rows also we have aggregated our property types in the dataset.**

library(dplyr)

##   
## Attaching package: 'dplyr'

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

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

library(tidyverse)

## -- Attaching packages -------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.0 v purrr 0.3.3  
## v tibble 2.1.3 v stringr 1.4.0  
## v tidyr 1.0.2 v forcats 0.4.0  
## v readr 1.3.1

## -- Conflicts ----------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggplot2)

library(readxl)  
berdo <- read\_excel("C:/Users/palash/Downloads/berdo-data-for-disclosure-calendar-year-2018 (1).xlsx")  
head(berdo)

## # A tibble: 6 x 25  
## `Property Name` Reported `Property Type` Address ZIP `Gross Area (sq~  
## <chr> <chr> <chr> <chr> <chr> <dbl>  
## 1 #2665 - West R~ Complia~ Retail Store 1213 V~ 02132 123596  
## 2 #2679 - South ~ Complia~ Retail Store 5 Alls~ 02125 130520  
## 3 00602-prudenti~ Complia~ Supermarket/Gr~ 53 Hun~ 02199 42918  
## 4 045-Holiday In~ Complia~ Hotel 69 R. ~ 02125 58051  
## 5 046-Courtyard ~ Complia~ Hotel 63 R. ~ 02125 99213  
## 6 06630-Boston, ~ Complia~ Retail Store 786 Bo~ 02199 110000  
## # ... with 19 more variables: `Site EUI (kBTU/sf)` <chr>, `Energy Star  
## # Score` <dbl>, `Energy Star Certified` <chr>, `Property Uses` <chr>, `Year  
## # Built` <chr>, `GHG Emissions (MTCO2e)` <chr>, `GHG Intensity  
## # (kgCO2/sf)` <chr>, `Total Site Energy (kBTU)` <chr>, `% Electricity` <dbl>,  
## # `% Gas` <dbl>, `% Steam` <dbl>, `Water Intensity (gal/sf)` <chr>, `Onsite  
## # Renewable (kWh)` <dbl>, `User Submitted Info` <chr>, `User Submitted  
## # Link` <chr>, `Tax Parcel` <chr>, `Years Reported` <chr>, `Covered by 5 Year  
## # Energy Action and Assessment Required in 2019` <chr>, `5 Year Energy Action  
## # and Assessment Compliance Status` <chr>

# eliminate unuseful columns  
clean\_berdo <- berdo[, -c(1, 4, 9, 17, 19, 20, 21, 22, 23, 25)]  
head(clean\_berdo)

## # A tibble: 6 x 15  
## Reported `Property Type` ZIP `Gross Area (sq~ `Site EUI (kBTU~  
## <chr> <chr> <chr> <dbl> <chr>   
## 1 Complia~ Retail Store 02132 123596 86.8   
## 2 Complia~ Retail Store 02125 130520 69.599999999999~  
## 3 Complia~ Supermarket/Gr~ 02199 42918 251.9   
## 4 Complia~ Hotel 02125 58051 117.6   
## 5 Complia~ Hotel 02125 99213 67.400000000000~  
## 6 Complia~ Retail Store 02199 110000 107.1   
## # ... with 10 more variables: `Energy Star Score` <dbl>, `Property Uses` <chr>,  
## # `Year Built` <chr>, `GHG Emissions (MTCO2e)` <chr>, `GHG Intensity  
## # (kgCO2/sf)` <chr>, `Total Site Energy (kBTU)` <chr>, `% Electricity` <dbl>,  
## # `% Gas` <dbl>, `Water Intensity (gal/sf)` <chr>, `Covered by 5 Year Energy  
## # Action and Assessment Required in 2019` <chr>

names(clean\_berdo) <- c("reported", "proptype", "zip", "area", "EUI", "ESScore", "propuse", "yearbuilt", "GHGEmission", "GHGIntensity", "totalenergy", "electricity", "gas", "water", "covered")  
clean\_berdo$EUI <- as.numeric(clean\_berdo$EUI)

## Warning: NAs introduced by coercion

clean\_berdo$GHGEmission <- as.numeric(clean\_berdo$GHGEmission)

## Warning: NAs introduced by coercion

clean\_berdo$GHGIntensity <- as.numeric(clean\_berdo$GHGIntensity)

## Warning: NAs introduced by coercion

clean\_berdo$totalenergy <- as.numeric(clean\_berdo$totalenergy)

## Warning: NAs introduced by coercion

clean\_berdo$water <- as.numeric(clean\_berdo$water)

## Warning: NAs introduced by coercion

summary(clean\_berdo)

## reported proptype zip area   
## Length:1753 Length:1753 Length:1753 Min. : 1   
## Class :character Class :character Class :character 1st Qu.: 37433   
## Mode :character Mode :character Mode :character Median : 71450   
## Mean : 153548   
## 3rd Qu.: 155880   
## Max. :4699442   
##   
## EUI ESScore propuse yearbuilt   
## Min. : 0.0 Min. : 1.00 Length:1753 Length:1753   
## 1st Qu.: 51.1 1st Qu.: 41.00 Class :character Class :character   
## Median : 71.5 Median : 68.00 Mode :character Mode :character   
## Mean : 194.3 Mean : 61.88   
## 3rd Qu.: 101.7 3rd Qu.: 86.00   
## Max. :54121.1 Max. :100.00   
## NA's :84 NA's :630   
## GHGEmission GHGIntensity totalenergy electricity   
## Min. : 0.0 Min. : 0.00 Min. :0.000e+00 Min. :0.0000   
## 1st Qu.: 129.4 1st Qu.: 3.20 1st Qu.:2.157e+06 1st Qu.:0.1966   
## Median : 298.1 Median : 4.50 Median :4.984e+06 Median :0.3612   
## Mean : 1519.5 Mean : 12.39 Mean :2.443e+07 Mean :0.4187   
## 3rd Qu.: 758.6 3rd Qu.: 6.30 3rd Qu.:1.224e+07 3rd Qu.:0.5925   
## Max. :347241.9 Max. :3593.90 Max. :5.229e+09 Max. :0.9209   
## NA's :65 NA's :66 NA's :84 NA's :106   
## gas water covered   
## Min. :0.0000 Min. : 0 Length:1753   
## 1st Qu.:0.4077 1st Qu.: 11 Class :character   
## Median :0.6413 Median : 27 Mode :character   
## Mean :0.5822 Mean : 73016   
## 3rd Qu.:0.8013 3rd Qu.: 47   
## Max. :1.0000 Max. :76520210   
## NA's :408 NA's :389

mean\_berdo <- clean\_berdo %>%   
 select(proptype, area, EUI, ESScore, GHGEmission, GHGIntensity, totalenergy, electricity, gas, water) %>%   
 group\_by(proptype) %>% summarise\_all(funs(mean), na.rm = TRUE)

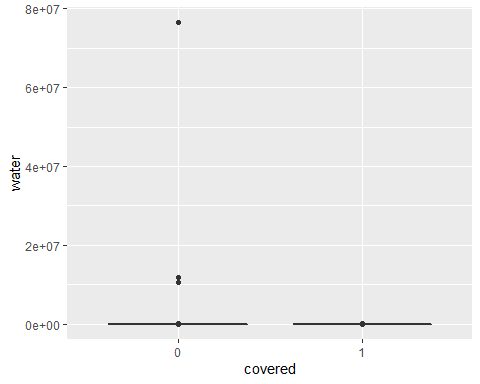
## Warning: funs() is soft deprecated as of dplyr 0.8.0  
## Please use a list of either functions or lambdas:   
##   
## # Simple named list:   
## list(mean = mean, median = median)  
##   
## # Auto named with `tibble::lst()`:   
## tibble::lst(mean, median)  
##   
## # Using lambdas  
## list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))  
## This warning is displayed once per session.

view(mean\_berdo)  
clean\_berdo <- clean\_berdo[-c(14,27 ,43,59,88,99,126,139,141,244,260,283,288,357,408,426,431,460,464,478,479,482,491,507,512,528,581,623,627,647,660,661,662,676,681,685,693,694,759,837,848,861,866,872,945,965,973,1017,1045,1074,1113,1116,1128,1149,1168,1180,1214,1279,1285,1296,1300,1308,1347,1348,1351,1354,1357,1438,1440,1441,1442,1490,1491,1522,1536,1542,1567,1574,1584,1656,1709,1718,1739,1754),]  
  
clean\_berdo <- clean\_berdo %>% filter(proptype != "Bar/Nightclub",   
 proptype != "Data Center",   
 proptype != "Ice/Curling Rink",   
 proptype != "Not Available",   
 proptype != "Race Track")  
  
mean\_berdo <- mean\_berdo[-c(4, 8, 18,30,49),]

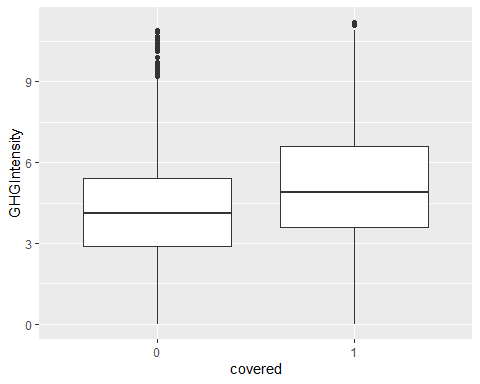
for(i in 1:length(clean\_berdo$proptype)){  
 for(j in 1:ncol(clean\_berdo)){  
 proptype1<-clean\_berdo[i,2]  
 if(is.na(clean\_berdo[i,j])){  
 clean\_berdo[i,j] <- mean\_berdo[which(mean\_berdo$proptype == as.character(proptype1)), which(names(mean\_berdo) == names(clean\_berdo[i,j]))]  
 }  
 }  
}  
  
# get rid of the ESSScore  
clean\_berdo <- clean\_berdo %>% select(-ESScore)  
  
clean\_berdo <- na.omit(clean\_berdo)  
summary(clean\_berdo)

## reported proptype zip area   
## Length:1655 Length:1655 Length:1655 Min. : 1   
## Class :character Class :character Class :character 1st Qu.: 37550   
## Mode :character Mode :character Mode :character Median : 71057   
## Mean : 151040   
## 3rd Qu.: 154172   
## Max. :3912529   
## EUI propuse yearbuilt GHGEmission   
## Min. : 0.0 Length:1655 Length:1655 Min. : 0.0   
## 1st Qu.: 51.7 Class :character Class :character 1st Qu.: 134.8   
## Median : 72.9 Mode :character Mode :character Median : 313.4   
## Mean : 167.4 Mean : 1401.7   
## 3rd Qu.: 104.3 3rd Qu.: 819.1   
## Max. :39792.8 Max. :347241.9   
## GHGIntensity totalenergy electricity gas   
## Min. : 0.00 Min. :0.000e+00 Min. :0.0000 Min. :0.0000   
## 1st Qu.: 3.20 1st Qu.:2.258e+06 1st Qu.:0.2035 1st Qu.:0.3600   
## Median : 4.50 Median :5.207e+06 Median :0.3672 Median :0.6127   
## Mean : 10.66 Mean :2.258e+07 Mean :0.4203 Mean :0.5672   
## 3rd Qu.: 6.40 3rd Qu.:1.385e+07 3rd Qu.:0.5925 3rd Qu.:0.7626   
## Max. :2642.50 Max. :5.229e+09 Max. :0.9209 Max. :1.0000   
## water covered   
## Min. : 0 Length:1655   
## 1st Qu.: 10 Class :character   
## Median : 24 Mode :character   
## Mean : 62762   
## 3rd Qu.: 47   
## Max. :76520210

#write.csv(clean\_berdo, "C:/Users/palash/Documents/Desktop/clean\_berdo.csv")  
#trying to clean the dataset  
library(ggplot2)  
ggplot(clean\_berdo,aes(y=water ,x=covered))+geom\_boxplot()



#by plotting various box plots between covered and various other variables we found many outliers, so we need to remove those outliers to properly visualize the data ,for that we are using interquartile ranges and eliminating the values above 1.25IQR and 0.25 below IQR.  
Q<-quantile(clean\_berdo$GHGIntensity,probs = c(0.25,0.75,na.rm= FALSE))  
iqr<- IQR(clean\_berdo$GHGIntensity)  
#Now that we know the IQR and the quantiles, we can find the cut-off ranges beyond which all data points are outliers.  
up <- Q[2]+1.5\*iqr # Upper Range   
low<- Q[1]-1.5\*iqr # Lower Range  
#now we can simply extract the part of our dataset between the upper and lower ranges leaving out the outliers and we can further plot the boxplot and we will get better visualization .  
eliminated\_GHGIntensity<- subset(clean\_berdo, clean\_berdo$GHGIntensity > (Q[1] - 1.5\*iqr) & clean\_berdo$GHGIntensity < (Q[2]+1.5\*iqr))  
ggplot(eliminated\_GHGIntensity,aes(y=GHGIntensity ,x=covered))+geom\_boxplot()



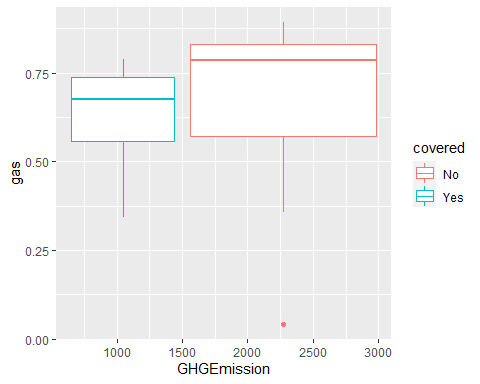
#similarly we can try plotting for other functions as well

# but we are not going with this cleaned dataset because we already have very less data in our dataset and it further reduces our dataset to 962 and further down so we are not going further with this.

ggplot(clean\_berdo\_new, aes(x=GHGEmission, y=gas, color=covered)) +   
 geom\_boxplot()

## Warning: Removed 1 rows containing missing values (stat\_boxplot).

## Warning: Removed 3 rows containing non-finite values (stat\_boxplot).

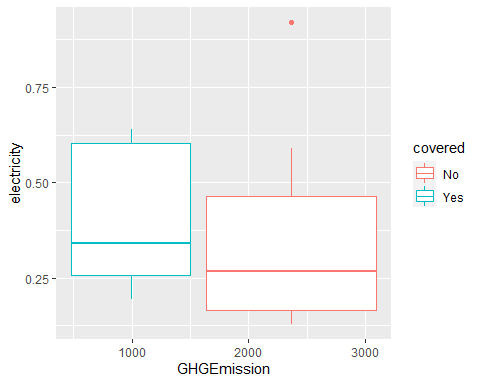


#In the comparison of GHE Emissions and gas, it can be seen that the boxplot of not covered has a higher median and a smaller range while as the box plot in the covered has a lower median and wider range. Wider boxplots indicate more variable data. In the above case, it can be seen that the covered one has a wider boxplot which means more variable data.

ggplot(clean\_berdo\_new, aes(x=GHGEmission, y=electricity, color=covered)) +   
 geom\_boxplot()

## Warning: Removed 1 rows containing missing values (stat\_boxplot).

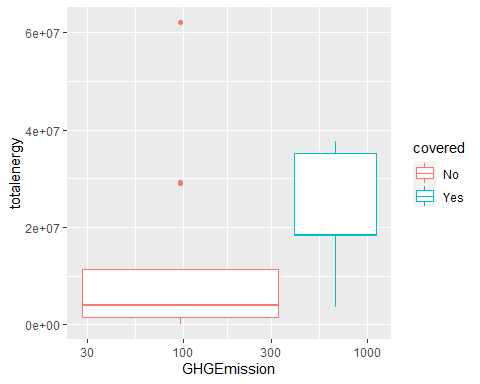
## Warning: Removed 1 rows containing non-finite values (stat\_boxplot).



#In the comparison of GHE Emissions and electricity, it can be seen that the median of not covered has a lower median and range while as the box plot in the covered has a higher median and range. Wider boxplots indicate more variable data. In the above case, it can be seen that the covered one has a wider boxplot which means more variable data.

ggplot(clean\_berdo\_new, aes(x=GHGEmission, y=totalenergy, color=covered)) +   
 geom\_boxplot() + scale\_x\_log10()

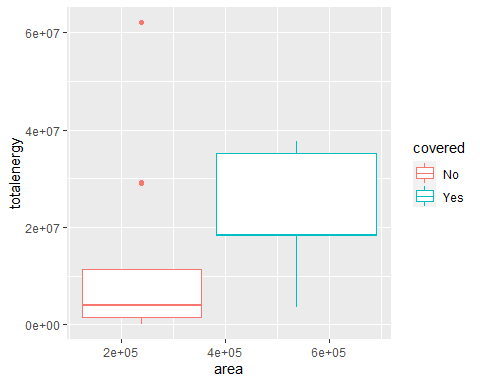
## Warning: Removed 1 rows containing missing values (stat\_boxplot).



#In the comparision of GHE Emissions and totalenergy, it can be seen that the boxplot of not covered has a higher median and range while as the box plot in the covered has a higher median and range. Wider boxplots indicate more variable data. In the above case also, it can be seen that the covered one has a wider boxplot which means more variable data.

ggplot(clean\_berdo\_new, aes(x=area, y=totalenergy, color=covered)) +   
 geom\_boxplot()

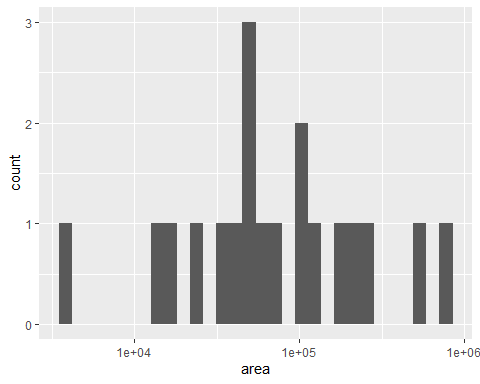
## Warning: Removed 1 rows containing non-finite values (stat\_boxplot).



#In the comparision of GHE Emissions and electricity, it can be seen that the boxplot of not covered has a lower median and range while as the box plot in the covered has a higher median and range. Wider boxplots indicate more variable data. In the above case also, it can be seen that the covered one has a wider boxplot which means more variable data.

ggplot(clean\_berdo\_new, aes(x = area)) +  
 geom\_histogram()+ scale\_x\_log10()

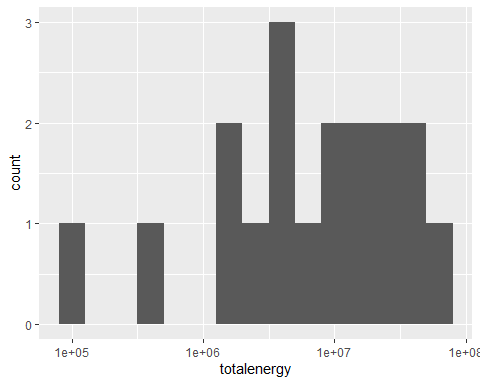
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#it can be seen that the area in the range 1\*10^5 is the highest with its range to plus minus 10^1

ggplot(clean\_berdo\_new, aes(x = totalenergy)) +  
 geom\_histogram(binwidth = .2) + scale\_x\_log10()

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

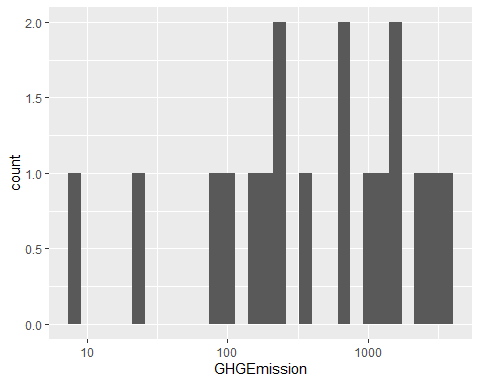


#it can be seen that the total energy lies mainly in the range 10\*6 to 10\*8 with its peak at 10\*7

ggplot(clean\_berdo\_new, aes(x = GHGEmission)) +  
 geom\_histogram()+ scale\_x\_log10()

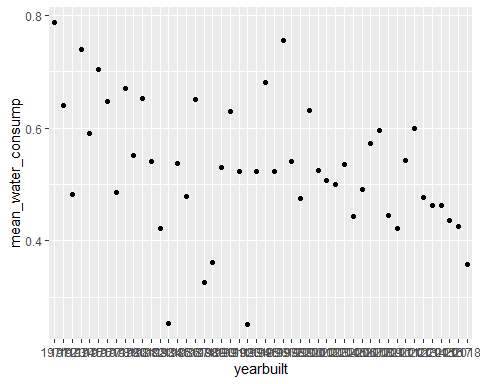
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

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

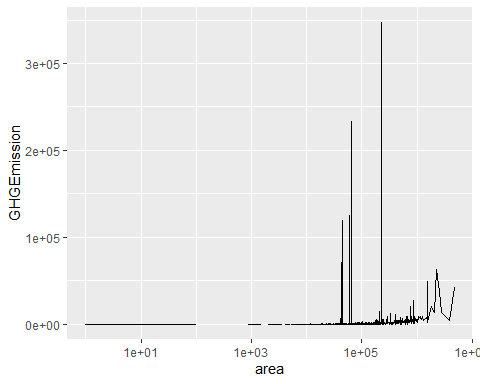


#it can be seen that in the GHGEmission histogram, the emission is high between 100 and 10000 having its peak near 1000

ggplot(by\_year\_clean\_berdo, aes(x= yearbuilt, y = mean\_water\_consump)) + geom\_point()

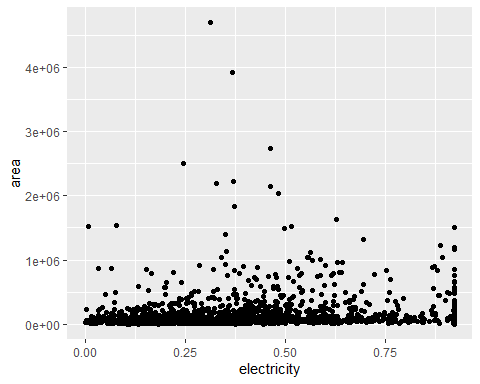


#it can be clearly seen that the mean water consumption changes rapidly during the time period  
  
ggplot(clean\_berdo, aes(x= area, y = GHGEmission)) + geom\_line()+ scale\_x\_log10()



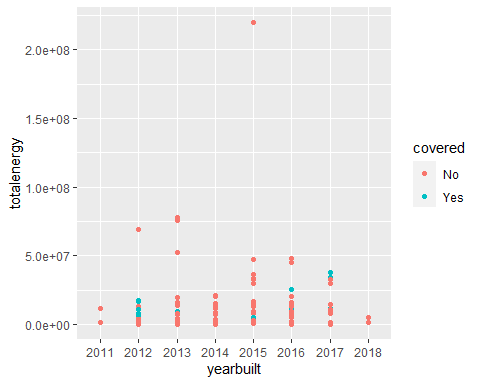
# It can be Cleary seen that as the area increases from 10\*5, the GHG Emissions starts to change rapidly.  
#above the areas 10\*5, there is a huge peak and fluctuations in GHG Emissions continue.  
  
  
  
ggplot(clean\_berdo, aes(x= electricity, y = area)) + geom\_point()

## Warning: Removed 106 rows containing missing values (geom\_point).



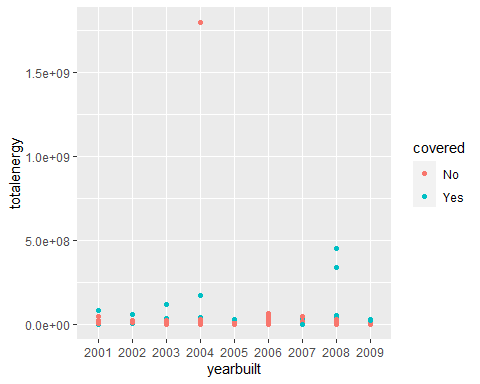
clean\_berdo\_new <- clean\_berdo %>%  
 filter( yearbuilt > 2010)  
  
ggplot(clean\_berdo\_new, aes(x= yearbuilt, y = totalenergy, color = covered))+geom\_point()

## Warning: Removed 3 rows containing missing values (geom\_point).



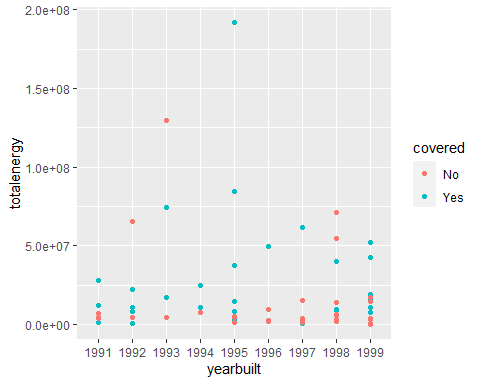
#This plot shows the relation between total energy and yearbuilt and also the color shows whether they are covered or not. It can be clearly seen that total energy above 5\*10^7 has all the plots as not covered under  
#five year compliance.  
  
  
  
clean\_berdo\_new <- clean\_berdo %>%  
 filter( yearbuilt > 2000, yearbuilt < 2010)  
  
ggplot(clean\_berdo\_new, aes(x= yearbuilt, y = totalenergy, color = covered))+geom\_point()

## Warning: Removed 16 rows containing missing values (geom\_point).



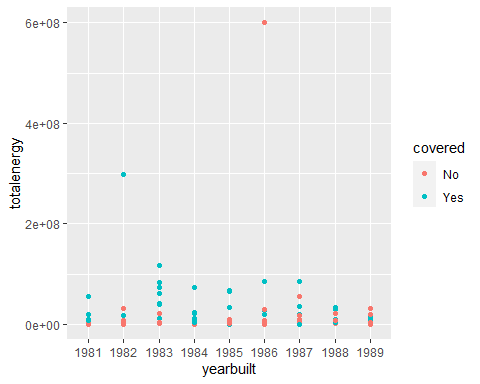
clean\_berdo\_new <- clean\_berdo %>%  
 filter( yearbuilt > 1990, yearbuilt < 2000)  
  
ggplot(clean\_berdo\_new, aes(x= yearbuilt, y = totalenergy, color = covered))+geom\_point()

## Warning: Removed 3 rows containing missing values (geom\_point).



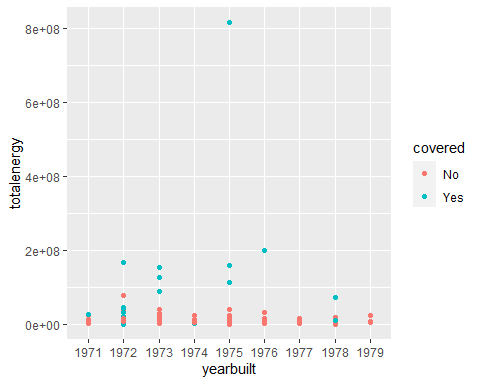
#this plot covers the data from 1990 to 2000 onwards, and it is clearly seen that between 2.5\*10^7 and   
#1\*10^8, the scatter plot is mostly covered while up to 2.5\*10^7, the area is mostly not covered  
  
  
clean\_berdo\_new <- clean\_berdo %>%  
 filter( yearbuilt >1980, yearbuilt < 1990)  
  
ggplot(clean\_berdo\_new, aes(x= yearbuilt, y = totalenergy, color = covered))+geom\_point()

## Warning: Removed 6 rows containing missing values (geom\_point).



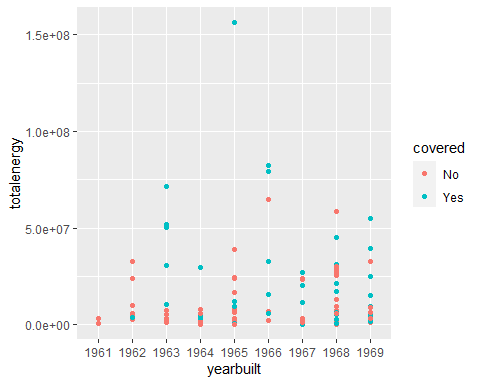
clean\_berdo\_new <- clean\_berdo %>%  
 filter( yearbuilt > 1970, yearbuilt < 1980)  
  
ggplot(clean\_berdo\_new, aes(x= yearbuilt, y = totalenergy, color = covered))+geom\_point()

## Warning: Removed 8 rows containing missing values (geom\_point).



#it can be seen that from the year 1970 to 1990, the total energy covered and not covered is not clear but those buildings having low energy are mostly under not covered region  
  
clean\_berdo\_new <- clean\_berdo %>%  
 filter( yearbuilt > 1960, yearbuilt < 1970)  
  
ggplot(clean\_berdo\_new, aes(x= yearbuilt, y = totalenergy, color = covered))+geom\_point()

## Warning: Removed 6 rows containing missing values (geom\_point).



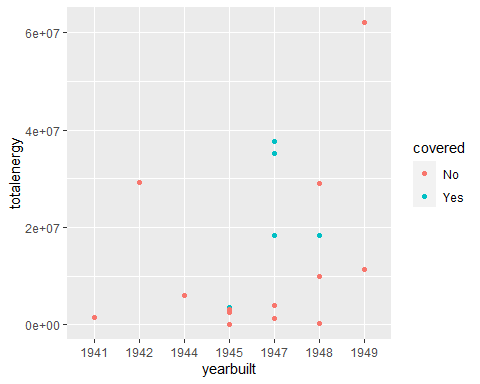
clean\_berdo\_new <- clean\_berdo %>%  
 filter( yearbuilt > 1950, yearbuilt <1960)  
   
ggplot(clean\_berdo\_new, aes(x= yearbuilt, y = totalenergy, color = covered))+geom\_point()

## Warning: Removed 1 rows containing missing values (geom\_point).

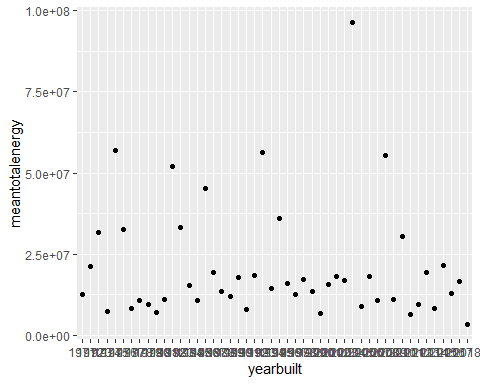


clean\_berdo\_new <- clean\_berdo %>%  
 filter( yearbuilt > 1940, yearbuilt <1950)  
  
ggplot(clean\_berdo\_new, aes(x= yearbuilt, y = totalenergy, color = covered))+geom\_point()

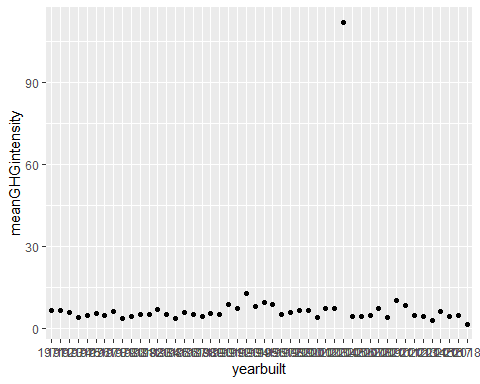
## Warning: Removed 1 rows containing missing values (geom\_point).



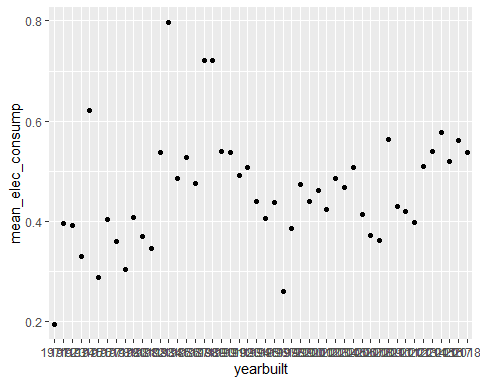
by\_year\_clean\_berdo <- clean\_berdo %>%  
filter(yearbuilt > 1970) %>%  
group\_by(yearbuilt) %>%  
summarise(meantotalenergy = mean(totalenergy, na.rm = TRUE),  
 meanGHGintensity = mean(GHGIntensity, na.rm = TRUE),  
 mean\_elec\_consump = mean(electricity, na.rm = TRUE),  
 mean\_gas\_consump = mean(water, na.rm = TRUE),  
 mean\_water\_consump = mean(gas, na.rm = TRUE))  
  
  
  
ggplot(by\_year\_clean\_berdo, aes(x= yearbuilt, y = meantotalenergy)) + geom\_point()



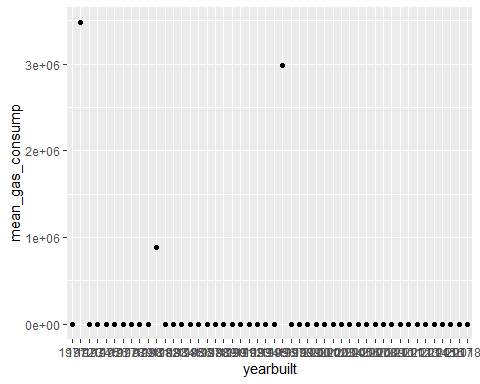
#plotting the yearbuilt and meantotalenergy, the mean total energy from 1971 to 2018 is mostly 2.5 \* 10^7  
  
  
ggplot(by\_year\_clean\_berdo, aes(x= yearbuilt, y = meanGHGintensity)) + geom\_point()



#the meanGHGintensity is uniform or does not change significantly as compared to total energy  
  
ggplot(by\_year\_clean\_berdo, aes(x= yearbuilt, y = mean\_elec\_consump)) + geom\_point()



#as seen from the scatterplot , the mean energy consumption from 1971 to 2018 is not same and changes from 0.3 to 0.6 with huge variations in between  
  
ggplot(by\_year\_clean\_berdo, aes(x= yearbuilt, y = mean\_gas\_consump)) + geom\_point()



#the mean gas consumption is constant during the time period

**Data Mining models and methods**

# k-NN Model  
# convert categorical variable to dummies  
library(fastDummies)

## Warning: package 'fastDummies' was built under R version 3.6.3

clean\_berdo <- dummy\_cols(clean\_berdo, select\_columns = c("reported" , "proptype") , remove\_first\_dummy = TRUE)  
clean\_berdo <- clean\_berdo[, -c(1,2)]  
  
# partition  
set.seed(123)  
train.index <- sample(row.names(clean\_berdo), 0.6\*dim(clean\_berdo)[1])  
valid.index <- setdiff(row.names(clean\_berdo), train.index)   
train.df <- clean\_berdo[train.index,]  
valid.df <- clean\_berdo[valid.index,]  
  
train.norm.df <- train.df  
valid.norm.df <- valid.df  
berdo.norm.df <- clean\_berdo  
  
norm.values <- preProcess (train.df, method=c("center", "scale"))

## Warning in preProcess.default(train.df, method = c("center", "scale")):  
## These variables have zero variances: proptype\_Ambulatory Surgical Center,  
## proptype\_Automobile Dealership, proptype\_Barracks, proptype\_Courthouse,  
## proptype\_Pre-school/Daycare, proptype\_Restaurant

train.norm.df <- as.data.frame(predict(norm.values, train.df))  
valid.norm.df <- as.data.frame(predict(norm.values, valid.df))  
berdo.norm.df <- as.data.frame(predict(norm.values, clean\_berdo))  
  
accuracy.df <- data.frame(k = seq(1, 20, 1),  
 Accuracy = rep(0, 20))  
# classify "covered"  
for(i in 1:20){  
 knn.pred<-knn(train = train.norm.df[, -10],   
 test = valid.norm.df[,-10],   
 cl = train.df[,10], k = i,prob = TRUE)  
 accuracy.df[i,2] <- confusionMatrix(knn.pred, valid.norm.df[,10],  
 positive = "1")$overall[1]  
}  
accuracy.df

## k Accuracy  
## 1 1 0.8308157  
## 2 2 0.8383686  
## 3 3 0.8610272  
## 4 4 0.8610272  
## 5 5 0.8595166  
## 6 6 0.8655589  
## 7 7 0.8716012  
## 8 8 0.8564955  
## 9 9 0.8519637  
## 10 10 0.8474320  
## 11 11 0.8489426  
## 12 12 0.8444109  
## 13 13 0.8429003  
## 14 14 0.8429003  
## 15 15 0.8383686  
## 16 16 0.8383686  
## 17 17 0.8338369  
## 18 18 0.8338369  
## 19 19 0.8338369  
## 20 20 0.8338369

# you can then pick the k with the best accuracy which is K=7 with accuracy=87 % but i am still not sure of its values   
# as k leads to values of 83 % after K=15 that its almost similar so lets try logistic regression model to verify the   
# accuracy values

# Logistic Regression Model  
model <- glm(covered ~.,family="binomial",data=train.norm.df)  
summary(model)

##   
## Call:  
## glm(formula = covered ~ ., family = "binomial", data = train.norm.df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4343 -0.2116 -0.1186 0.5850 3.6166   
##   
## Coefficients: (6 not defined because of singularities)  
## Estimate  
## (Intercept) -3.055e+00  
## area 3.074e-01  
## EUI -1.005e+01  
## yearbuilt 8.633e-02  
## GHGEmission -5.371e+00  
## GHGIntensity 1.013e+01  
## totalenergy 5.324e+00  
## electricity -2.826e-01  
## gas -3.731e-02  
## water -1.232e-01  
## reported\_Municipal -2.636e+00  
## reported\_Voluntary -2.341e+00  
## `proptype\_Ambulatory Surgical Center` NA  
## `proptype\_Automobile Dealership` NA  
## proptype\_Barracks NA  
## `proptype\_College/University` 4.323e+00  
## proptype\_Courthouse NA  
## `proptype\_Distribution Center` 1.659e+00  
## `proptype\_Enclosed Mall` 1.164e+00  
## `proptype\_Energy/Power Station` 9.629e-01  
## `proptype\_Financial Office` 1.567e+00  
## `proptype\_Fire Station` 9.883e-01  
## `proptype\_Fitness Center/Health Club/Gym` 1.374e+00  
## `proptype\_Food Service` 4.288e-03  
## `proptype\_Hospital (General Medical & Surgical)` 1.911e+00  
## proptype\_Hotel 3.494e+00  
## `proptype\_Indoor Arena` 1.670e+00  
## `proptype\_K-12 School` 5.337e+00  
## proptype\_Laboratory 2.449e+00  
## proptype\_Library 2.375e+00  
## `proptype\_Manufacturing/Industrial Plant` 1.668e+00  
## `proptype\_Medical Office` 1.614e+00  
## `proptype\_Mixed Use Property` 1.571e+00  
## `proptype\_Multifamily Housing` 7.031e+00  
## proptype\_Museum 1.059e+00  
## `proptype\_Non-Refrigerated Warehouse` 1.069e+00  
## proptype\_Office 7.540e+00  
## proptype\_Other 3.210e+00  
## `proptype\_Other - Education` 1.334e+00  
## `proptype\_Other - Entertainment/Public Assembly` 1.209e+00  
## `proptype\_Other - Lodging/Residential` -1.306e-02  
## `proptype\_Other - Mall` 1.188e+00  
## `proptype\_Other - Recreation` 2.487e+00  
## `proptype\_Other - Specialty Hospital` 1.334e+00  
## `proptype\_Other - Stadium` 1.176e+00  
## `proptype\_Other - Technology/Science` 2.376e+00  
## proptype\_Parking 1.980e+00  
## `proptype\_Performing Arts` 1.187e+00  
## `proptype\_Police Station` 7.564e-01  
## `proptype\_Pre-school/Daycare` NA  
## `proptype\_Repair Services (Vehicle, Shoe, Locksmith, etc.)` 2.244e-01  
## `proptype\_Residence Hall/Dormitory` 2.955e+00  
## proptype\_Restaurant NA  
## `proptype\_Retail Store` 2.356e+00  
## `proptype\_Self-Storage Facility` 1.737e+00  
## `proptype\_Senior Care Community` 1.213e+00  
## `proptype\_Single Family Home` 1.179e+00  
## `proptype\_Social/Meeting Hall` 1.660e+00  
## `proptype\_Strip Mall` 1.191e+00  
## `proptype\_Supermarket/Grocery Store` 4.061e-03  
## `proptype\_Urgent Care/Clinic/Other Outpatient` 1.450e+00  
## `proptype\_Vocational School` 2.625e-03  
## `proptype\_Worship Facility` 1.771e+00  
## Std. Error z value  
## (Intercept) 5.791e+01 -0.053  
## area 1.726e-01 1.780  
## EUI 6.103e+00 -1.647  
## yearbuilt 1.126e-01 0.767  
## GHGEmission 2.682e+00 -2.003  
## GHGIntensity 6.120e+00 1.655  
## totalenergy 2.758e+00 1.930  
## electricity 1.589e-01 -1.779  
## gas 1.636e-01 -0.228  
## water 1.284e+00 -0.096  
## reported\_Municipal 5.094e-01 -5.176  
## reported\_Voluntary 1.580e+02 -0.015  
## `proptype\_Ambulatory Surgical Center` NA NA  
## `proptype\_Automobile Dealership` NA NA  
## proptype\_Barracks NA NA  
## `proptype\_College/University` 1.480e+03 0.003  
## proptype\_Courthouse NA NA  
## `proptype\_Distribution Center` 3.582e+02 0.005  
## `proptype\_Enclosed Mall` 2.927e+02 0.004  
## `proptype\_Energy/Power Station` 3.582e+02 0.003  
## `proptype\_Financial Office` 5.057e+02 0.003  
## `proptype\_Fire Station` 1.003e+03 0.001  
## `proptype\_Fitness Center/Health Club/Gym` 5.057e+02 0.003  
## `proptype\_Food Service` 2.927e+02 0.000  
## `proptype\_Hospital (General Medical & Surgical)` 6.830e+02 0.003  
## proptype\_Hotel 1.170e+03 0.003  
## `proptype\_Indoor Arena` 3.582e+02 0.005  
## `proptype\_K-12 School` 1.745e+03 0.003  
## proptype\_Laboratory 8.217e+02 0.003  
## proptype\_Library 7.418e+02 0.003  
## `proptype\_Manufacturing/Industrial Plant` 5.834e+02 0.003  
## `proptype\_Medical Office` 5.834e+02 0.003  
## `proptype\_Mixed Use Property` 5.460e+02 0.003  
## `proptype\_Multifamily Housing` 3.145e+03 0.002  
## proptype\_Museum 3.582e+02 0.003  
## `proptype\_Non-Refrigerated Warehouse` 3.582e+02 0.003  
## proptype\_Office 2.452e+03 0.003  
## proptype\_Other 1.080e+03 0.003  
## `proptype\_Other - Education` 4.619e+02 0.003  
## `proptype\_Other - Entertainment/Public Assembly` 4.619e+02 0.003  
## `proptype\_Other - Lodging/Residential` 6.517e+02 0.000  
## `proptype\_Other - Mall` 2.927e+02 0.004  
## `proptype\_Other - Recreation` 7.694e+02 0.003  
## `proptype\_Other - Specialty Hospital` 2.927e+02 0.005  
## `proptype\_Other - Stadium` 2.927e+02 0.004  
## `proptype\_Other - Technology/Science` 4.618e+02 0.005  
## proptype\_Parking 6.185e+02 0.003  
## `proptype\_Performing Arts` 4.134e+02 0.003  
## `proptype\_Police Station` 7.696e+02 0.001  
## `proptype\_Pre-school/Daycare` NA NA  
## `proptype\_Repair Services (Vehicle, Shoe, Locksmith, etc.)` 6.694e+02 0.000  
## `proptype\_Residence Hall/Dormitory` 9.816e+02 0.003  
## proptype\_Restaurant NA NA  
## `proptype\_Retail Store` 7.960e+02 0.003  
## `proptype\_Self-Storage Facility` 6.185e+02 0.003  
## `proptype\_Senior Care Community` 4.619e+02 0.003  
## `proptype\_Single Family Home` 2.927e+02 0.004  
## `proptype\_Social/Meeting Hall` 5.057e+02 0.003  
## `proptype\_Strip Mall` 2.927e+02 0.004  
## `proptype\_Supermarket/Grocery Store` 2.927e+02 0.000  
## `proptype\_Urgent Care/Clinic/Other Outpatient` 5.057e+02 0.003  
## `proptype\_Vocational School` 2.927e+02 0.000  
## `proptype\_Worship Facility` 6.516e+02 0.003  
## Pr(>|z|)   
## (Intercept) 0.9579   
## area 0.0750 .   
## EUI 0.0996 .   
## yearbuilt 0.4433   
## GHGEmission 0.0452 \*   
## GHGIntensity 0.0979 .   
## totalenergy 0.0536 .   
## electricity 0.0752 .   
## gas 0.8196   
## water 0.9236   
## reported\_Municipal 2.27e-07 \*\*\*  
## reported\_Voluntary 0.9882   
## `proptype\_Ambulatory Surgical Center` NA   
## `proptype\_Automobile Dealership` NA   
## proptype\_Barracks NA   
## `proptype\_College/University` 0.9977   
## proptype\_Courthouse NA   
## `proptype\_Distribution Center` 0.9963   
## `proptype\_Enclosed Mall` 0.9968   
## `proptype\_Energy/Power Station` 0.9979   
## `proptype\_Financial Office` 0.9975   
## `proptype\_Fire Station` 0.9992   
## `proptype\_Fitness Center/Health Club/Gym` 0.9978   
## `proptype\_Food Service` 1.0000   
## `proptype\_Hospital (General Medical & Surgical)` 0.9978   
## proptype\_Hotel 0.9976   
## `proptype\_Indoor Arena` 0.9963   
## `proptype\_K-12 School` 0.9976   
## proptype\_Laboratory 0.9976   
## proptype\_Library 0.9974   
## `proptype\_Manufacturing/Industrial Plant` 0.9977   
## `proptype\_Medical Office` 0.9978   
## `proptype\_Mixed Use Property` 0.9977   
## `proptype\_Multifamily Housing` 0.9982   
## proptype\_Museum 0.9976   
## `proptype\_Non-Refrigerated Warehouse` 0.9976   
## proptype\_Office 0.9975   
## proptype\_Other 0.9976   
## `proptype\_Other - Education` 0.9977   
## `proptype\_Other - Entertainment/Public Assembly` 0.9979   
## `proptype\_Other - Lodging/Residential` 1.0000   
## `proptype\_Other - Mall` 0.9968   
## `proptype\_Other - Recreation` 0.9974   
## `proptype\_Other - Specialty Hospital` 0.9964   
## `proptype\_Other - Stadium` 0.9968   
## `proptype\_Other - Technology/Science` 0.9959   
## proptype\_Parking 0.9974   
## `proptype\_Performing Arts` 0.9977   
## `proptype\_Police Station` 0.9992   
## `proptype\_Pre-school/Daycare` NA   
## `proptype\_Repair Services (Vehicle, Shoe, Locksmith, etc.)` 0.9997   
## `proptype\_Residence Hall/Dormitory` 0.9976   
## proptype\_Restaurant NA   
## `proptype\_Retail Store` 0.9976   
## `proptype\_Self-Storage Facility` 0.9978   
## `proptype\_Senior Care Community` 0.9979   
## `proptype\_Single Family Home` 0.9968   
## `proptype\_Social/Meeting Hall` 0.9974   
## `proptype\_Strip Mall` 0.9968   
## `proptype\_Supermarket/Grocery Store` 1.0000   
## `proptype\_Urgent Care/Clinic/Other Outpatient` 0.9977   
## `proptype\_Vocational School` 1.0000   
## `proptype\_Worship Facility` 0.9978   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1232.98 on 992 degrees of freedom  
## Residual deviance: 544.29 on 936 degrees of freedom  
## AIC: 658.29  
##   
## Number of Fisher Scoring iterations: 17

# using logistic regression model on validation data  
p <- predict(model, valid.norm.df[, -10])

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from a rank-deficient fit may be misleading

#p

# stepwide elimination  
library(MASS)  
library(dplyr)

##   
## Attaching package: 'dplyr'

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

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

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

step.model <- model %>% stepAIC(trace = FALSE)  
coef(step.model)

## (Intercept)   
## -2.4577360   
## area   
## 0.3312611   
## EUI   
## -10.9550890   
## GHGEmission   
## -5.6177158   
## GHGIntensity   
## 11.0278668   
## totalenergy   
## 5.5864080   
## electricity   
## -0.2509137   
## reported\_Municipal   
## -2.6166618   
## reported\_Voluntary   
## -2.2167150   
## `proptype\_College/University`   
## 1.0142238   
## `proptype\_Distribution Center`   
## 0.9633042   
## `proptype\_Enclosed Mall`   
## 0.6692167   
## `proptype\_Energy/Power Station`   
## 0.1655677   
## `proptype\_Financial Office`   
## 0.4388962   
## `proptype\_Fitness Center/Health Club/Gym`   
## 0.2509940   
## `proptype\_Hospital (General Medical & Surgical)`   
## 0.3811252   
## proptype\_Hotel   
## 0.8943410   
## `proptype\_Indoor Arena`   
## 0.9730329   
## `proptype\_K-12 School`   
## 1.4354526   
## proptype\_Laboratory   
## 0.6193196   
## proptype\_Library   
## 0.7206867   
## `proptype\_Manufacturing/Industrial Plant`   
## 0.3650100   
## `proptype\_Medical Office`   
## 0.3094892   
## `proptype\_Mixed Use Property`   
## 0.3531168   
## proptype\_Museum   
## 0.2584701   
## `proptype\_Non-Refrigerated Warehouse`   
## 0.2663417   
## proptype\_Office   
## 2.0569552   
## proptype\_Other   
## 0.7945020   
## `proptype\_Other - Education`   
## 0.3001393   
## `proptype\_Other - Entertainment/Public Assembly`   
## 0.1799240   
## `proptype\_Other - Mall`   
## 0.6933192   
## `proptype\_Other - Recreation`   
## 0.7679016   
## `proptype\_Other - Specialty Hospital`   
## 0.8375436   
## `proptype\_Other - Stadium`   
## 0.6851115   
## `proptype\_Other - Technology/Science`   
## 1.3908071   
## proptype\_Parking   
## 0.6015968   
## `proptype\_Performing Arts`   
## 0.2592999   
## `proptype\_Residence Hall/Dormitory`   
## 0.7642178   
## `proptype\_Retail Store`   
## 0.5797778   
## `proptype\_Self-Storage Facility`   
## 0.3568424   
## `proptype\_Senior Care Community`   
## 0.1813419   
## `proptype\_Single Family Home`   
## 0.6835947   
## `proptype\_Social/Meeting Hall`   
## 0.5339519   
## `proptype\_Strip Mall`   
## 0.6993451   
## `proptype\_Urgent Care/Clinic/Other Outpatient`   
## 0.3263513   
## `proptype\_Worship Facility`   
## 0.3110536

#step.model  
  
# using logistic regression model on validation data  
step\_p <- predict(step.model, valid.norm.df[, -10])  
#step\_p   
  
# use 0 as cutoff for classification  
step\_p1 <- ifelse(step\_p >= 0, "1", "0")  
confusionMatrix(valid.norm.df[, 10], as.factor(step\_p1))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 408 48  
## 1 29 177  
##   
## Accuracy : 0.8837   
## 95% CI : (0.8568, 0.9071)  
## No Information Rate : 0.6601   
## P-Value [Acc > NIR] : < 2e-16   
##   
## Kappa : 0.7354   
##   
## Mcnemar's Test P-Value : 0.04024   
##   
## Sensitivity : 0.9336   
## Specificity : 0.7867   
## Pos Pred Value : 0.8947   
## Neg Pred Value : 0.8592   
## Prevalence : 0.6601   
## Detection Rate : 0.6163   
## Detection Prevalence : 0.6888   
## Balanced Accuracy : 0.8602   
##   
## 'Positive' Class : 0   
##

#so we get accuracy values of approx 88 % from logistic regression thus we can say that for k=7 the acuracy of   
#87 % sensitivity =93% and specificity =78% was good lets just confirm it again using

#Repeated K-fold cross validation to know the accuracy of the model  
  
#The process of splitting the data into k-folds can be repeated a number of times,  
#this is called repeated k-fold cross validation.  
  
# The final model error is taken as the mean error from the number of repeats.  
  
#The following code uses 10-fold cross validation with 3 repeats   
  
# Define training control  
set.seed(123)  
train.control <- trainControl(method = "repeatedcv",   
 number = 10, repeats = 3)  
# Train the model  
model.cross <- train(covered ~., data = clean\_berdo, method = "glm",  
 trControl = train.control)

#predict  
predict.cross<-predict(model.cross,valid.norm.df[ ,-10])  
  
#accuracy  
confusionMatrix(valid.norm.df[ ,10],predict.cross)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 380 76  
## 1 108 98  
##   
## Accuracy : 0.7221   
## 95% CI : (0.6862, 0.7559)  
## No Information Rate : 0.7372   
## P-Value [Acc > NIR] : 0.82333   
##   
## Kappa : 0.3228   
##   
## Mcnemar's Test P-Value : 0.02229   
##   
## Sensitivity : 0.7787   
## Specificity : 0.5632   
## Pos Pred Value : 0.8333   
## Neg Pred Value : 0.4757   
## Prevalence : 0.7372   
## Detection Rate : 0.5740   
## Detection Prevalence : 0.6888   
## Balanced Accuracy : 0.6710   
##   
## 'Positive' Class : 0   
##

#Repeated K-fold cross validation to know the accuracy of the model  
  
#The process of splitting the data into k-folds can be repeated a number of times,  
#this is called repeated k-fold cross validation.  
  
# The final model error is taken as the mean error from the number of repeats.  
  
#The following code uses 10-fold cross validation with 3 repeats   
  
# Define training control  
set.seed(123)  
train.control <- trainControl(method = "repeatedcv",   
 number = 10, repeats = 3)  
# Train the model  
model.cross <- train(covered ~., data = clean\_berdo, method = "glm",  
 trControl = train.control)

# Summarize the results  
print(model.cross)

## Generalized Linear Model   
##   
## 1655 samples  
## 62 predictor  
## 2 classes: '0', '1'   
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 3 times)   
## Summary of sample sizes: 1489, 1489, 1490, 1490, 1489, 1489, ...   
## Resampling results:  
##   
## Accuracy Kappa   
## 0.864055 0.698311

#predict  
predict.cross<-predict(model.cross,valid.norm.df[ ,-10])  
  
#accuracy  
confusionMatrix(valid.norm.df[ ,10],predict.cross)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 380 76  
## 1 108 98  
##   
## Accuracy : 0.7221   
## 95% CI : (0.6862, 0.7559)  
## No Information Rate : 0.7372   
## P-Value [Acc > NIR] : 0.82333   
##   
## Kappa : 0.3228   
##   
## Mcnemar's Test P-Value : 0.02229   
##   
## Sensitivity : 0.7787   
## Specificity : 0.5632   
## Pos Pred Value : 0.8333   
## Neg Pred Value : 0.4757   
## Prevalence : 0.7372   
## Detection Rate : 0.5740   
## Detection Prevalence : 0.6888   
## Balanced Accuracy : 0.6710   
##   
## 'Positive' Class : 0   
##

#evaluate for logistic regression model   
library(pROC)

## Type 'citation("pROC")' for a citation.

##   
## Attaching package: 'pROC'

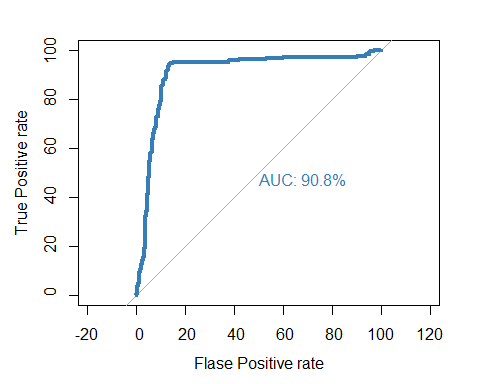
## The following object is masked from 'package:Metrics':  
##   
## auc

## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

roc(valid.norm.df$covered,p,  
 plot = TRUE,legacy.axes=TRUE,percent=TRUE,  
 xlab="Flase Positive rate",ylab="True Positive rate",  
 col="#377EB8",lwd=4,print.auc=TRUE)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases



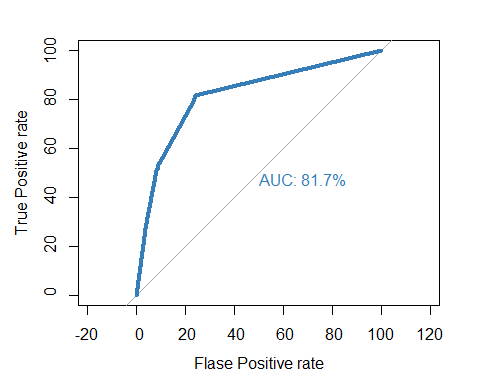
##   
## Call:  
## roc.default(response = valid.norm.df$covered, predictor = p, percent = TRUE, plot = TRUE, legacy.axes = TRUE, xlab = "Flase Positive rate", ylab = "True Positive rate", col = "#377EB8", lwd = 4, print.auc = TRUE)  
##   
## Data: p in 456 controls (valid.norm.df$covered 0) < 206 cases (valid.norm.df$covered 1).  
## Area under the curve: 90.82%

#from here we get that area under curve was 91 %

#for KNN  
knn.pred<-knn(train = train.norm.df[, -10],  
 test = valid.norm.df[,-10],  
 cl = train.df[,10], k = 7,  
 prob = TRUE)  
knn.pred <- attributes(knn.pred)  
knn.pred <- knn.pred$prob  
roc(valid.norm.df$covered,knn.pred ,  
 plot = TRUE,legacy.axes=TRUE,percent=TRUE,  
 xlab="Flase Positive rate",ylab="True Positive rate",  
 col="#377EB8",lwd=4,print.auc=TRUE)

## Setting levels: control = 0, case = 1

## Setting direction: controls > cases



##   
## Call:  
## roc.default(response = valid.norm.df$covered, predictor = knn.pred, percent = TRUE, plot = TRUE, legacy.axes = TRUE, xlab = "Flase Positive rate", ylab = "True Positive rate", col = "#377EB8", lwd = 4, print.auc = TRUE)  
##   
## Data: knn.pred in 456 controls (valid.norm.df$covered 0) > 206 cases (valid.norm.df$covered 1).  
## Area under the curve: 81.67%

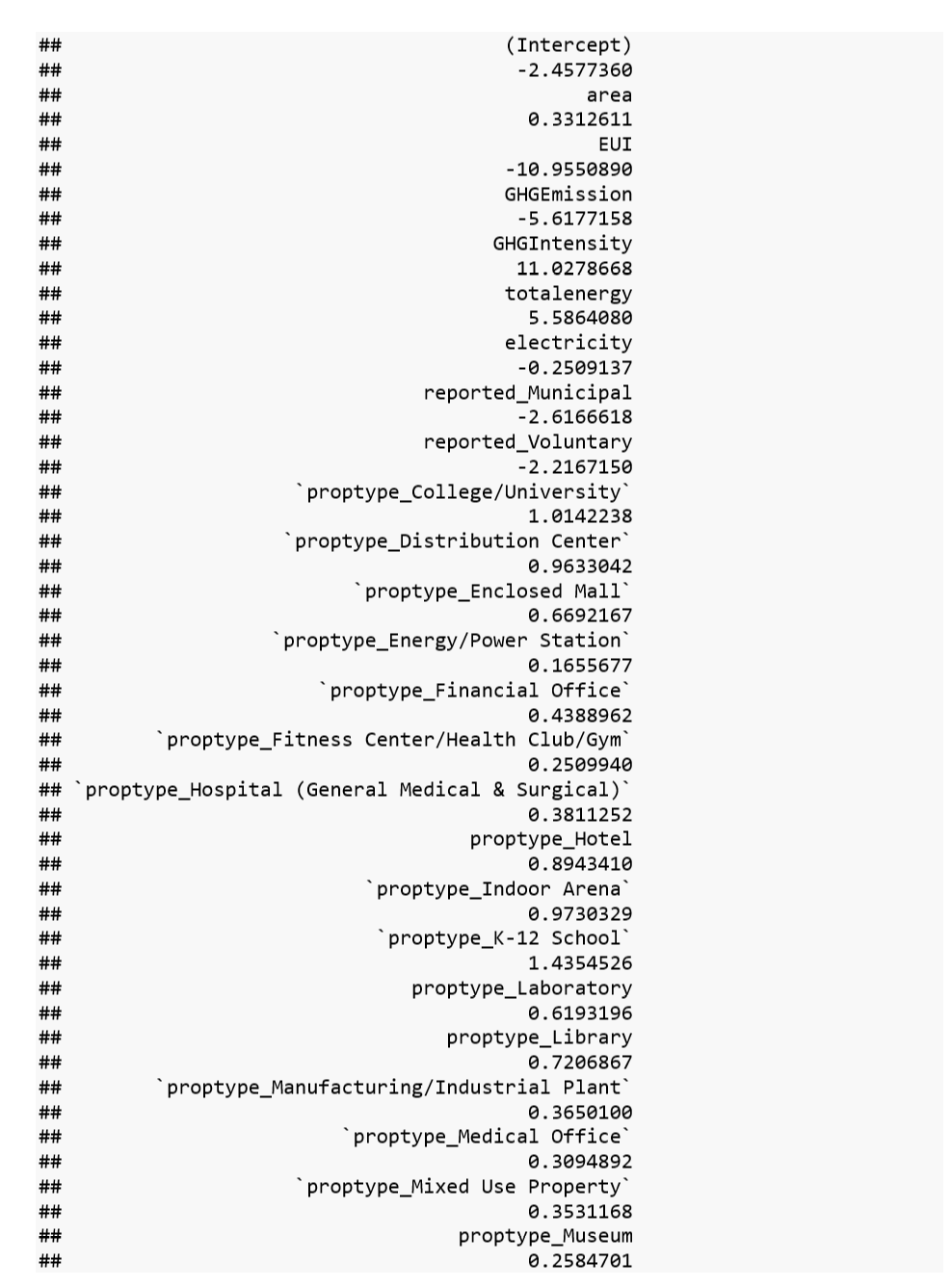
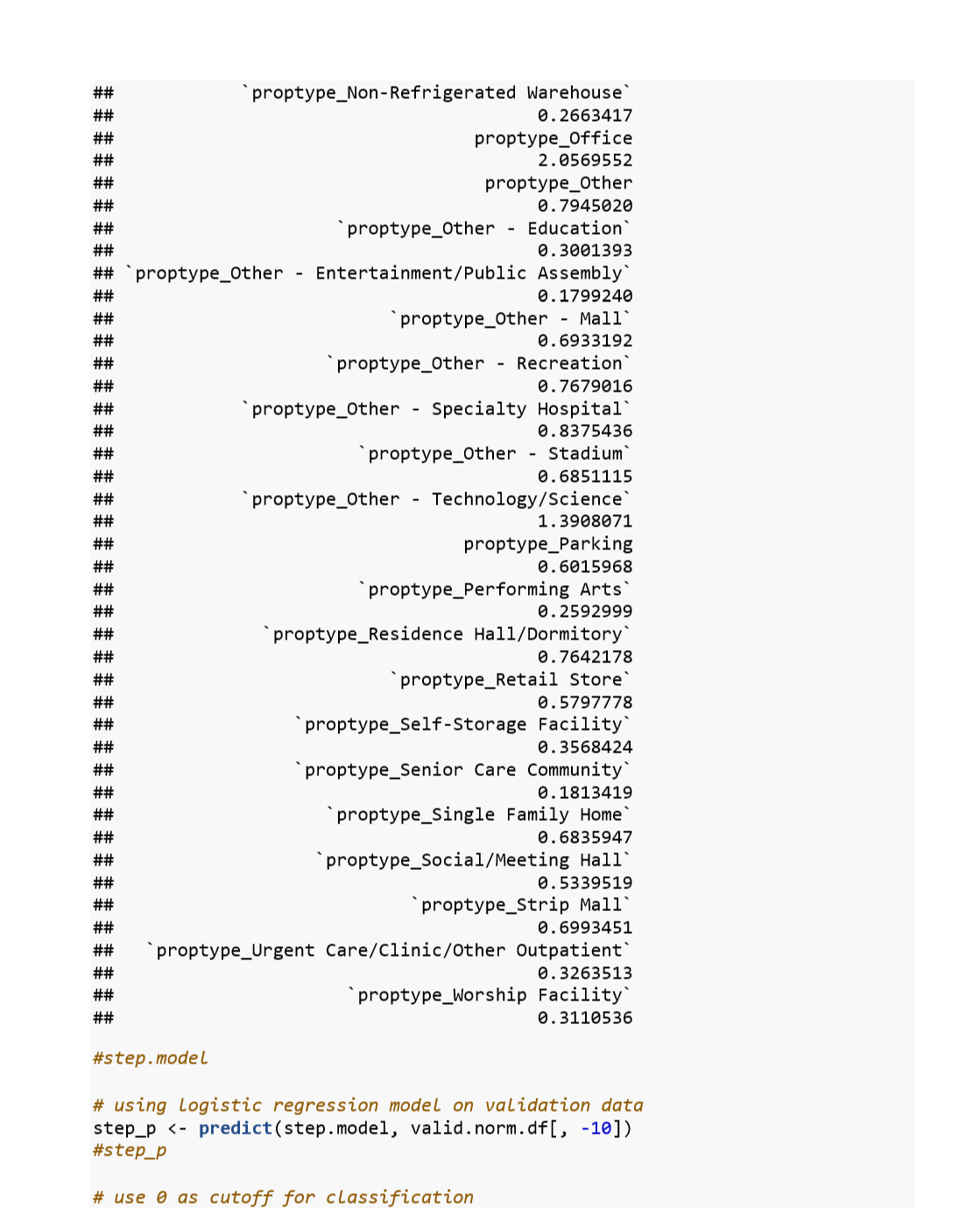
#from above we get that auc was 82%

***MODEL CONCLUSION***

Thus we conclude that logistic regression model is better compared to knn  
 #since area covered is more as well as the accuract ,  
# we can also see that from the 10 fold cross validation that we did above where  
# accuracy of logistic was better than accuracy of knn .

#  
  
  
##Insight for decision making   
#thus logistic regression model can be used to clasify whether the certain facility is   
#covered in the 5 year energy accessment plan making the work of the deparment easier

**Insights for decision making**

#More the GHG(greenhouse gas emission ) less chances of bring covered .

#more the GHG intensity more the chances of being covered .(it has the highest impact of all )

#more the total energy consumption more the chances of being covered and it has good effect on the model .

#if the facility is reported by municipal or voluntarily then there is more chance of them being not being covered in the assessment program .

#More the EUI less will be the chance of being covered in the model and it will decrease rapidly as it increases.

#More the electricity consumption less will be the chance of being covered

**Possible outcomes :**

From the logistic regression coefficients we get to know that office ,K-12 School , Technology /science facility, Specialty hospital, College university and hotels have more chances of being covered under the 5 year energy assessment plan rather than other buildings .

And senior care community center ,Entertainment and public facility non-refrigetated warehouse has less chance of being covered for the 5 year assessment plan

So the department of innovation can ensure that for the next year they focus more on the facilities which have less chances of being covered this year from our model and less focus on facilities which have higher coefficient since they are already covered in assessment plan.

Thus the department can ensure the maximum compliance from the facilities thus ensuring almost accurate demand of power ,water from the sources and hence ensuring the facility reduce their energy usage and reduced GHG emissions.