Building Energy Analysis of Western Pennsylvania Municipal Buildings

Karthik Seeganahalli
Yash Kumar

Introduction

There are 139 municipal buildings in Western PA, which regularly upload their building energy data to the Western PA Datacenter online for free and non-commercial use. We are attempting to analyze this energy data through various predictive, inferential and machine-learning techniques. The data spans 6 years from 2009-14 and includes data on electricity use, natural gas (NG) use, energy costs, Greenhouse Gas (GHG) Emissions. The analysis has been done using R. Through the results of this analysis, recommendations will be made to the Western Pennsylvania Municipalities on how they should try to improve their emission rates and control their electricity use. Also, we will try to predict their energy costs so that they can better allocate financial resources according to their requirements.

Data Source: Western Pennsylvania Regional Data Center

Data Cleaning and Manipulation

Initialization:

```
#Yash PC
# setwd("C:/Users/Yash Kumar/Desktop/Studies/Fall 2017/Data Analytics/Assignments/Fina

# Karthik PC
setwd("~/Desktop/Fall_Courses/Data_Analytics/Final_Project")

library(dplyr)
library(stringr)
library(readxl)
library(readr)
library(Hmisc)
library(gridExtra)
library(knitr)
```

The raw data is mostly clean. The null values have "Not Available" written on them. So, we will convert them to NA values. Then the data columns were converted to the appropriate numeric and factor levels. Most of the data columns were numeric in nature. Only the Property ID, Property Name and the Address are factors. Another modification that is required is that the yearly data is specified as a date. So, using data modification, that has been changed into the number format and then into a numeric category.

```
B_E1 <- read.csv(file="MBEU_2009-14.csv",header=TRUE,sep="\t",
                 na.strings = c("Not Available"))

B_E <- B_E1

B_E$Property_Id <- as.factor(B_E$Property_Id)
B_E$Property_Name <- as.factor(B_E$Property_Name)
B_E$Address_1 <- as.factor(B_E$Address_1)
B_E$Property_GFA_ftsq <- as.numeric(B_E$Property_GFA_ftsq)
B_E$Site_EUI_kBtu.ftsq <- as.numeric(B_E$Site_EUI_kBtu.ftsq)
B_E$WN_Site_EUI_kBtu.ftsq <- as.numeric(B_E$WN_Site_EUI_kBtu.ftsq)
B_E$Source_EUI_kBtu.ftsq <- as.numeric(B_E$Source_EUI_kBtu.ftsq)
B_E$WN_Source_EUI_kBtu.ftsq <- as.numeric(B_E$WN_Source_EUI_kBtu.ftsq)
B_E$Electricity_Use_GP_kBtu <- as.numeric(B_E$Electricity_Use_GP_kBtu)
B_E$Electricity_Use_GP_kWh <- as.numeric(B_E$Electricity_Use_GP_kWh)
B_E$WN_Site_Electricity_kWh <- as.numeric(B_E$WN_Site_Electricity_kWh)
B_E$NG_Use_kBtu <- as.numeric(B_E$NG_Use_kBtu)
B_E$NG_Use_therms <- as.numeric(B_E$NG_Use_therms)
B_E$WN_Site_NG_Use_therms <- as.numeric(B_E$WN_Site_NG_Use_therms)
B_E$District_Steam_Use_kBtu <- as.numeric(B_E$District_Steam_Use_kBtu)
B_E$GHG_tonnes_CO2 <- as.numeric(B_E$GHG_tonnes_CO2)
B_E$GHG_Emissions_Intensity_kgCO2e.ftsq <- as.numeric(B_E$GHG_Emissions_Intensity_kgCO2e
B_E$Energy_Cost_USD <- as.numeric(B_E$Energy_Cost_USD)
B_E$Energy_Cost_Intensity_USD.ftsq <- as.numeric(B_E$Energy_Cost_Intensity_USD.ftsq)
```

```
B_E$Year_Ending <- str_replace(B_E$Year_Ending, "30-09-","20")
B_E$Year_Ending <- as.numeric(B_E$Year_Ending)

write.csv(B_E,file = "Modified_data.csv")
```

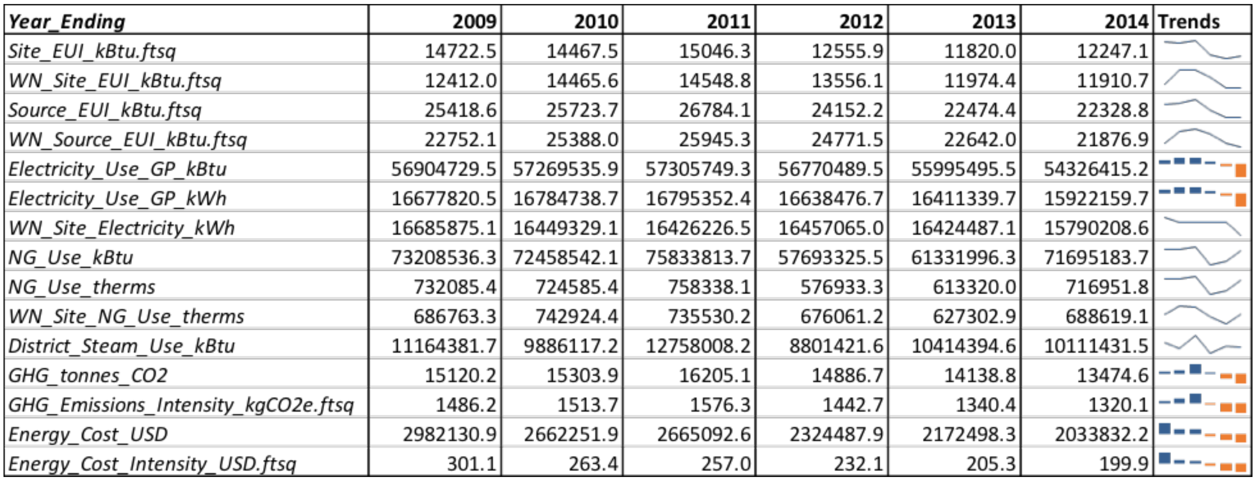
Exploratory Data Analysis

Yearly Trends

Firstly, we would like to do a descriptive analysis to understand how the building energy data is changing throughout the six years. Our initial data analysis question is: "What are the yearly trends for the energy intensities, energy use and energy costs for the Western PA Municipal Buildings?" For this, we took the sum of the data over the years, then compared each value to the mean data point and accordingly figured out the trends over the six years. This was done using the following R code:

```
Descriptive <- B_E[,c(4:ncol(B_E))] %>%
  group_by(Year_Ending) %>%
  summarise_all(funs(sum), na.rm = TRUE)
```

As it is descriptive data, we would like there to be the appropriate values mentioned along with the general trends. For this, we will use sparklines in Excel to denote the specific trends. The electricity use, emissions and energy costs have been shown as diverging [Figure 1].



Year_Ending	2009	2010	2011	2012	2013	2014	Trends
Site_EUI_kBtu.ftsq	14722.5	14467.5	15046.3	12555.9	11820.0	12247.1	
WN_Site_EUI_kBtu.ftsq	12412.0	14465.6	14548.8	13556.1	11974.4	11910.7	
Source_EUI_kBtu.ftsq	25418.6	25723.7	26784.1	24152.2	22474.4	22328.8	
WN_Source_EUI_kBtu.ftsq	22752.1	25388.0	25945.3	24771.5	22642.0	21876.9	
Electricity_Use_GP_kBtu	56904729.5	57269535.9	57305749.3	56770489.5	55995495.5	54326415.2	
Electricity_Use_GP_kWh	16677820.5	16784738.7	16795352.4	16638476.7	16411339.7	15922159.7	
WN_Site_Electricity_kWh	16685875.1	16449329.1	16426226.5	16457065.0	16424487.1	15790208.6	
NG_Use_kBtu	73208536.3	72458542.1	75833813.7	57693325.5	61331996.3	71695183.7	
NG_Use_therms	732085.4	724585.4	758338.1	576933.3	613320.0	716951.8	
WN_Site_NG_Use_therms	686763.3	742924.4	735530.2	676061.2	627302.9	688619.1	
District_Steam_Use_kBtu	11164381.7	9886117.2	12758008.2	8801421.6	10414394.6	10111431.5	
GHG_tonnes_CO2	15120.2	15303.9	16205.1	14886.7	14138.8	13474.6	
GHG_Emissions_Intensity_kgCO2e.ftsq	1486.2	1513.7	1576.3	1442.7	1340.4	1320.1	
Energy_Cost_USD	2982130.9	2662251.9	2665092.6	2324487.9	2172498.3	2033832.2	
Energy_Cost_Intensity_USD.ftsq	301.1	263.4	257.0	232.1	205.3	199.9	

Figure 1: Year wise trends of Key Characteristics

Most of the data is showing a net downwards trend. However, for the years 2010-11, there is a slight increasing trend for the Energy Use Intensities and GHG emissions. Also, this trend was not applicable to for total natural gas use or total electricity use. This points to the fact that GHG emissions are not only calibrated by total energy use or total natural gas, but there is some other underlying reason for the GHG emissions to go up, which might have to do with the energy mix of the region. It is interesting to note that PA retired coal plants with a total rating of 588 MW in 2011 which may have contributed to the decline in GHG emissions later.

Dashboard

To do proper exploratory analysis, it is important that we are able to abstract the proper data and explore it dynamically. For this purpose, we have created a dashboard in Tableau which will abstract the important data and allow dynamic filters for uncomplicated trend analysis. For this, we first need to observe which are the important parameters governing building energy usage.

Energy Use compared to Property Area

The biggest predictor of energy use should be property area. The bigger the property area, the larger will be the electrical and natural gas consumption on average. This is because more electricity will be required for lighting and powering the area. Also, the higher the area, more will be the heating requirements which will increase the natural gas usage. So, using this chart, it will be easy to check the trends of how increasing property area will increase requirements of energy, for both natural gas and electricity. Also, as there are 139 municipal buildings, the chart is a histogram type, with bins created for a range of average property GFA against which the energy use has been compared. Also, accordingly, we have calculated the median and interquartile ranges. Those bins belonging above the median line use more than the median building energy and sets a measurable benchmark for each average building to follow. Thus, it follows the SMART criteria as it is a specific value against which the comparison can be made.

The same follows for the interquartile ranges. Also, there is a filter for each year, so the trends can be observed to see how the energy use for each bin is changing over the years. Energy Cost compared to Energy Use Another metric which should be kept in mind is how much the annual energy cost is compared to their energy use. This is because there are many buildings which may be charged higher rate of electricity or natural gas depending on the utility they use. Thus, this graph will provide a benchmark for the buildings to see how much their energy cost is coming out compared to the energy use. This energy is assumed to be the sum of electricity use and Natural Gas use. Both elements are in kBTU (1000 British Thermal Units), so no conversion is required. Again, this can be filtered for each year to understand how the cost is changing over the years for each energy range.

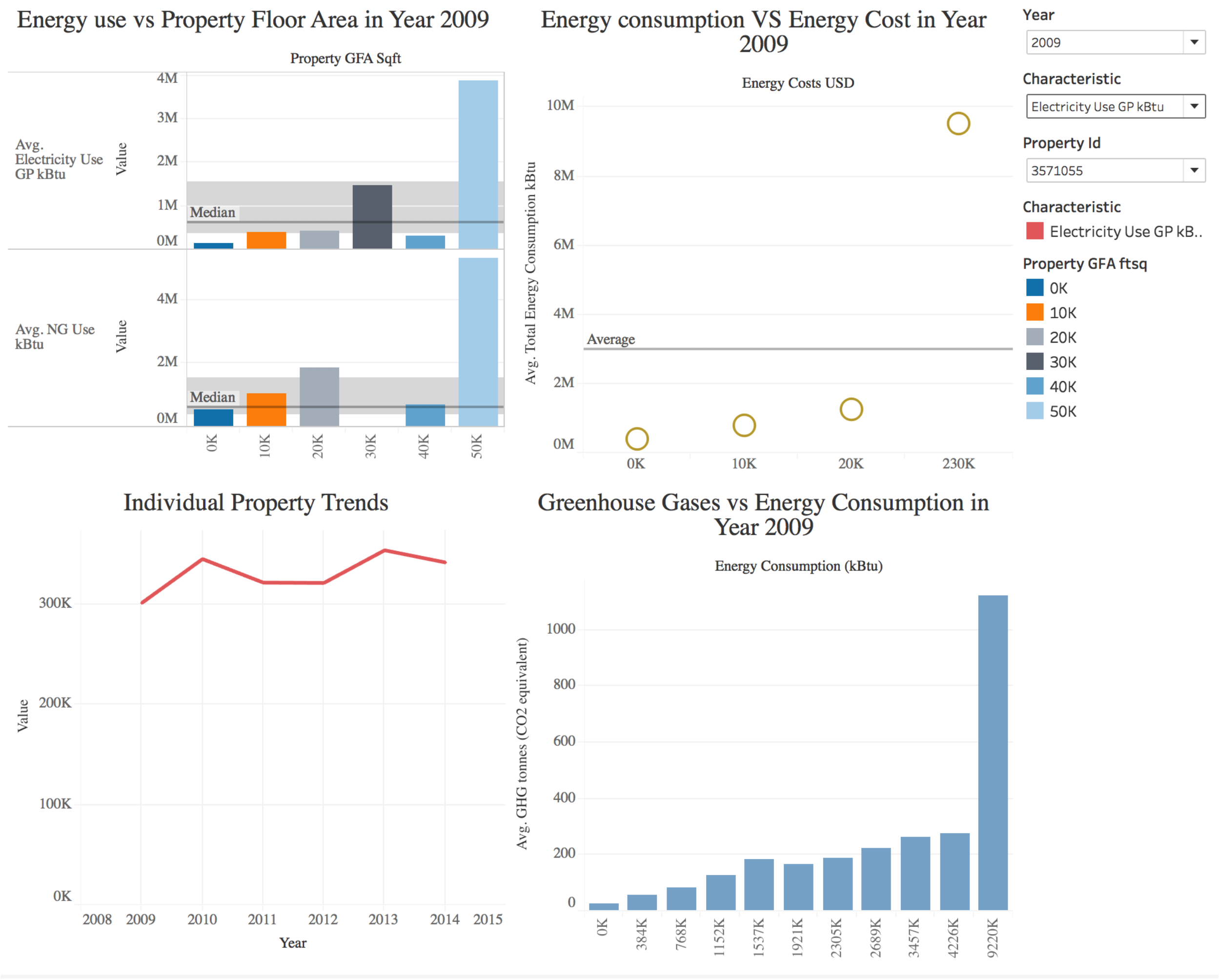
Year-wise Individual Property Trends

Up till now, all the charts, referred to a range of properties. Many people are not concerned with the general trends and want to check trends of an individual property over the years. Also, it might be that someone wants to check how their individual property compared to other properties in a similar range. For this, we have created a graph which shows the individual characteristics of one property. The characteristics include: Electricity Use, Energy Cost, GHG emissions, Natural Gas Use, Property GFA and Total Energy Consumption.

GHG Emissions vs Energy Use

The final most important characteristic is the GHG emissions. The most important reason for its increase is the energy use. So, we are comparing the GHG emissions with the overall energy consumption (sum of Natural Gas use and electricity use). Again, we are filtering through years to get a year-wise trend for each energy range.

A snapshot of the dashboard has been included in figure 2.



Year
2009

Characteristic
Electricity Use GP kBtu

Property Id
3571055

Figure 2: Dashboard

Statistical Analysis

Electricity Use analysis

Now, we wanted to analyze whether it is possible to predicted using a single parameter how the electricity use changes. The most obvious in the given parameters is the property Gross Feet Area (GFA). It is plausible to assume that higher the area, the higher the electricity required will be. This can be seen with the following graph:

```r
ggplot(data = B_E, aes(x = Property_GFA_ftsq, y = WN_Site_Electricity_kWh)) +
  geom_point(colour="blue3") + theme_classic() +
  labs(x="Property Gross Floor Area Sqft",
       y="Weather Normalized Site Electricity (kWh)") +
  theme(axis.title = element_text("Times","bold.italic","black",9,hjust = 0.5),
        legend.text = element_text("Times"),
        legend.title = element_text("Times"),
        axis.text = element_text("Times"))
```

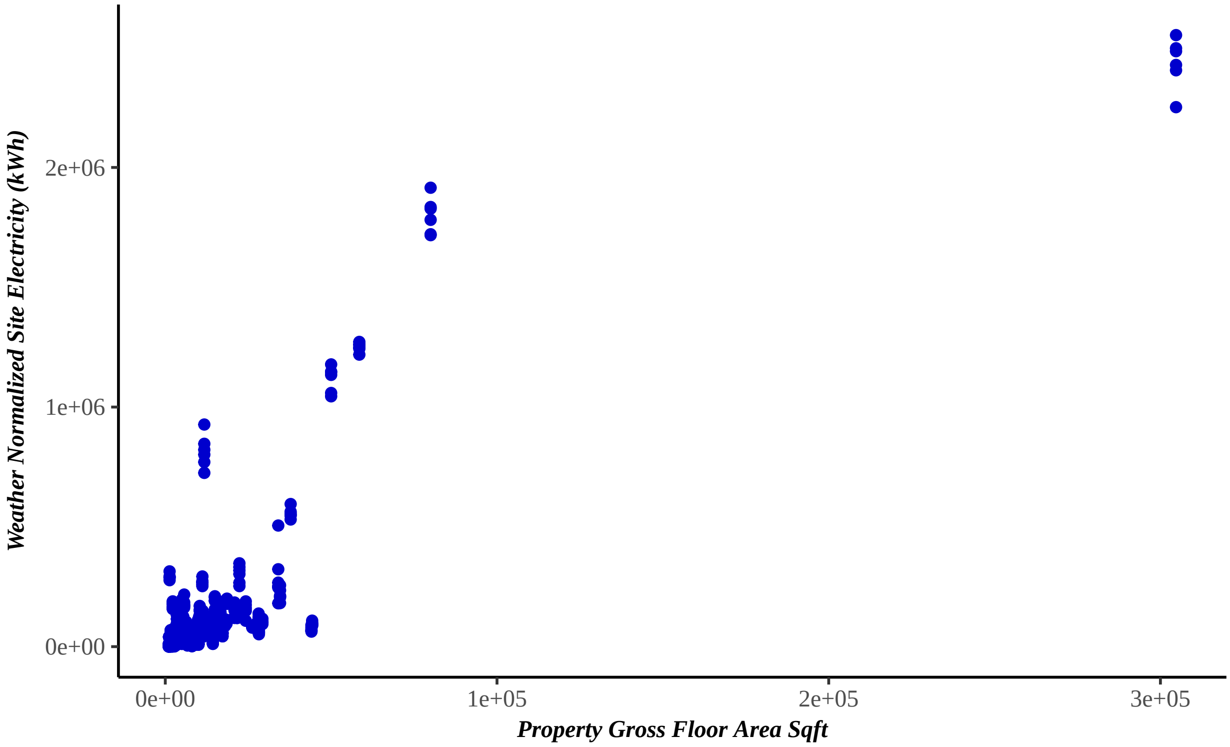


Figure 3: Energy Cost Variation with Property Area

There is clearly an upward trend. Thus, the first of our statistical analysis question is: "Can it be predicted with statistical certainty, how the property area affects the electricity use for Western PA municipal buildings, and what is the predictive model?" The important thing to consider here is whether it can be predicted accurately what the electricity use will be based on the property GFA. This can be tested in the following way: Firstly, we divide the original data into training and test data. Then, we test the value predicted by the training data onto the test data and see if it's good enough.

Data Manipulation

Firstly, as can be seen in the plot above, the data has some vertical aberrations. This is due to the same building having different electricity uses over different years. To normalize this data, we can either do trendline analysis for the average of each building over all the years or we can extract the data for a single year and try to work with that. We will go with the first option as it works over a broader spectrum.

```
new <- B_E %>%
  group_by(Property_Id) %>%
  summarise(mean(WN_Site_Electricity_kWh, na.rm = TRUE),
            mean(Property_GFA_ftsq, na.rm = TRUE))

names(new) <- c("Property_ID", "WN_Site_Electricity_kWh", "Property_GFA_ftsq")

ggplot(data = new, aes(x = Property_GFA_ftsq, y = WN_Site_Electricity_kWh)) +
  geom_point(colour="blue3") + theme_classic() +
  labs(x="Property Gross Floor Area Sqft",
       y="Weather Normalized Site Electricity  (kWh)") +
  theme(axis.title = element_text("Times","bold.italic","black",9,hjust = 0.5),
        legend.text = element_text("Times"), legend.title = element_text("Times"),
        axis.text = element_text("Times"))
```

Now, the data looks cleaner. Then after extracting the relevant data, we separate the data into test and training data. We have chosen 70% and 30% divisions for the total data into training and test data. After that we fit the data for the 70% of training data and fit it to the test data. We will go with the simple linear regression model.

```
mysample <- sample(2, nrow(new), replace=TRUE, prob=c(0.7, 0.3))
trainData <- new[mysample==1,]
```

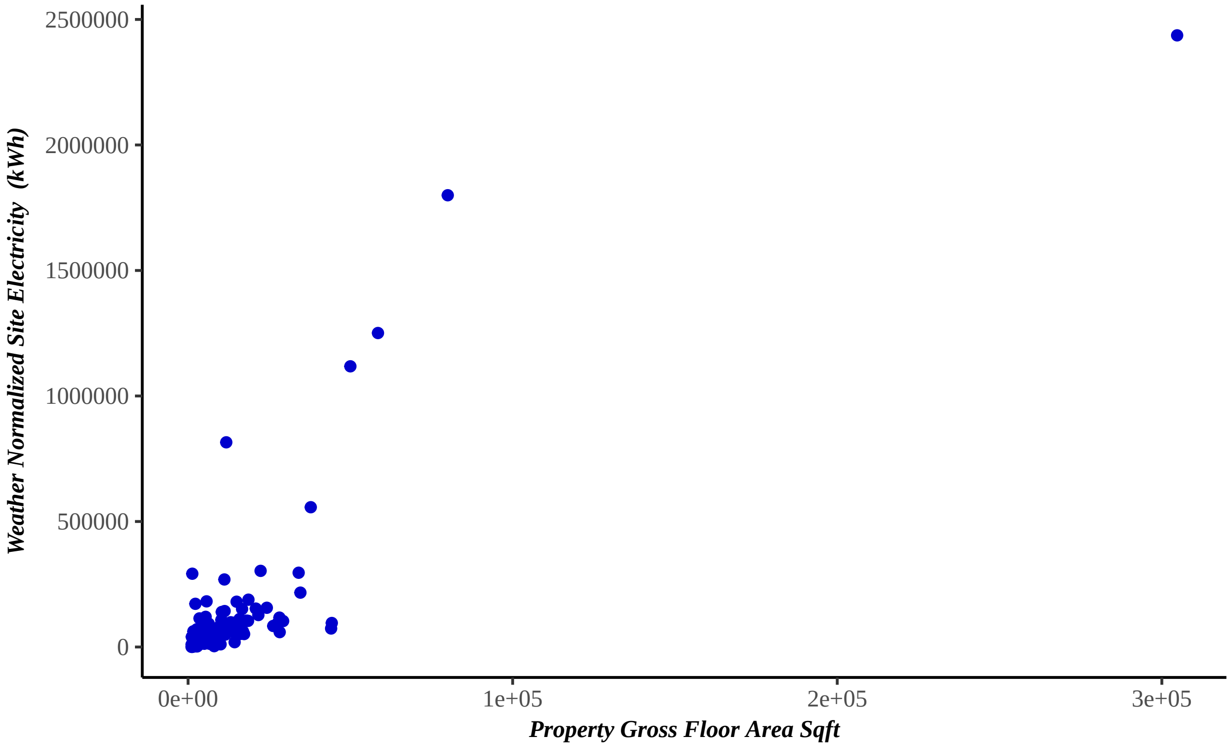


Figure 4: Energy Cost Variation with Property Area

```
testData <- new[mysample==2,]


fit <- lm(trainData$WN_Site_Electricity_kWh ~ trainData$Property_GFA_ftsq)
```

After doing the fitting, we find out how well it fits the training data. This can be done using the summary and plot function.

```
summary(fit)
```

```
##
## Call:
## lm(formula = trainData$WN_Site_Electricity_kWh ~ trainData$Property_GFA_ftsq)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -373507  -56365  -27771    4588 1053948
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.078e+04  2.079e+04   0.518    0.605
## trainData$Property_GFA_ftsq 9.187e+00  5.811e-01  15.811   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 184600 on 93 degrees of freedom
## Multiple R-squared:  0.7289, Adjusted R-squared:  0.7259
## F-statistic:   250 on 1 and 93 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(fit)
```

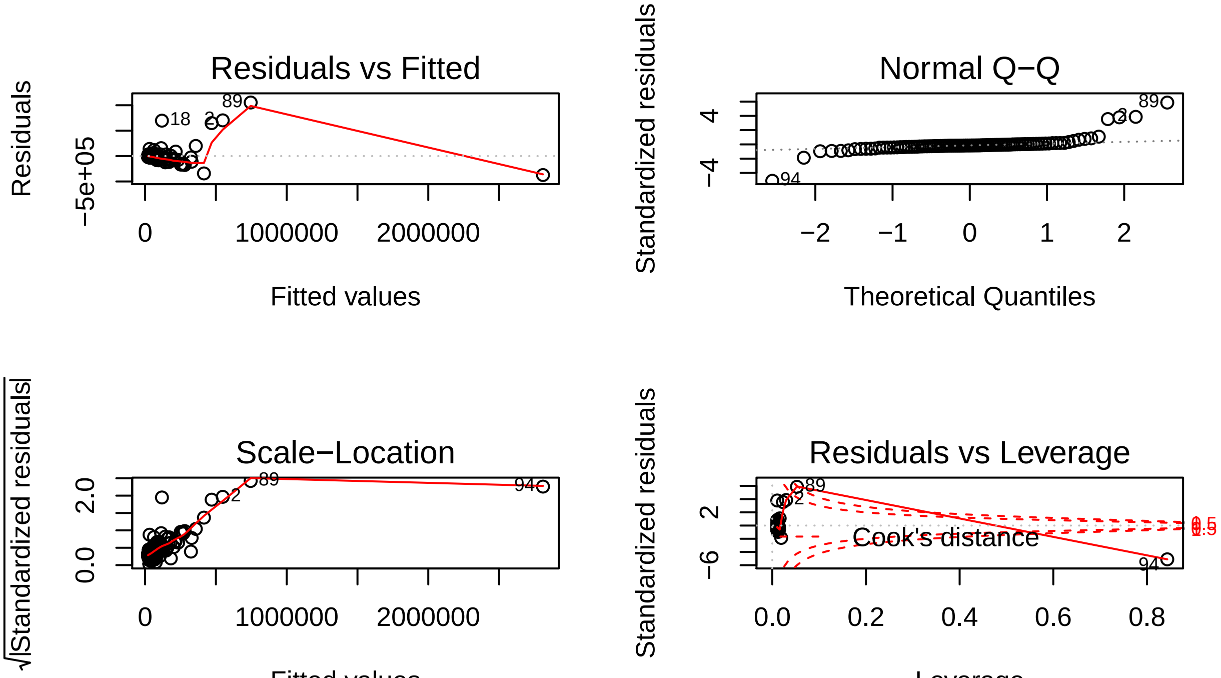


Figure 5: Training Data Validation

It can be seen that the data fits mostly well, considering it is a single parameter. Also, most of the data points are away outside the cook's distance, which makes for a good fit. Then we fit it on the test data and calculate the error between the predicted and the actual values. Using that, we calculate the error bias (mean) and the standard deviation to find the fitting.

```
Co<- coefficients(fit)

testData <- mutate(testData, estimate = (Co[1]+Co[2]*Property_GFA_ftsq))
testData <- mutate(testData, errorsq = (estimate - WN_Site_Electricity_kWh)^2)
```

```
# valueerr = (sum(testData$errorsq,na.rm=T)/nrow(testData))^0.5

rmse<-sqrt(mean(testData$errorsq,na.rm = T))

std_dev <- sd(sqrt(testData$errorsq), na.rm = TRUE)
```

The root mean square error is way too high. Also, the standard deviation is very high. Thus, it seems that although correlations are very strong between electricity use and property area. However, this does not lead to good predictive models. We can further see how the actual points are different from the estimates.

```
p1<-plot(testData$estimate,col='blue', xlab="Index",
         ylab="Weather Normalized Electricity (kWh)",family="Times")
points(testData$WN_Site_Electricity_kWh,col='red',family="Times")

legend("topright",legend=c("Estimate Values", "Actual Values"),pch=c(1,1),
       col=c("blue", "red"), cex=0.8)
```

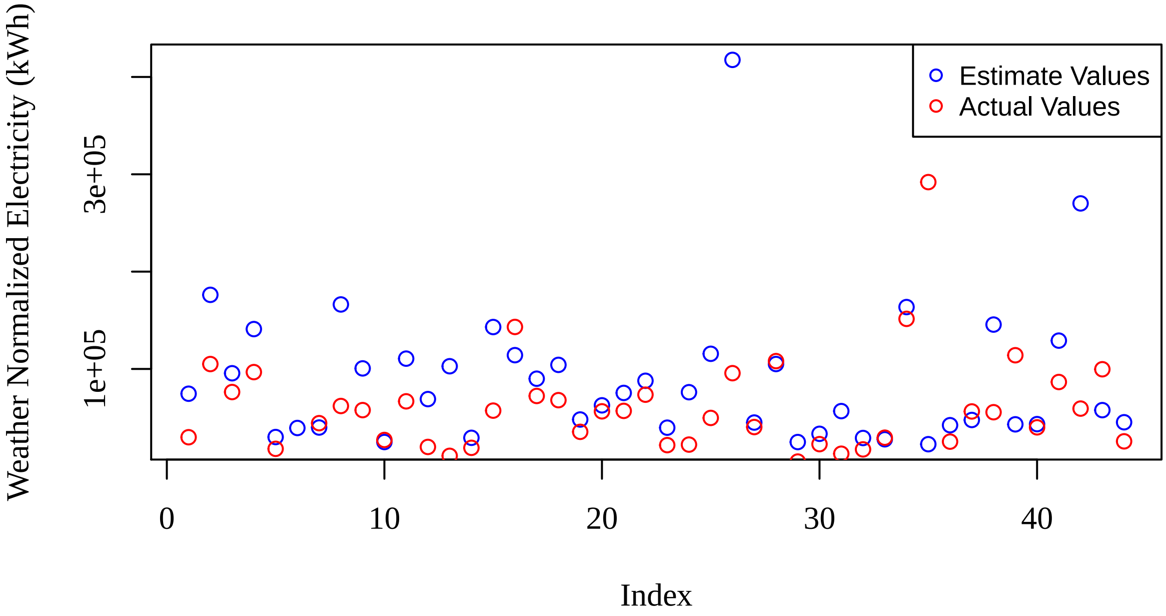


Figure 6: Test Data Validation

As can be seen, although many points seem to overlap and coincide well. However, the error magnitudes are of the order 10^6, which is way too high. Thus, it seems very unlikely that you can predict the Electricity Use, based on Property Area alone.

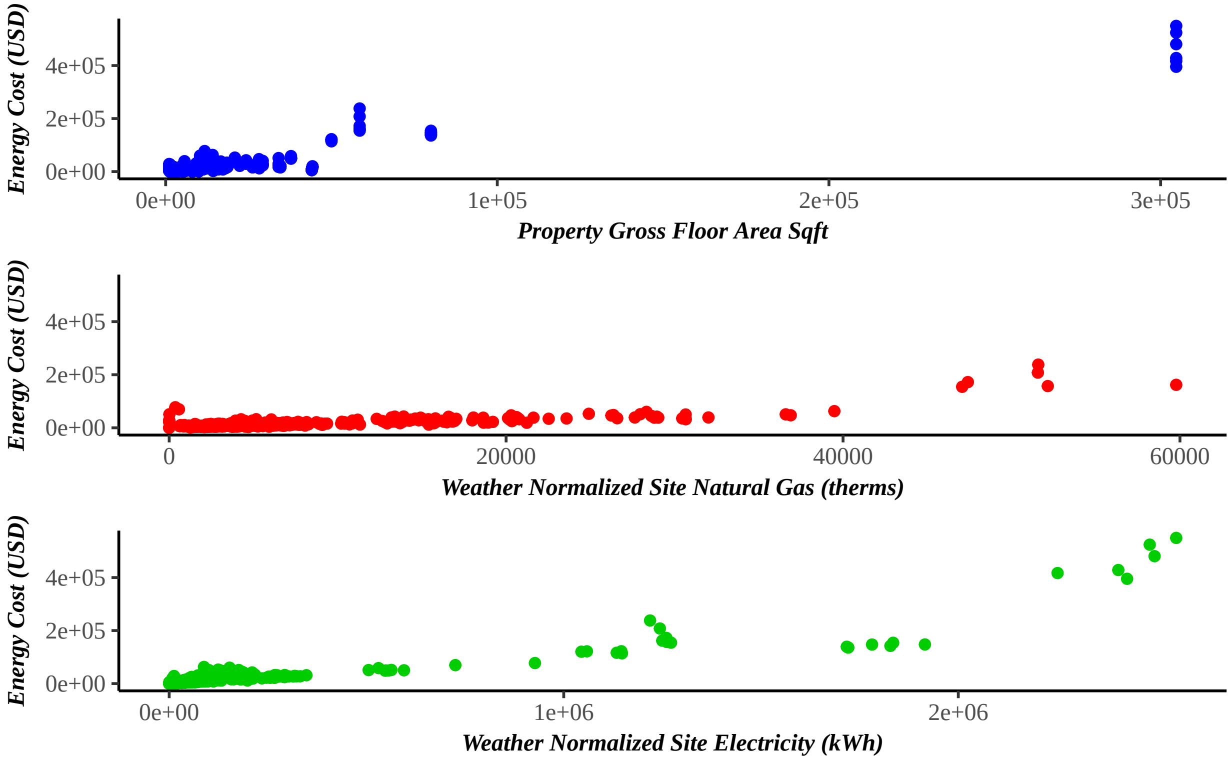
Energy Cost Analysis

The second question we will try to answer is to infer how the the energy cost changes with electricity use, natural gas use and/or property area. The initial exploratory analysis for each component shows the following plots:

```r
p1 <- ggplot(data = B_E, aes(x = Property_GFA_ftsq, y = Energy_Cost_USD)) +
  geom_point(colour="blue") +theme_classic() +
  labs(x="Property Gross Floor Area Sqft",y="Energy Cost (USD)") +
  theme(axis.title = element_text("Times","bold.italic","black",9,hjust = 0.5),
        legend.text = element_text("Times"), legend.title = element_text("Times"),
        axis.text = element_text("Times"))

p2 <- ggplot(data = B_E, aes(x = WN_Site_NG_Use_therms, y = Energy_Cost_USD)) +
  geom_point(colour="red") + theme_classic() +
  labs(x="Weather Normalized Site Natural Gas (therms)",
       y="Energy Cost (USD)") +
  theme(axis.title = element_text("Times","bold.italic","black",9,hjust = 0.5),
        legend.text = element_text("Times"), legend.title = element_text("Times"),
        axis.text = element_text("Times"))

p3 <- ggplot(data = B_E, aes(x = WN_Site_Electricity_kWh, y = Energy_Cost_USD)) +
  geom_point(colour="green3")+ theme_classic()+
  labs(x="Weather Normalized Site Electricity (kWh)",y="Energy Cost (USD)") +
  theme(axis.title = element_text("Times","bold.italic","black",9,hjust = 0.5),
        legend.text = element_text("Times"), legend.title = element_text("Times"),
        axis.text = element_text("Times"))
```



There is a linear trend seen for the Property Area and Natural Gas Use, also there seems to be a quadratic or exponential trend for electricity usage. Thus, we deem our second question will be:

"Can anything be inferred about the annual energy cost from electricity use, natural gas use and property area and what is the inference?"

For this, we will try to fit the energy cost with the property area, electricity use and natural gas use. We will do three individual fits with each of the parameters. Then, we try pairwise fits with the three parameters. Then, finally, we a fit with all three parameters together. The model which gives us the best fit will be our inference.

There are a total of 7 fits required. We found out that that there are two models which stand out as the best. The other fits have been described in the Appendix. The best fit is, unexpectedly, fitting the energy cost along with electricity use and natural gas use. It does not involve the Property GFA at all.

```
New <- B_E[,c(19,15,12,4)]

New$Energy_Cost_USD[which(!is.finite(New$Energy_Cost_USD))] = NA
New$WN_Site_NG_Use_therms[which(!is.finite(New$WN_Site_NG_Use_therms))] = NA
New$Property_GFA_ftsq[which(!is.finite(New$Property_GFA_ftsq))] = NA
```

```r
New$WN_Site_Electricity_kWh[which(!is.finite(New$WN_Site_Electricity_kWh))] = NA

fit7 <- lm(Energy_Cost_USD ~ WN_Site_NG_Use_therms+WN_Site_Electricity_kWh^2,
           na.action=na.exclude, data = New)
summary(fit7)
```

```
## 
## Call:
## lm(formula = Energy_Cost_USD ~ WN_Site_NG_Use_therms + WN_Site_Electricity_kWh^2, 
##     data = New, na.action = na.exclude)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -26231  -1467   -469   1271  61891 
## 
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             1.967e+02  2.759e+02   0.713    0.476    
## WN_Site_NG_Use_therms   1.092e+00  3.137e-02  34.797   <2e-16 ***
## WN_Site_Electricity_kWh 9.817e-02  1.723e-03  56.963   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4483 on 521 degrees of freedom
##   (310 observations deleted due to missingness)
## Multiple R-squared:  0.9549, Adjusted R-squared:  0.9548 
## F-statistic:  5522 on 2 and 521 DF,  p-value: < 2.2e-16
```

When we create a fit for Energy Cost using only electricity usage and natural gas usage, we get the the best R-squared value, which suggests that NG use and Electricity use totally describe the energy cost. It is interesting to note that the t-value of electricity is higher than that of Natural Gas usage. This can indicate the fact that precision of electricity use's coefficient is more as compared to the coefficient of the NG use. This is despite the fact that Natural Gas costs are higher i.e. a therm of NG (approx USD 0.40/therm) costs more than a kWh of electricity. (approx USD 0.10/kWh)

The second best fit is that which uses all the three parameters.

```r
fit3 <- lm(Energy_Cost_USD ~ WN_Site_NG_Use_therms + Property_GFA_ftsq +
             WN_Site_Electricity_kWh^2, na.action=na.exclude, data = New)
summary(fit3)
```

```
##
## Call:
## lm(formula = Energy_Cost_USD ~ WN_Site_NG_Use_therms + Property_GFA_ftsq +
##     WN_Site_Electricity_kWh^2, data = New, na.action = na.exclude)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -26270  -1468   -482   1274  61877
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             225.927831 314.988379   0.717    0.474
## WN_Site_NG_Use_therms     1.093891   0.033710  32.450   <2e-16 ***
## Property_GFA_ftsq        -0.005670   0.029330  -0.193    0.847
## WN_Site_Electricity_kWh   0.098327   0.001912  51.425   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4487 on 520 degrees of freedom
##   (310 observations deleted due to missingness)
## Multiple R-squared:  0.955,  Adjusted R-squared:  0.9547
## F-statistic:  3674 on 3 and 520 DF,  p-value: < 2.2e-16
```

It can be seen that the R-squared value is very good. However, the property GFA is not very useful in the calculation. The t-value for Property area is very bad. This gives the sense that Energy cost is also not highly dependent on Property Area, which may arise from different energy efficiencies, insulation and weather conditions of the area. This means that large properties can also have a low energy cost.

Categorization of buildings

For our final part of the analysis, we would like employ machine learning techniques to do some useful analysis. We will employ the K-means clustering analysis technique to categorize buildings. This will be done based on property area and annual energy cost. The major reason for using property area is that real estate values area very highly. Thus, a premium is given to each incremental square feet area of real estate. Thus, it is a very important feature of buildings. Also, we have ascertained that energy prices are the standard to compare as it is allocated a significant chunk of financial resources. From this categorization of clusters, we will finally find out the mean natural gas usage, electricity usage, property area, energy costs and GHG emissions for each cluster category, which will in turn help agencies such as the Green Building Alliance to build benchmarks and check which buildings are performing worse than average in emissions or electricity consumption.

For this, we need to average out the data for all the years, so that some properties do not fall in separate clusters for separate years. Then, the NA values are omitted and data clustering is done. The clusters looks something like this. These cluster factors are later binded to the original dataset so that grouping by clusters can be done.

```
new1 <- B_E[!(is.na(B_E$Energy_Cost_USD)),]

new2<-new1 %>%
  group_by(Property_Id) %>%
  summarise_all(mean)




cluster_data<-new2

datacluster<-kmeans(cluster_data[,c(4,19)],3)

datacluster$cluster<-as.factor(datacluster$cluster)

c1<-cbind(cluster_data,datacluster$cluster)

ggplot(cluster_data,aes(Property_GFA_ftsq,Energy_Cost_USD,color=datacluster$cluster)) +
  geom_point() + theme_classic() +
  labs(x="Property Gross Floor Area (Sqft)",y="Energy Cost (USD)", colour="Clusters") +
  theme(plot.title = element_text("Times","bold.italic","black",12,hjust = 0.5),
        axis.title =element_text("Times","bold.italic","black",9,hjust = 0.5),
        legend.text = element_text("Times"), legend.title = element_text("Times"),
        axis.text = element_text("Times"))
```

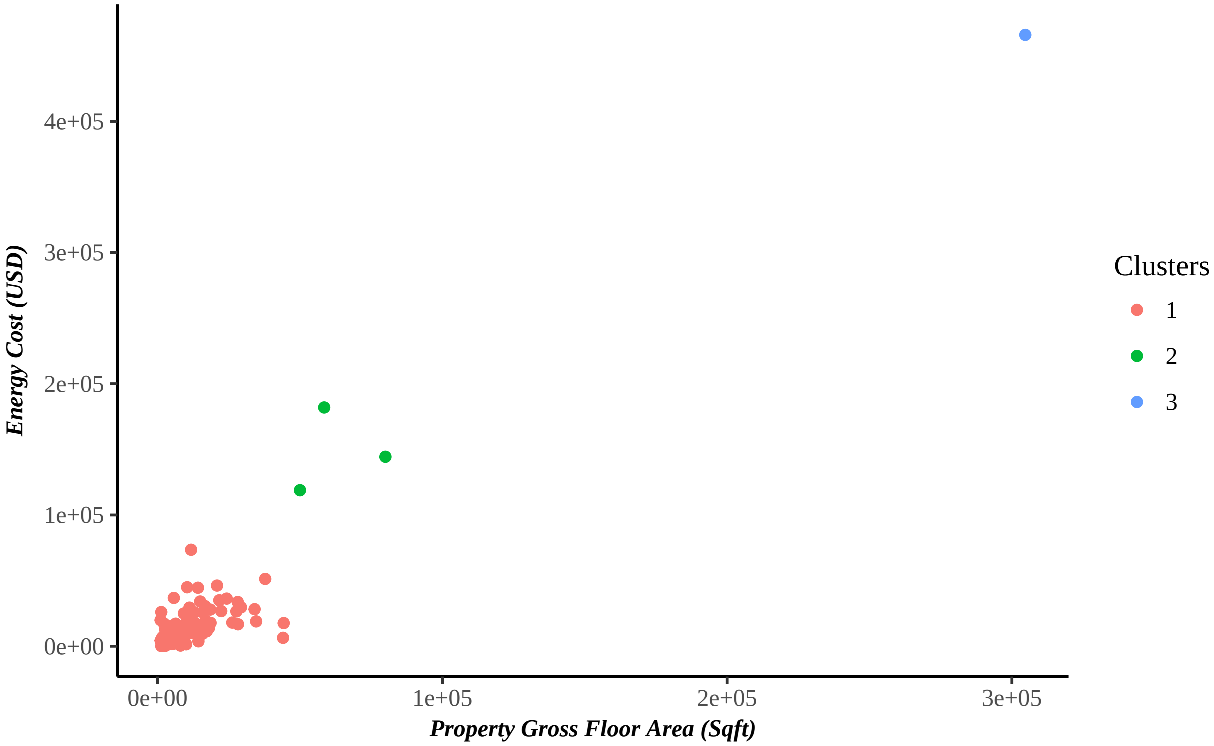


Figure 7: Cluster Analysis

```
new1 <- B_E[!(is.na(B_E$Energy_Cost_USD)),]
```

Three clusters seems appropriate looking at the analysis, as there are three distinct clusters which can be seen in the plot. The second and third cluster have much lesser data points, but the technique to perform the analysis is the same and can be replicated on large datasets. Now, we can get the mean values for each of the data clusters.

```
c1 %>%
  group_by(datacluster$cluster)%>%
  summarise(mean(Electricity_Use_GP_kBtu,na.rm=TRUE),mean(NG_Use_kBtu,na.rm=TRUE),
            mean(GHG_tonnes_CO2,na.rm=TRUE),mean(Energy_Cost_USD,na.rm=TRUE),
            mean(Property_GFA_ftsq,na.rm=TRUE))->Cluster_means
colnames(Cluster_means)<-c('Cluster','Avg_Electricity_Use(kBtu)',
                           'Avg_Natural_Gas_Use(kBtu)','Avg_GHG_Emissions ((tonnes) CO2
                           'Avg_Energy_Cost(USD)','Avg_Gross_Floor_Area(Sqft)')
kable(Cluster_means[,1:3],
      caption="Average Characteristics of the Building Categories")
```

Table 1: Average Characteristics of the Building Categories

Cluster	Avg_Electricity_Use(kBtu)	Avg_Natural_Gas_Use(kBtu)
1	252365.4	715548.6
2	4764597.6	5120039.7
3	8387387.1	NaN

```
kable(Cluster_means[,4:6],
      caption="Average Characteristics of the Building Categories")
```

Table 2: Average Characteristics of the Building Categories

Avg_GHG_Emissions ((tonnes) CO2 e)	Avg_Energy_Cost(USD)	Avg_Gross_Floor_Area(Sqft)
76.8998	12818.3	9459.015
1064.1078	148394.6	62833.333
2343.3333	465867.6	304710.000

There is an NA value for the NG use because the the second cluster has only one building, and its Natural Gas use has not been mentioned. However, these results may prove as a useful benchmark for the various building sizes and their energy costs, emissions etc.

Conclusion and Recommendations

After an extensive analysis, we figured that the correlations are not straightforward between costs, energy use and property area. This is because there are various additional parameters which influence energy use, costs and emissions. However, there are some distinct trends which stand out and allow for analysis. This analysis is by no means perfect and there is always room for improvement.

Based on our analysis, we were able to prove that predictions regarding electricity use and property GFA are very hard to make. However, we were able to find out very accurate estimates regarding energy costs using total natural gas use and electricity use. This makes sense as the electricity cost per kWh and energy costs per therm of Natural Gas throughout the Western PA should be constant. Notably, we categorized buildings along the Property GFA and Energy Costs and created a benchmark for other buildings to be compared against.

Based on our analysis, we have the following conclusions and recommendations:

1) Buildings should follow LEED certification. The property area is not a good indiccator of energy costs. These means if buildings follow the right practices. They can immensely control their energy costs.

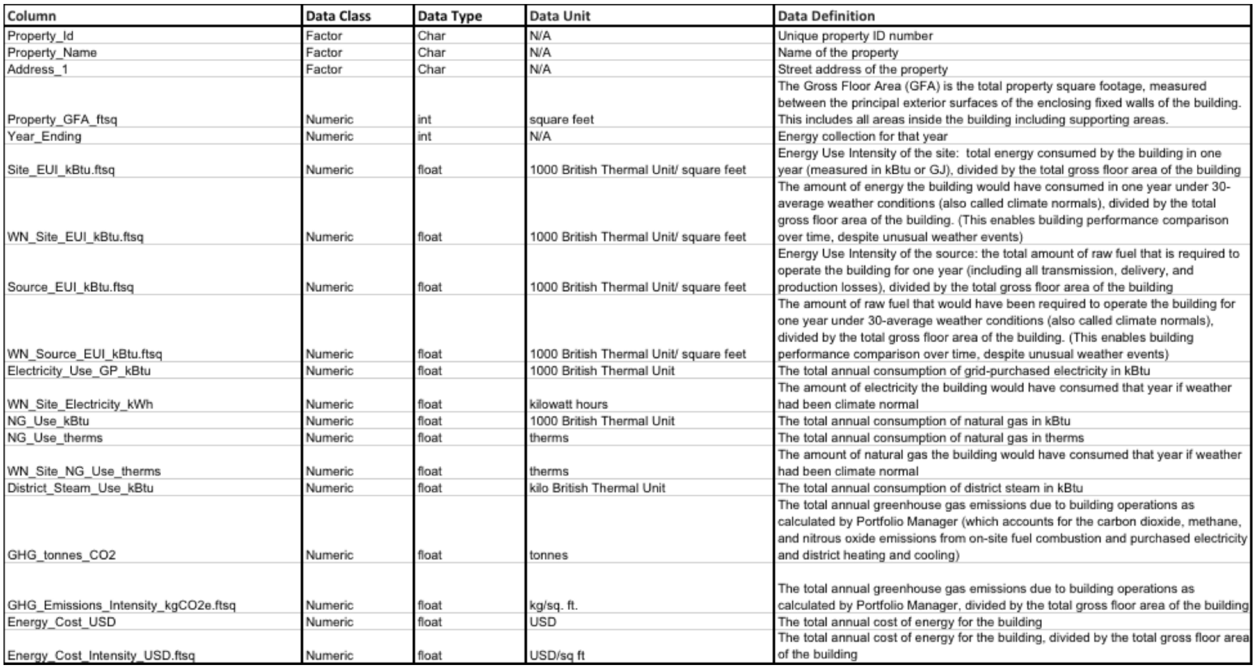
2) It is interesting to note that GHG emissions are not completely correlated with the energy use, when looking at the dashboard. There is some other underlying calculation which is being used to find the GHG emissions and it is only loosely based on the energy mix. This is because the energy mix should be the same for the whole of Western PA and still the GHG emissions are not completely similar.

3) Electricity costs are more predictive of annual energy costs as compared to natural gas. This can be seen from the t-values of the second inferential question. Thus, it is more important to reduce electricity consumption to control energy costs

4) Cluster 3 has 5 times the area of cluster 2. However, its energy costs are only three times and emissions are only two times that of cluster 2. This points to the fact that larger buildings are much more carefully built to conserve energy as compared to larger buildings. However, this inference is made from a very limited dataset.

Appendix

Data Dictionary



Column	Data Class	Data Type	Data Unit	Data Definition
Property_Id	Factor	Char	N/A	Unique property ID number
Property_Name	Factor	Char	N/A	Name of the property
Address_1	Factor	Char	N/A	Street address of the property
Property_GFA_ftsq	Numeric	int	square feet	The Gross Floor Area (GFA) is the total property square footage, measured between the principal exterior surfaces of the enclosing fixed walls of the building. This includes all areas inside the building including supporting areas.
Year_Ending	Numeric	int	N/A	Energy collection for that year
Site_EUI_kBtu.ftsq	Numeric	float	1000 British Thermal Unit/ square feet	Energy Use Intensity of the site: total energy consumed by the building in one year (measured in kBtu or GJ), divided by the total gross floor area of the building
WN_Site_EUI_kBtu.ftsq	Numeric	float	1000 British Thermal Unit/ square feet	The amount of energy the building would have consumed in one year under 30-average weather conditions (also called climate normals), divided by the total gross floor area of the building. (This enables building performance comparison over time, despite unusual weather events)
Source_EUI_kBtu.ftsq	Numeric	float	1000 British Thermal Unit/ square feet	Energy Use Intensity of the source: the total amount of raw fuel that is required to operate the building for one year (including all transmission, delivery, and production losses), divided by the total gross floor area of the building
WN_Source_EUI_kBtu.ftsq	Numeric	float	1000 British Thermal Unit/ square feet	The amount of raw fuel that would have been required to operate the building for one year under 30-average weather conditions (also called climate normals), divided by the total gross floor area of the building. (This enables building performance comparison over time, despite unusual weather events)
Electricity_Use_GP_kBtu	Numeric	float	1000 British Thermal Unit	The total annual consumption of grid-purchased electricity in kBtu
WN_Site_Electricity_kWh	Numeric	float	kilowatt hours	The amount of electricity the building would have consumed that year if weather had been climate normal
NG_Use_kBtu	Numeric	float	1000 British Thermal Unit	The total annual consumption of natural gas in kBtu
NG_Use_therms	Numeric	float	therms	The total annual consumption of natural gas in therms
WN_Site_NG_Use_therms	Numeric	float	therms	The amount of natural gas the building would have consumed that year if weather had been climate normal
District_Steam_Use_kBtu	Numeric	float	kilo British Thermal Unit	The total annual consumption of district steam in kBtu
GHG_tonnes_CO2	Numeric	float	tonnes	The total annual greenhouse gas emissions due to building operations as calculated by Portfolio Manager (which accounts for the carbon dioxide, methane, and nitrous oxide emissions from on-site fuel combustion and purchased electricity and district heating and cooling)
GHG_Emissions_Intensity_kgCO2e.ftsq	Numeric	float	kg/sq. ft.	The total annual greenhouse gas emissions due to building operations as calculated by Portfolio Manager, divided by the total gross floor area of the building
Energy_Cost_USD	Numeric	float	USD	The total annual cost of energy for the building
Energy_Cost_Intensity_USD.ftsq	Numeric	float	USD/sq ft	The total annual cost of energy for the building, divided by the total gross floor area of the building

Figure 8: Data Dictionary

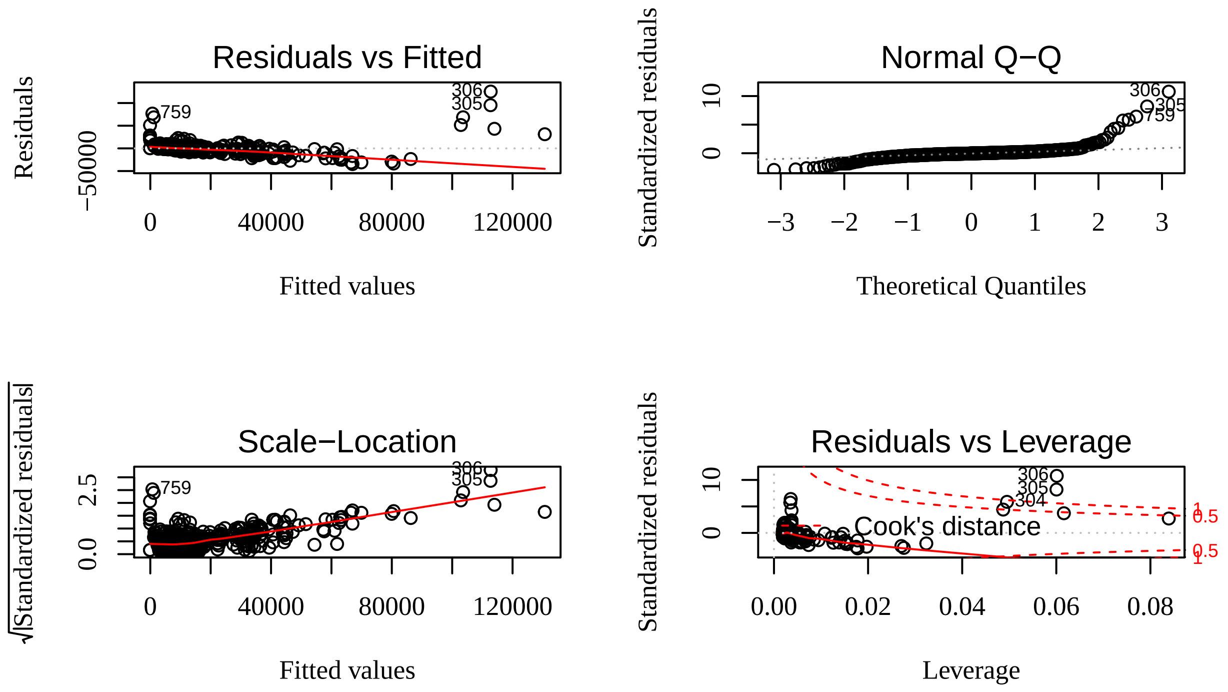
Other Linear Models

```
fit4 <- lm(Energy_Cost_USD ~ WN_Site_NG_Use_therms, na.action=na.exclude, data = New)
summary(fit4)
```

```
##
## Call:
## lm(formula = Energy_Cost_USD ~ WN_Site_NG_Use_therms, data = New,
##     na.action = na.exclude)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34659  -3199   -423   1880 125292
##
## Coefficients:
```

```
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -110.10314  733.09441   -0.15    0.881
## WN_Site_NG_Use_therms      2.18773    0.06607   33.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11990 on 526 degrees of freedom
##   (306 observations deleted due to missingness)
## Multiple R-squared:  0.6758, Adjusted R-squared:  0.6751
## F-statistic:  1096 on 1 and 526 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(fit4,family="Times")
```

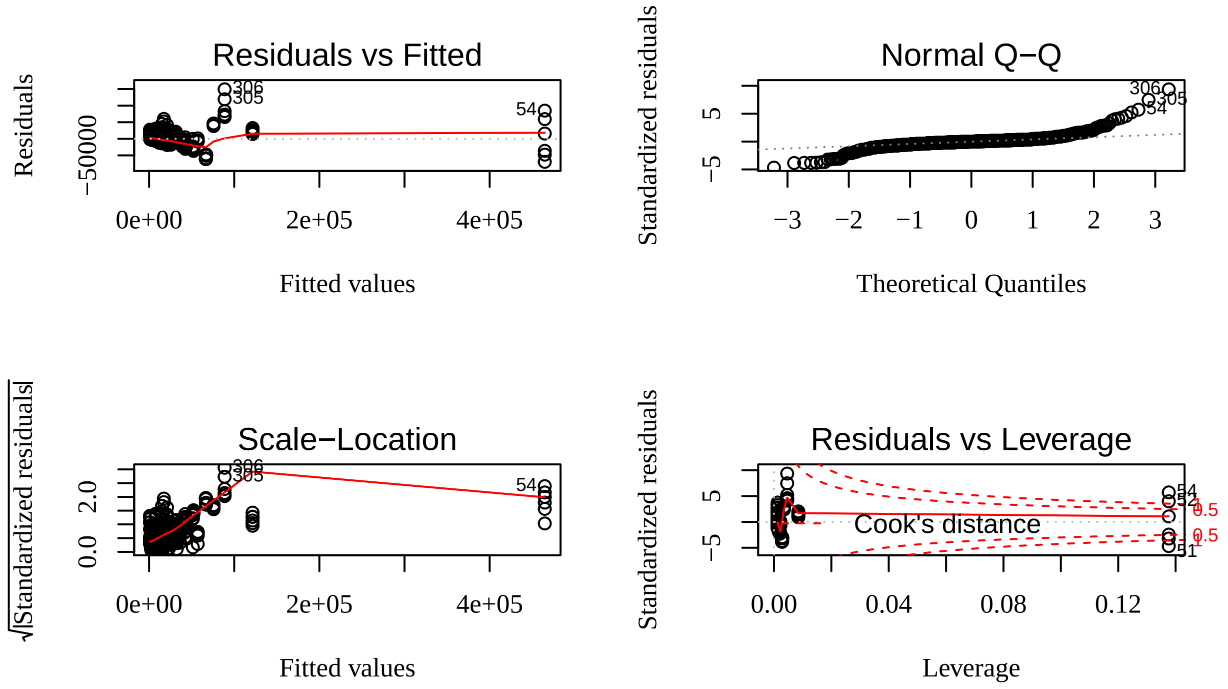


```r
fit5 <- lm(Energy_Cost_USD ~ Property_GFA_ftsq, na.action=na.exclude, data = New)
summary(fit5)
```

```
##
## Call:
## lm(formula = Energy_Cost_USD ~ Property_GFA_ftsq, data = New,
##     na.action = na.exclude)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -69376  -4745   -174   4009 149303
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -596.92340  627.50571  -0.951    0.342
## Property_GFA_ftsq      1.52716    0.02022  75.539   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15980 on 780 degrees of freedom
##   (52 observations deleted due to missingness)
## Multiple R-squared:  0.8797, Adjusted R-squared:  0.8796
## F-statistic:  5706 on 1 and 780 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(fit5,family="Times")
```

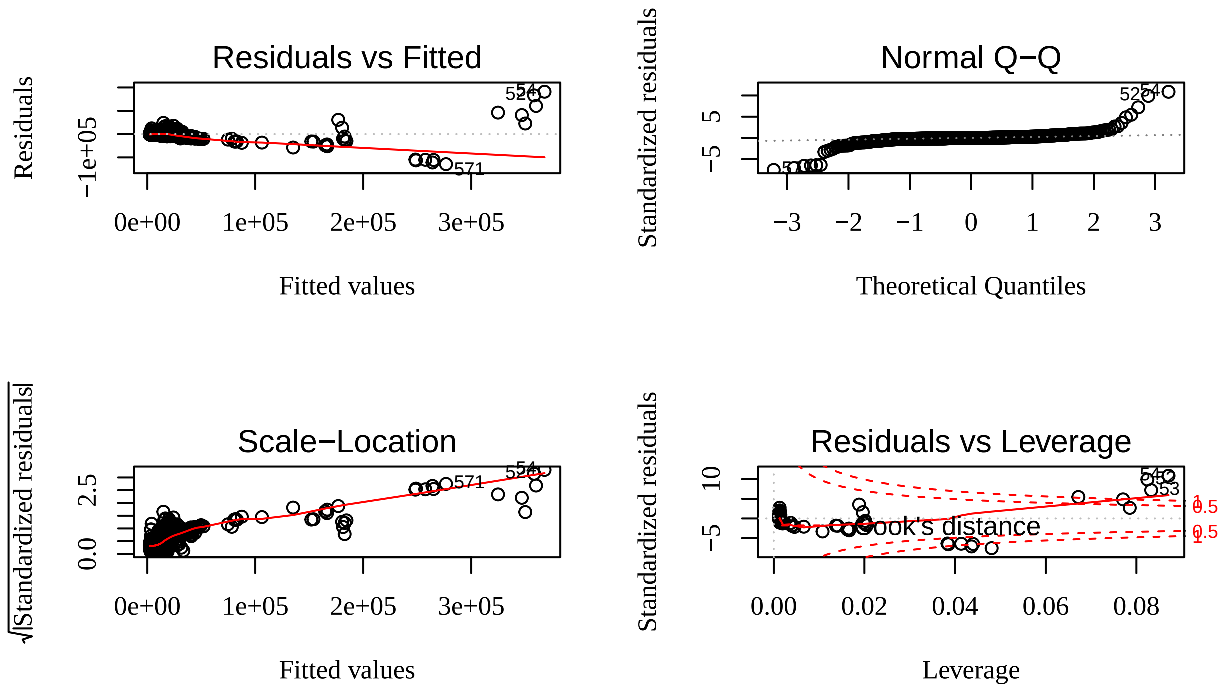


```
fit6 <- lm(Energy_Cost_USD ~ WN_Site_Electricity_kWh^2, na.action=na.exclude, data = New
summary(fit6)
```

```
##
## Call:
```

```
## lm(formula = Energy_Cost_USD ~ WN_Site_Electricity_kWh^2, data = New,
##     na.action = na.exclude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -128981   -2332    -264    2608  181788
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.241e+03  6.751e+02   3.319 0.000945 ***
## WN_Site_Electricity_kWh 1.432e-01  2.106e-03  68.010  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17510 on 775 degrees of freedom
##   (57 observations deleted due to missingness)
## Multiple R-squared:  0.8565, Adjusted R-squared:  0.8563
## F-statistic:  4625 on 1 and 775 DF,  p-value: < 2.2e-16
```

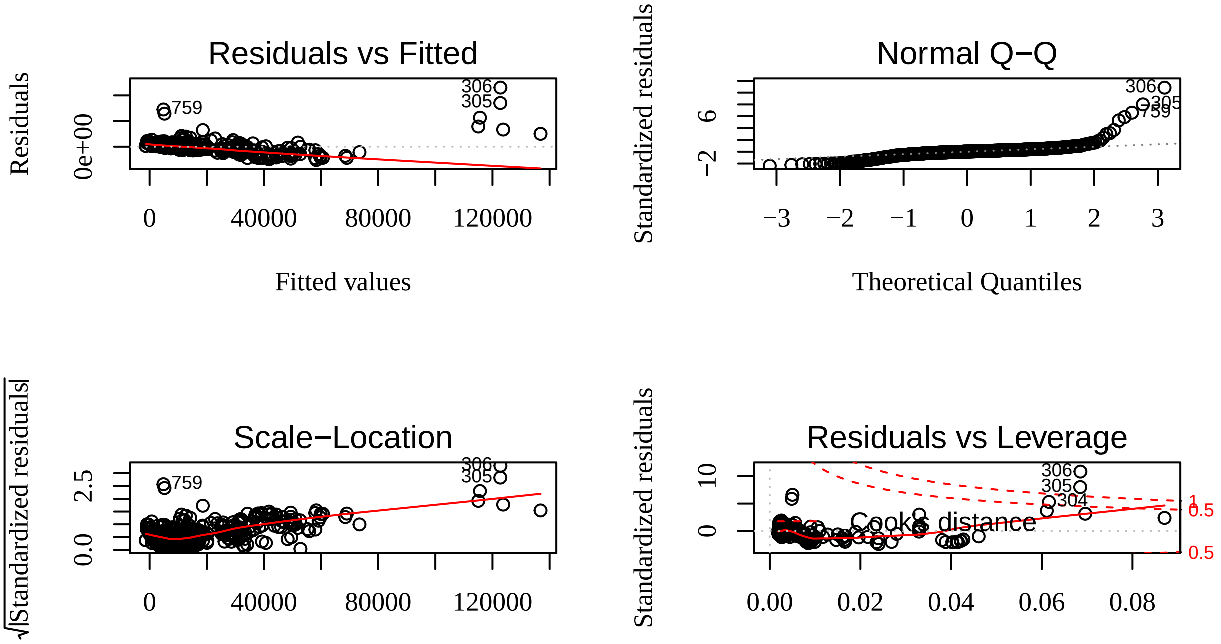
```
par(mfrow=c(2,2))
plot(fit6,family="Times")
```



```
fit8 <- lm(Energy_Cost_USD ~ WN_Site_NG_Use_therms+Property_GFA_ftsq, na.action=na.exclu
summary(fit8)
```

```
##
## Call:
## lm(formula = Energy_Cost_USD ~ WN_Site_NG_Use_therms + Property_GFA_ftsq,
##     data = New, na.action = na.exclude)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -26026  -3164    351   3056 115257
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -3.360e+03  7.487e+02  -4.487 8.88e-06 ***
## WN_Site_NG_Use_therms  1.714e+00  7.719e-02  22.210  < 2e-16 ***
## Property_GFA_ftsq      6.446e-01  6.494e-02   9.925  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11020 on 525 degrees of freedom
##   (306 observations deleted due to missingness)
## Multiple R-squared:  0.727,  Adjusted R-squared:  0.726
## F-statistic:   699 on 2 and 525 DF,  p-value: < 2.2e-16
```

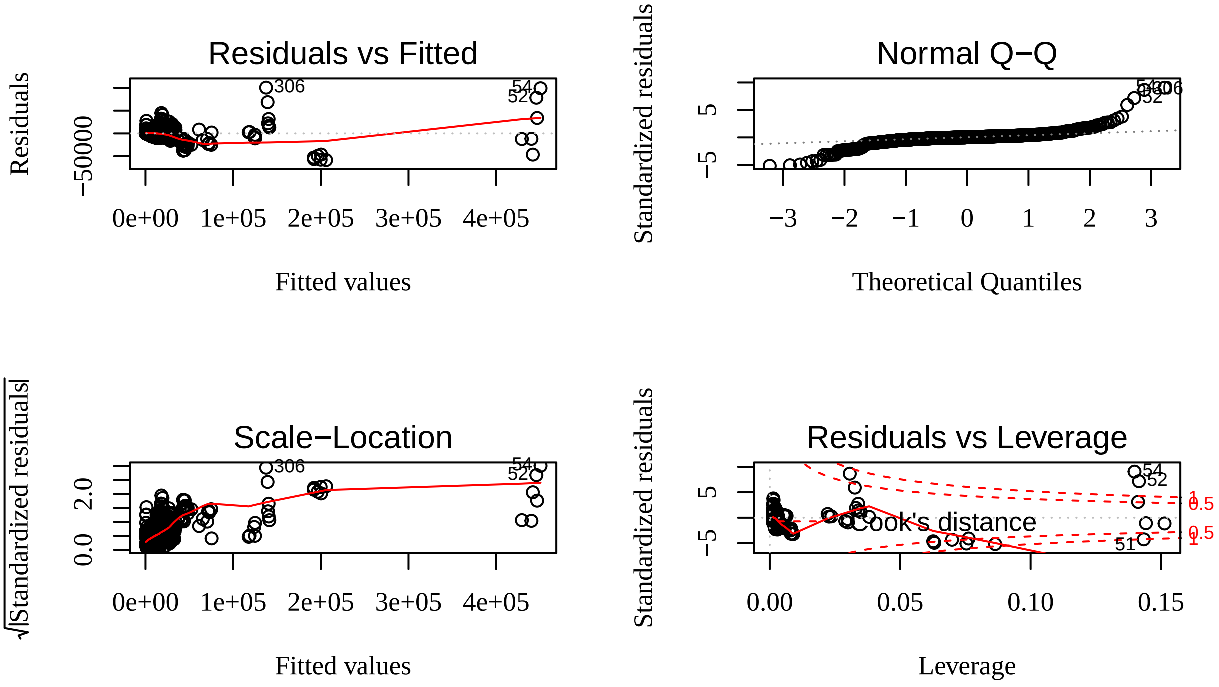
```
par(mfrow=c(2,2))
plot(fit8,family="Times")
```



```
fit9 <- lm(Energy_Cost_USD ~ WN_Site_Electricity_kWh^2+Property_GFA_ftsq, na.action=na.e
summary(fit9)
```

```
##
## Call:
## lm(formula = Energy_Cost_USD ~ WN_Site_Electricity_kWh^2 + Property_GFA_ftsq,
##     data = New, na.action = na.exclude)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -58406  -2594    298   3207  100447
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -6.741e+02  4.647e+02  -1.451    0.147
## WN_Site_Electricity_kWh  7.085e-02  2.761e-03  25.657   <2e-16 ***
## Property_GFA_ftsq        8.876e-01  2.904e-02  30.561   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11790 on 774 degrees of freedom
##   (57 observations deleted due to missingness)
## Multiple R-squared:  0.935,  Adjusted R-squared:  0.9348
## F-statistic:  5564 on 2 and 774 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(fit9,family="Times")
```


Residuals vs Fitted
Residuals
−50000
0e+00
1e+05
2e+05
3e+05
4e+05
Fitted values
306
54
52
Normal Q−Q
Standardized residuals
−5
5
−3
−2
−1
0
1
2
3
Theoretical Quantiles
54
306
52
Scale−Location
√|Standardized residuals|
0.0
2.0
0e+00
1e+05
2e+05
3e+05
4e+05
Fitted values
306
54
52
Residuals vs Leverage
Standardized residuals
−5
5
0.00
0.05
0.10
0.15
Leverage
Cook's distance
54
52
51
1
0.5
0.5
1