Analysis of Factors that Affect Gas and Electricity Usage in Chicago

Aamodini Gupta (gupta88) and Sahiti Kolli (kolli4)

12/13/2018

Introduction and Literature Review

Buildings represent 40% of energy consumption and 38% of CO2 emissions in the United States (Amasyali & El-Gohary, 2018). Understanding energy consumption can aid in implementing energy efficient planning and create infrastructure and environments that reduce inefficient energy usage. Energy saving practices can aid in the effort to reduce air pollution, especially in large urban areas. Additionally, understanding energy consumption can minimize economic losses since forecasting has become a tool for optimizing energy resources. Predicting and understanding energy consumption is dependent on a variety of diverse factors including properties of the building, weather, and behavior of the occupants. Thus, it is challenging to accurately determine the specific factors of variations in energy usage.

This analysis aims to understand the factors that predict energy expenditure by comparing the results obtained from linear regression and basis spline models. The analysis also aims to understand if building types can be accurately classified using various factors including energy expenditure through the use of random forests. Through this analysis, the aim is to have a better understanding of what specific factors of buildings and occupants affect energy expenditure as well as how various factors differ across building types.

The data comes from open source data from the city of Chicago found here. It displays units of energy consumption for households, businesses, and industries in the City of Chicago during 2010 measured by electricity (kWh) and gas (Therm) expenditure. Each observation is a building, and for each observation, total and average energy expenditure for 2010 is provided along with the average expenditure for each month in the year. The data is aggregated from ComEd and Peoples Natural Gas by Accenture. Electrical and gas usage data comprises 88% of Chicago's buildings in 2010. The electricity data comprises 68% of overall electrical usage in the city while gas data comprises 81% of all gas consumption in Chicago for 2010. The analysis will compare both electricity and gas usage. The data also contains variables such as Census block population, building characteristics, and occupancy for each observation. The data has 67.1K observations and 73 variables which are both categorical and numeric in nature with a size of 10 MB.

Although there are no publicly available analyses on this specific data, there have been numerous studies to understand energy consumption using analytical and predictive methods. Previous research has successfully predicted future expenditure based on previous years' consumption through artificial neural networks and modified Newton's method. However, they do not assess factors affecting these expenditures (Ozoh et al., 2014). Research that does look into the factors affecting energy consumption use artificial neural networks to find that price and temperature are significant when predicting expenditure (Ozoh et al., 2016). Analysis comparing the performance between random forests and neural networks finds that the latter performed better (Tso and Yao, 2007). Improved predictive performance from using a random forest model compared to a neural network or other machine learning techniques has guided this analysis to include this type of method.

Analysis has also been done to predict both short term and long term energy expenditure. Short term expenditure measures hours and days and is beneficial in understanding generator capacity and short term maintenance. Long term expenditure measures yearly or longer data and is beneficial in developing more permanent generation strategies (Rahman et al., 2018). A recent analysis of existing models in predicting

energy finds that very few models focus on long term analysis and only 19% of reviewed studies focus on residential housing (Amasyali & El-Gohary, 2018). Since the implication of this analysis should help guide and understand more long term energy efficient practices, it is important to conduct analysis on data that covers an entire year of energy expenditure and includes information on residential housing.

There has been considerable research into this topic and different machine learning algorithms have been used to analyze energy related data. However, through this analysis, more potential factors such as building type, age of the building, and size of the unit are included. These factors are essential in affecting the efficiency and expenditure of energy usage and have been lacking in previous analyses. Doing so creates a more comprehensive understanding of energy usage and consolidates the different analyses that have previously been done. Additionally, by analyzing both kWh and Therms, there is an assessment if certain factors are more important for one energy type over the other, thus providing more insight into understanding expenditure.

Data Exploration

Exploratoration of data is done to better understand the data and its variables prior to conducting analysis. The first step is to remove the observations in the data with any missing values. Since the data has an immense number of observations, doing so reduces it by 31.5% from 67,051 to 45,854 observations. This is still a large number of observations to conduct the analysis on.

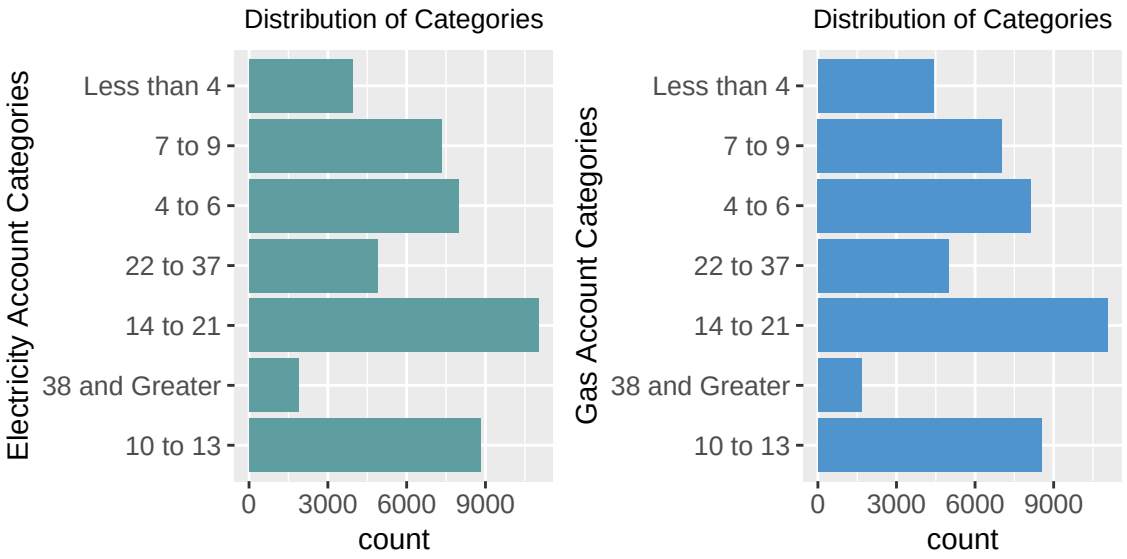
Every observation in the data is a building with 68 numeric and 5 categorical features. The categorical variables are Building Type, Building Subtype, Electricity Accounts, Gas Accounts, and Community Area.

The categories and number of observations in each category for Building Type and Building Subtype are shown in the table below. Building Subtypes are sub-categories of Building Types. For Building Type, there are only 2 observations for Industrial buildings. These observations are removed to prevent any model from attempting to classify a category with too few observations and to instead focus on categories with a larger sample size. The Building Subtypes show a similar result for Municipal buildings (28 observations) and are also removed. The dominating category for Building Subtype is Single Family which accounts for 51.1% of the observations.

Commercial	Residential
5879	39975

Commercial	Multi < 7	Multi 7+	Single Family
3388	17640	1377	23449

Electricity Accounts have 273 categories and Gas Accounts have 150 categories with an uneven distribution in each category. Each variable is consolidated into six categories displayed in the figure below.

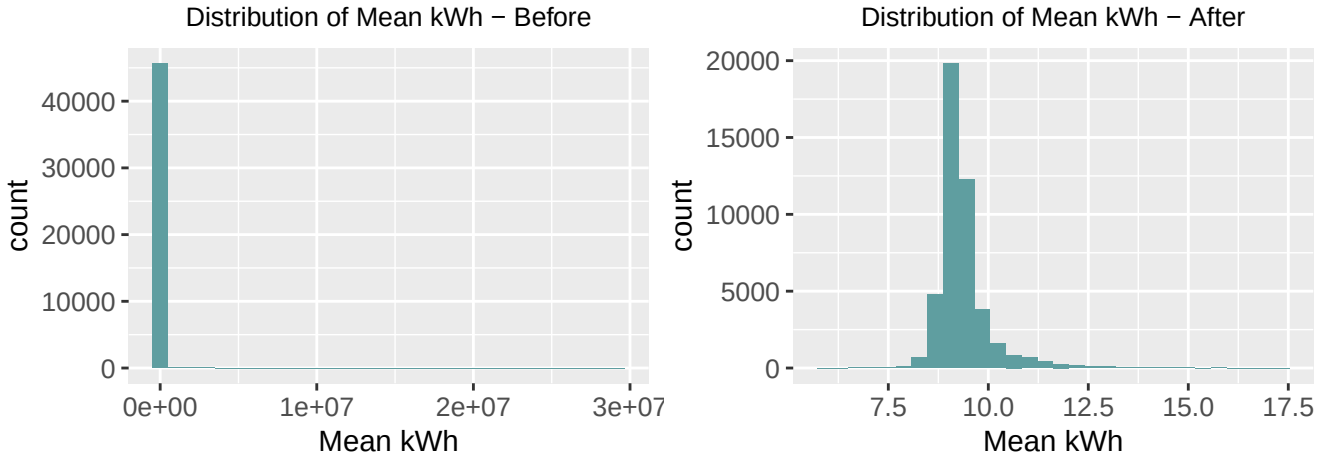


Community Area also has a large number of categories. The original categories have the names of 77 neighborhoods of Chicago which are now consolidated into 9 geographic regions based on this information.

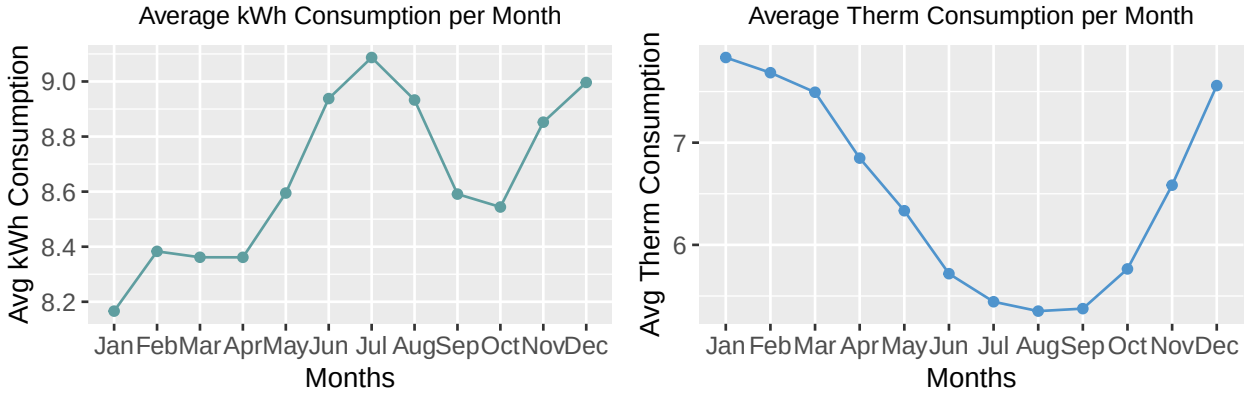
There are a large number of numeric variables in the data. Several split the energy expenditures into quartiles, standard deviations, total, minimum and maximum values for each observation. Since the analysis is focused on understanding and predicting average energy usage over the year, these variables are removed. There are also 12 additional columns, 1 for each month, that give the average energy usage for every observation per month.

The average values of kWh usage have a very skewed distribution, as seen in the left plot below. This indicates that the range of values for the electric energy usage is very high. The log scale is taken to account for this. The right plot below shows the distribution of kWh expenditure after scaling, where the range is

more reasonable. A similar pattern is seen in the average values of Therms, and the same scaling is applied.



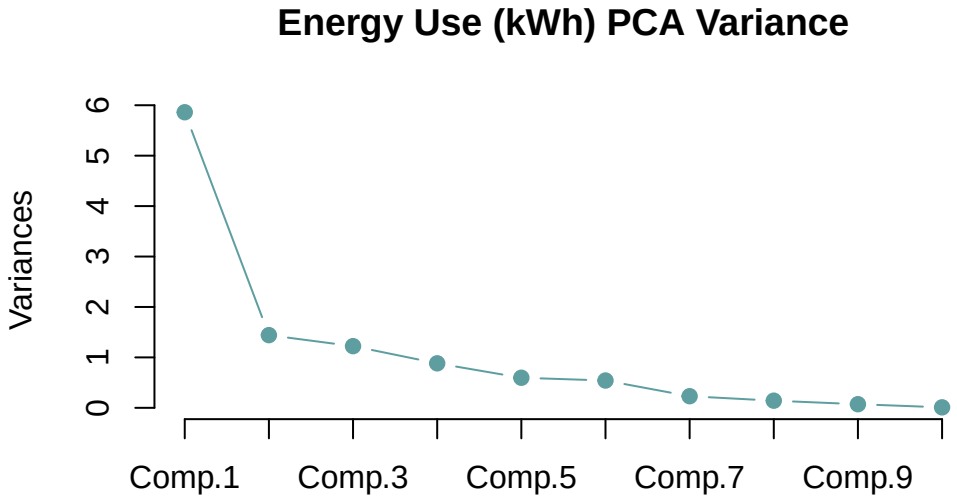
The figures below show the average kWh and Therm consumption by month. There is a sharp increase in kWh consumption and decrease in Therm consumption during the months of June through August. Likewise, less kWh consumption and higher Therm consumption is observed in the colder months. This indicates that month and time of the year differ in terms of energy usage, so it is important to take this into account during the analysis.



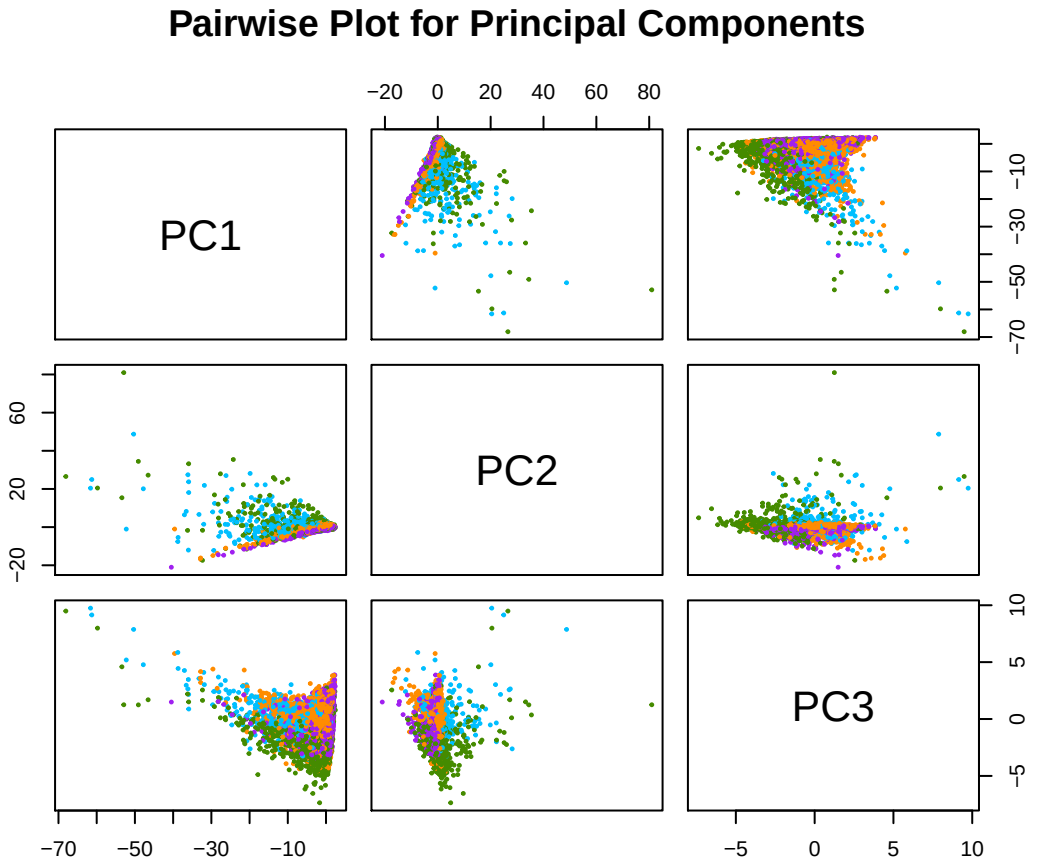
Previous research has found temperature to be significant in predicting energy expenditure (Ozoh et al., 2016), thus providing further evidence to include the effects of month into the analysis. In addition to splitting the data into separate data sets for kWh and Therms, data sets are also created to account for months. Each of these data sets have two additional columns indicating the month and the month's corresponding average energy for each observation. This is done by recording each observation twelve times with each energy value per month in one column and the category of the month in the next column. Creating and manipulating the original data set to obtain this format is essential for the analysis so that the effect of the months can be taken into account.

To understand how the numeric variables contribute to the variation in the data, principal components analysis (PCA) is conducted.

PCA aids in understanding which variables contribute to the most variation in the data. After centering the variables and doing PCA for the kWh data, the variance explained by each of the components is seen in the figure below. The first component explains the largest amount of variance in the data and the first 8 components explain 81% of the cumulative variance.



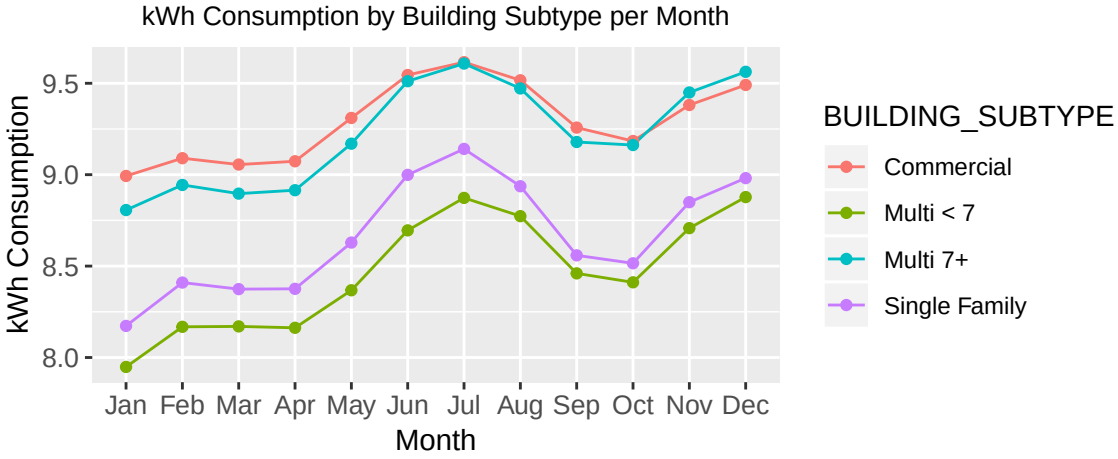
The two dimensional pairwise plots of the first three components show the greatest separation by Building Subtype, seen through the color separation. This provides evidence that there may be factors that vary across building subtypes, which is explored in the analysis.

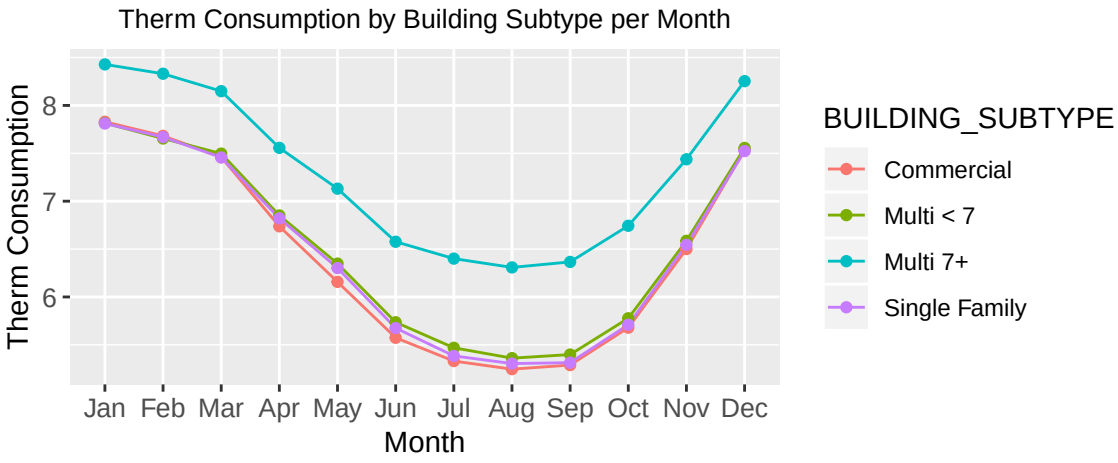


The loadings of the first three components are shown below. Analyzing the loadings provides feature contributions of the components. The first component presents the largest absolute values of loadings in the variables describing the occupied units and total population indicating that this component is largely describing the building occupancy. The second component has the largest loading values in average square foot and average stories and is thus describing the building size. Lastly, the third component is describing building age. This is repeated for the data containing Therms and similar results are seen.

	PC1	PC2	PC3
ZERO.KWH.ACCOUNTS	-0.2666980	0.2658322	0.1419777
KWH.MEAN.2010	-0.2113631	0.2867015	-0.4153159
KWH.SQFT.MEAN.2010	-0.2220238	0.5728332	-0.0161422
TOTAL.POPULATION	-0.3674730	-0.2369151	0.1366228
TOTAL.UNITS	-0.3988684	-0.1734555	0.0361936
AVERAGE.STORIES	-0.2587175	0.5435467	0.1428996
AVERAGE.BUILDING.AGE	0.0372729	0.1033180	0.7331811
AVERAGE.HOUSESIZE	0.1314431	-0.0035037	0.4735677
OCCUPIED.UNITS	-0.3971785	-0.2007545	0.0231882
RENTER.OCCUPIED.HOUSING.UNITS	-0.3737219	-0.2145864	0.0616615
OCCUPIED.HOUSING.UNITS	-0.3971785	-0.2007545	0.0231882

The two plots below show average kWh and Therm consumption by Building Subtype. For kWh, the Commercial buildings have the highest expenditure, and households with less than 7 occupants have the lowest expenditure. For Therms, households with over 7 occupants have the highest expenditure and single family homes have the lowest expenditure. It should be noted that the energy expenditure varies across building subtype. There is evidence that energy expenditure varies across building subtype, and this provides justification for using building subtype to understand energy usage in the analysis.





The exploration of the data has helped identify variables that needed further processing and cleaning. It also aided in visualizing underlying trends in the data that propagated the analysis.

Analysis

To determine the factors that affect both average kWh and Therm expenditure per year, two different models are analyzed and compared.

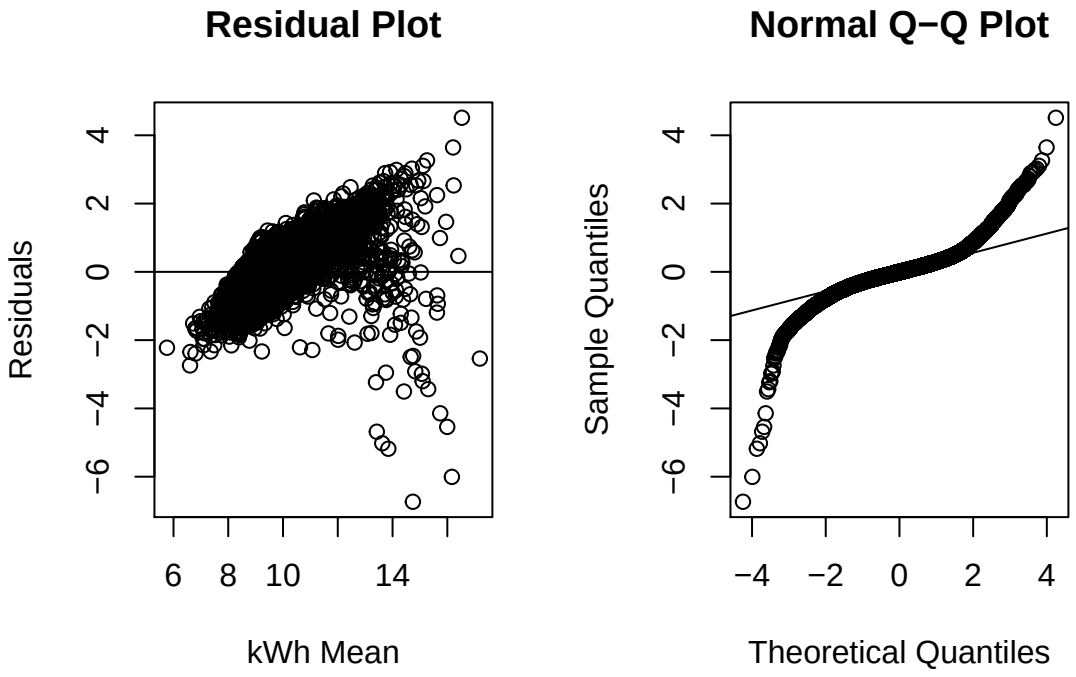
Linear Regression with Principal Components

Analysis of kWh

The first model is a linear regression with the first three principal components and the four categorical variables as predictors that predict the average kWh. Principal components are used as predictors to reduce the complexity of the model, which effectively decreases the variance but also increases the bias of the model.

```
nums <- unlist(lapply(kwh.dat, is.numeric))
pca.linearfit.kwh <- data.frame("PC1" = prcomp.kwh$x[,1],"PC2" = prcomp.kwh$x[,2],
                                "PC3" = prcomp.kwh$x[,3], kwh.dat[, (nums == F)],
                                "KWH.MEAN.2010" = kwh.dat$KWH.MEAN.2010)
pca.fit.kwh <- lm(KWH.MEAN.2010 ~., dat = pca.linearfit.kwh)
```

To assess whether a linear model is appropriate for the data, the residual plots and Q-Q plots are observed and shown below. For linearity to be assumed, residuals should be homoscedastic. Based on the figure, there appears to be a slight pattern in the residuals. However, the points are scattered almost evenly around 0 and linearity can be assumed. The Q-Q plot will assess if the normality assumption is met. This plot shows that there is only deviation towards the tails and the rest of the pattern fits closely to a straight line, so the normality assumption is held.



The first 4 coefficients of the variables and the R^2 values are shown below. Assessing the fit of this model shows an R^2 value of .6658. This is considerably high and implies that 66.58% of variation in average kWh expenditure is explained by this model. The mean squared error is .156, which is also quite low and this

provides evidence for good model fit. All of the variables are significant except certain categories of electricity accounts. This shows that building occupancy (PC1), building size (PC2), building age (PC3), building type and subtype, community area, and electricity accounts to an extent are significant in their relation to average electricity usage.

```
summary.kwh$coefficients[1:5,]
```

```
##                              Estimate  Std. Error    t value    Pr(>|t|)
## (Intercept)                10.3297726 0.023306607  443.21220 0.00000e+00
## PC1                        -0.1128193 0.001024279 -110.14509 0.00000e+00
## PC2                         0.1676188 0.001608508  104.20761 0.00000e+00
## PC3                        -0.2127421 0.001994288 -106.67573 0.00000e+00
## BUILDING.TYPEResidential -0.1840984 0.009159210  -20.09981 1.80292e-89
```

```
summary.kwh$r.squared
```

```
## [1] 0.6657651
```

```
mean(pca.fit.kwh$residuals^2)
```

```
## [1] 0.1561605
```

Analyzing the coefficients shows some novel insights into the relationships between these variables and electricity. The second principal component has a positive coefficient, indicating that for an increase in building size, there is an increase in average kWh usage. This is not particularly surprising because larger buildings are expected to use more energy than smaller ones. Principal components 1 and 3 have negative coefficients indicating that an increase in building occupancy and building age result in a decrease in average kWh expenditure. This is surprising because older building and buildings with more occupants are assumed to use more electricity. Upon further research, it is understood that efficiency of appliances and electronics is more indicative of energy efficiency than building age. Therefore, it is possible for older buildings and buildings with larger occupancy to be more energy efficient and use less energy than newer ones. For all community areas, coefficients are negative with the most negative being the Far Southeast Side and the least negative being the North Side. This indicates that buildings in the Far Southeast Side may use, on average, less electricity amongst all community areas.

Analysis of Therms

This regression is repeated with Therms as the predictor. The Residual plot and Q-Q plot show a similar result from the analysis of kWh. Thus, although a slight trend is observed, it can be said that the linearity and normality assumptions are upheld.

The first 4 coefficients of the variables and the R^2 values are shown below. The R^2 value is .623. This is also a considerably large value and indicates that 62.33% of variation in average Therm expenditure is explained by this model. The mean squared error for this model is .123. This is even lower than the previous model indicating that there is a good model fit. All the variables are significant except for certain community area and gas accounts categories. The North Side and Southeast Side communities are not significant in modeling average Therm usage, but they are significant in modeling average kWh usage. Thus, all the same factors that are significant in analyzing kWh are also significant in analyzing Therms, but community areas differ in how they affect gas versus electricity usage.

```
summary.therms$coefficients[1:5,]
```

```
##                              Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)                7.88645285 0.0202959603  388.57254 0.000000e+00
## PC1                       -0.11065983 0.0008729203 -126.76968 0.000000e+00
```

```
## PC2                       -0.15316332 0.0014765688 -103.72921 0.000000e+00
## PC3                        0.07032231 0.0017625171   39.89879 0.000000e+00
## BUILDING.TYPEResidential -0.14032067 0.0081011322  -17.32112 5.331966e-67
```

```
summary.therms$r.squared
```

```
## [1] 0.623258
```

```
mean(pca.fit.therms$residuals^2)
```

```
## [1] 0.123067
```

Principal components 1 and 2 have negative coefficients and component 3 has a positive coefficient. Thus building size (PC2) and building age (PC3) have the opposite effect on gas usage than they do on electricity usage. Additionally, the coefficients for community area are all positive whereas they are negative for the kWh model.

These two linear models identified key features which are significant with average energy expenditures. These same features are significant with both kWh and Therms, and model fit measures provide strong evidence that the predictors explain the energy expenditure well. Certain variations do exist between how the variables relate to the different energy expenditures. Community areas significant in the kWh model are not significant in the Therm model and the sign of certain coefficients differ across the two models, indicating a reverse relationship. There is a better understanding of which variables contribute to understanding different energy expenditures and how they affect these expenditures. It can also be understood that the way in which these variables are related to energy expenditure depends on the type of energy being assessed.

Cubic Spline

To take into account the effects of months on energy expenditure, a non parametric model is analyzed. For this analysis, the constructed data described in the data exploration that shows the average values for each month for each observation is used. The months were changed to a numeric value between 1-12 and a cubic spline is done on months to predict the monthly energy usage for each building. The cubic spline creates non linear functions of months as new features. A linear regression is fit on these new features as well as the other numeric and categorical variables in the data set. This should result in more flexibility of the model.

There are 3 knots initially chosen, which results in a basis for four regions meant to correspond to each season. Increasing the number of knots will increase the variance but decrease the bias of the model. Interaction terms are included to understand how the spline on months interacts with the other covariates in the data.

Analysis of kWh

The R^2 value that results from this initial model is .58 with a mean squared error of .47 when predicting monthly average energy usage.

```
bs.kwh <- lm(KWH.USE ~ .*bs(months, df=3, knots = 3), data=kwh.ms)
```

```
summary.splines.kwh$coefficients[1:4,]
```

```
##                             Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)               10.1018081 0.04797037 210.584334 0.000000e+00
## BUILDING.TYPEResidential  -0.1375983 0.01582331  -8.695925 3.449355e-18
```

```
## BUILDING_SUBTYPEMulti < 7 -1.4250881 0.01969985 -72.340032 0.000000e+00
## BUILDING_SUBTYPEMulti 7+  -1.6084233 0.02612541 -61.565485 0.000000e+00
```

```
summary.splines.kwh$r.squared
```

```
## [1] 0.5839179
```

```
mean(bs.kwh$residuals^2)
```

```
## [1] 0.4736285
```

However, the ultimate goal is to predict the yearly average energy while including the effects of months. Therefore, the average of 12 predicted values per observation is calculated to get the fitted values of the average energy use per year for each observation from the model.

```
fitted.kwh <- predict(bs.kwh)

kwh.mean.hat <- data.frame("KWH.MEAN.2010" = kwh.dat$KWH.MEAN.2010, "FITTED.MEAN" = 0)
for (obs in 1:nrow(therms.dat)){
  kwh.mean.hat[obs,2] <- sum(sapply(seq(0,11),
                                    function(x) fitted.kwh[(obs + x*nrow(kwh.dat))]))/12
}
```

The mean squared error for the average yearly usage is 1.18.

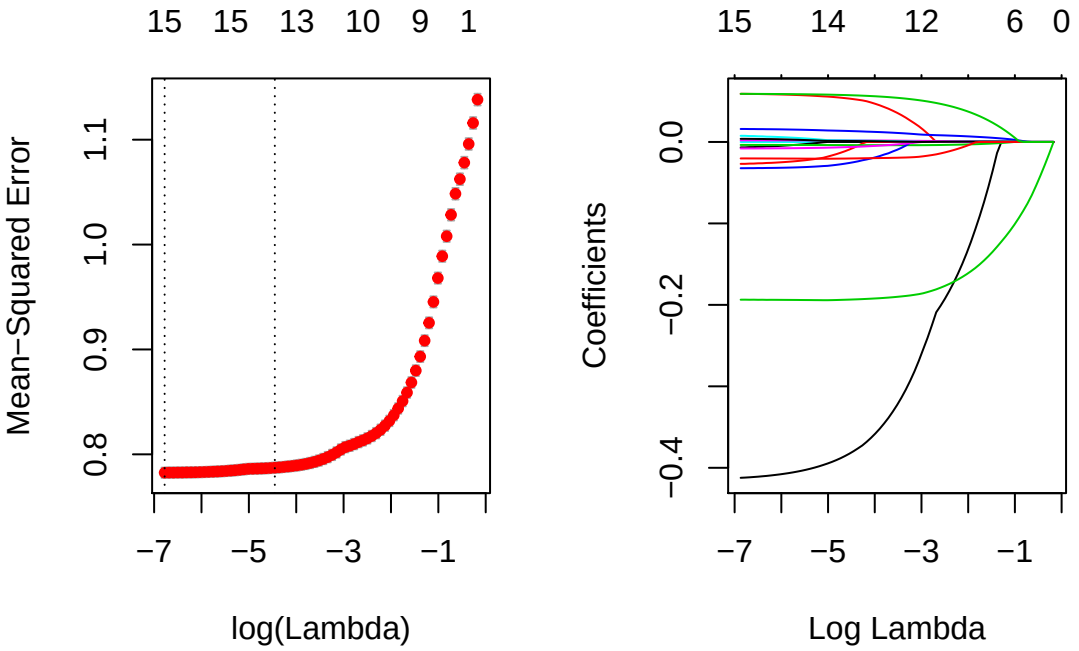
```
mean((kwh.mean.hat$KWH.MEAN.2010 - kwh.mean.hat$FITTED.MEAN)^2)
```

```
## [1] 1.180769
```

To improve this model, any collinearity in the variables in the model are penalized by Elastic Net regularization. An Elastic Net balances Ridge and Lasso regularization methods. Ridge uses an l_2 penalty which shrinks the coefficients in a model and is beneficial for addressing collinearity. Lasso regression also shrinks the coefficients in a model but uses an l_1 penalization. Lasso will also shrink some coefficients to 0, effectively doing variable selection in addition to shrinkage. A limitation of Lasso is that it tends to only pick one predictor when predictors are highly correlated. Additionally, the predictive performance of Lasso is less than Ridge when the number of observations are greater than the number of predictors. Using Elastic Net overcomes these challenges. An elastic net will use both an l_1 and l_2 penalization to penalize the coefficients. An elastic net thus serves as both a variable selection method and shrink the coefficients. The amount of shrinkage is chosen through tuning the lambda parameter. As the value of lambda increases, the coefficients are shrunk more, thus decreasing the variance but increasing the bias of the model.

```
set.seed(3)
elasticnet_kwh <- cv.glmnet(data.matrix(kwh.ms[,-which(colnames(kwh.ms) == "KWH.USE")]),
                            kwh.ms$KWH.USE, nfolds = 10, alpha = .5)
```

The left plot below shows the cross validation errors for increasing values of lambda. The right plot below shows the changes in the coefficients for increasing lambda values.



Tuning this lambda parameter results in a removal of the Total Units variable.

	1
(Intercept)	9.8416823
BUILDING.TYPE	-0.3795818
BUILDING_SUBTYPE	0.0525481
ELECTRICITY.ACCOUNTS	-0.1933444
ZERO.KWH.ACCOUNTS	0.0132740
KWH.SQFT.MEAN.2010	0.0000027
TOTAL.POPULATION	0.0019054
TOTAL.UNITS	0.0000000
AVERAGE.STORIES	-0.0070940
AVERAGE.BUILDING.AGE	-0.0038983
AVERAGE.HOUSESIZE	-0.0248386
OCCUPIED.UNITS	0.0018256
RENTER.OCCUPIED.HOUSING.UNITS	-0.0060500
OCCUPIED.HOUSING.UNITS	0.0011472
Community.Area	-0.0202285
months	0.0572478

The best lambda value minimizes the cross validation error, and this value is .001.

```
min(elasticnet_kwh$lambda)
```

```
## [1] 0.001138571
```

The model is again fit by removing this variable and by using 5 knots, which results in a basis being created for six regions. This will allow for the consideration of the effect of 2 months at a time. After increasing the number of knots and taking into consideration the elastic net penalization, the final model results in an R^2 value of .59 with a mean squared error of .47 in predicting the average energy use per month, and a mean squared error of 1.18 in predicting the average energy use per year.

```
bs.kwh <- lm(KWH.USE~(.- TOTAL.UNITS)*bs(months, df=3, knots = 5), data=kwh.ms)
```

```
summary.splines.kwh$coefficients[1:3,]
```

```
##                               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)                 10.0275876 0.04605396 217.735629 0.000000e+00
## BUILDING.TYPEResidential    -0.1281907 0.01527166  -8.394026 4.708626e-17
## BUILDING_SUBTYPEMulti < 7 -1.4106230 0.01901306 -74.192316 0.000000e+00
```

```
summary.splines.kwh$r.squared
```

```
## [1] 0.5853638
```

```
mean(bs.kwh$residuals^2)
```

```
## [1] 0.4719826
```

```
mean((kwh.mean.hat$KWH.MEAN.2010 - kwh.mean.hat$FITTED.MEAN)^2)
```

```
## [1] 1.181405
```

Analysis of Therms

The procedure is repeated for Therms. The initial model results in an R^2 value of .81 with a mean squared error of .30 in predicting the average energy use per month, and a mean squared error of 1.73 in predicting the average energy use per year.

```
bs.therms <- lm(THERM.USE ~ .*bs(months, df=3, knots = 3), data=therms.ms)
```

```
summary.splines.therms$r.squared
```

```
## [1] 0.811555
```

```
mean(bs.therms$residuals^2)
```

```
## [1] 0.303295
```

```
mean((therms.mean.hat$THERM.MEAN.2010 - therms.mean.hat$FITTED.MEAN)^2)
```

```
## [1] 1.72532
```

The elastic net resulted in the removal of more variables including Occupied Units, Occupied Housing Units, Building Subtype, Total Units, Average House Size, and Renter Occupied Housing Units as seen in the table below. Removing these variables greatly reduces the model complexity and variance in the model but also increases the bias.

The lambda from the elastic net is also tuned to .001.

```
min(elasticnet_therms$lambda)
```

```
## [1] 0.001059574
```

This final model results in an R^2 value of .80 with a mean squared error of .32 in predicting the average energy use per month, and a mean squared error of 1.70 in predicting the average energy use per year.

	1
(Intercept)	7.8389679
BUILDING.TYPE	0.0127306
BUILDING_SUBTYPE	0.0000000
GAS.ACCOUNTS	-0.1848332
THERMS.SQFT.MEAN.2010	0.0000016
TOTAL.POPULATION	0.0022902
TOTAL.UNITS	0.0000000
AVERAGE.STORIES	0.0728322
AVERAGE.BUILDING.AGE	-0.0015888
AVERAGE.HOUSESIZE	0.0000000
OCCUPIED.UNITS	0.0000000
RENTER.OCCUPIED.HOUSING.UNITS	0.0000000
OCCUPIED.HOUSING.UNITS	0.0000000
Community.Area	-0.0140553
months	-0.1187017

```
bs.therms <- lm(THERM.USE~(.-OCCUPIED.UNITS - OCCUPIED.HOUSING.UNITS
                - BUILDING_SUBTYPE - TOTAL.UNITS - AVERAGE.HOUSESIZE
                - RENTER.OCCUPIED.HOUSING.UNITS)*bs(months, df=3, knots = 5),
                data=therms.ms)
```

```
summary.splines.therms$r.squared
```

```
## [1] 0.802375
```

```
mean(bs.therms$residuals^2)
```

```
## [1] 0.3180699
```

```
mean((therms.mean.hat$THERM.MEAN.2010 - therms.mean.hat$FITTED.MEAN)^2)
```

```
## [1] 1.69802
```

In both cases, the model does not improve much from the cubic spline model prior to regularization and is also comparatively worse than the linear regression using the principal components. Predicting energy usage while including the effects of months was expected to create a more accurate result of understanding energy usage at the higher level of a year. Therefore, it is surprising that the value of the mean squared error increased. One of the reasons for this increase may be due to a loss of information when calculating the average usage based on the months. Although this mean squared error is comparatively worse than the original linear model, almost all the variables are significant. This indicates that in addition to the numeric and categorical data, the effect of months is also related to average energy expenditure. Therefore, seasonality can be said to have an effect on average energy expenditure which supports the exploratory data analysis and prior literature.

Random Forests

A secondary goal of this analysis is to understand what factors are crucial when predicting building subtype. This is done through a random forest classification.

A major limitation of decision trees is that they are prone to overfitting and they are not robust, so a small change in the training data results in a different tree. Random forests generate many decision

trees and aggregate their predictions into one single prediction to overcome these limitations. Additionally, random forest models tend to have better predictive performance than a single decision tree. Prior research attempting to understand and classify energy expenditure also determined that random forests perform better than other machine learning methods such as neural networks (Tso and Yao, 2007). For these reasons, a random forest model is used for classification of building subtype.

Bootstrap aggregating, also known as bagging, will decorrelate the trees. Bagging will generate new training sets by sampling the original training set with replacement. Each bootstrapped data set will construct each tree and this decreases the variance without increasing the bias. At each splitting rule in a random forest model, a random subset of variables are considered. Additionally, the distance used in random forest is adaptive to the true underlying structure of the model.

Tuning is crucial when doing random forests and varying tuning parameters affect the bias and variance of the model. The terminal node size, mtry (the number of variables randomly chosen at each split from the full set of features), and the number of trees are tuned. The terminal node sizes being evaluated for tuning are (1,20,40,60), mtry are (1,3,6), and number of trees are (200, 500,1000). Tuning is done through a bootstrapped cross validation where the In-Bag samples are the training samples and the Out-of-Bag samples are testing samples. Combinations of the values are assessed and tuned by choosing the values that minimize the Out-of-Bag error estimate.

Analysis of kWh

Random forest is done on all the numeric variables mentioned earlier including average kWh as the features. Categorical variables are excluded since including categorical variables of different levels results in the random forests being biased towards the attributes with more levels. Prior to tuning, an Out-of-Bag error of 14.18% was achieved.

Tuning of the parameters is done using the code displayed below:

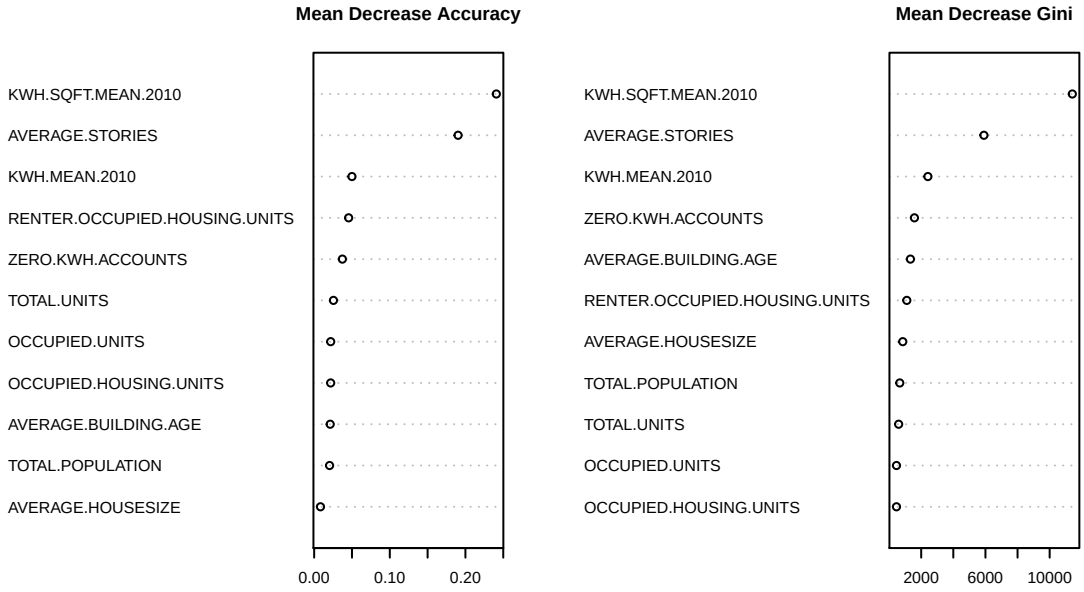
```
nodes <- c(1, 20, 40, 60) # picking some node values to test
mtrys <- c(1, 3, 6) # keeping the center value as the default
ntrees <- c(200, 500, 1000)
best <- c(10,10,10, 10) # initating the best(error, mtry, nodesize)

for (ntree in ntrees){
  for (nodesize in nodes){
    for (mtry in mtrys){
      rf.fit <- randomForest(dat.numeric, kwh.dat$BUILDING_SUBTYPE,
                             ntree = ntree, mtry = mtry, nodesize = nodesize)
      OOB.err <- mean(predict(rf.fit) != kwh.dat$BUILDING_SUBTYPE)
      # retain the information of only those predictors
      # that have the lowest prediction error
      ifelse(OOB.err < best[1], best <- c(OOB.err, mtry, nodesize, ntree), best <- best)
      print(best)
    }
  }
}
```

After tuning the parameters, this error decreases to 9.7%. This is achieved at a node size of 1, tree size of 1000, and mtry of 6. The predictions resulting from the model are shown below. The predictions show a relatively low classification error, indicating that the random forest is able to classify building subtypes relatively well.

```
rf.kwh.tune <- randomForest(num.kwh, kwh.dat$BUILDING_SUBTYPE,
                            ntree = 1000, mtry = 6, nodesize = 1, importance = TRUE)
```

```
##                Commercial Multi < 7 Multi 7+ Single Family class.error
## Commercial           2244       374       59           711  0.33766234
## Multi < 7             330     16149      141          1020  0.08452381
## Multi 7+               58       261     1028            30  0.25344953
## Single Family         315      1140        0         21994  0.06204955
```

The dominating group for building subtypes is single family subtypes that accounted for 51.1% of observations as described in the data exploration. The random forest model has a better classification accuracy than 51.1%, indicating that using these features to classify building subtype did much better than if everything is classified into the dominating group. Additionally, the Mean Decrease in Accuracy and Mean Decrease in Gini are observed to determine the most important variables in the classifications. These values are shown in the plot below.



The mean decrease in accuracy assesses how much the accuracy of the random forest decreases due to the exclusion of each variable. The variables with a larger mean decrease in accuracy are therefore more important for classifying the Building Subtype.

The Gini impurity measure is essentially the probability of a new record being incorrectly classified at a given node based on the training data. Mean Decrease in Gini is effectively a measure of how important a variable is for estimating the value of the target variable across all of the trees that make up the forest. A higher Mean Decrease in Gini therefore indicates higher variable importance. For both Mean Decrease in Accuracy and Mean Decrease in Gini, the variables with the largest value is average kWh per square foot followed by average stories. Since these variables are most crucial in the classification, they can be interpreted as having the most building subtype class discriminatory information. Average kWh per square foot expenditure having the largest variable importance means that this variable is distinctly different across building types.

Analysis of Therms

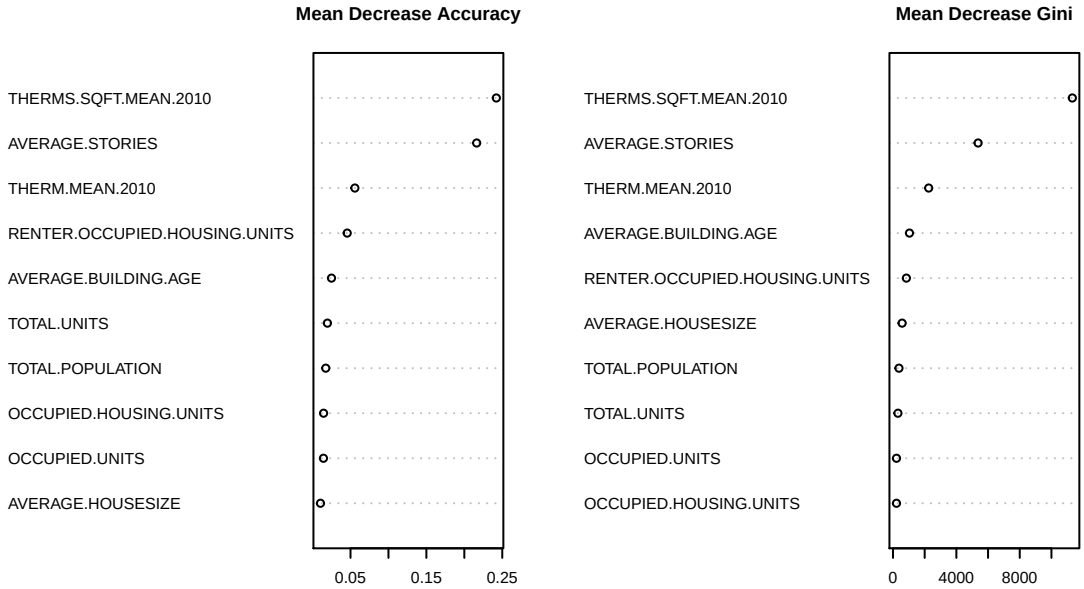
This random forest is done again but with Therms used in the features instead of kWh. Prior to tuning, an Out-of-Bag error of 15.21% was achieved.

After tuning the parameters, this error decreases to 9.7%. This is achieved at a node size of 20, tree size of 500, and mtry of 6. The classification errors are similar to the previous random forest using kWh and are still relatively low. The classification accuracy in this case is still better than 51.1%, indicating that using Therms in the features still classifies the building subtype better than if everything is classified into the dominating group of single family buildings.

```
rf.therms.tune <- randomForest(num.therms, therms.dat$BUILDING_SUBTYPE,
                               ntree = 500, mtry = 6, nodesize = 20, importance = TRUE)
```

```
##                Commercial Multi < 7 Multi 7+ Single Family class.error
## Commercial           2012       517       88           771  0.40613932
## Multi < 7             387     16068      143          1042  0.08911565
## Multi 7+               77       251     1021            28  0.25853304
## Single Family         274      1358        3         21814  0.06972579
```

For both Mean Decrease in Accuracy and Mean Decrease in Gini, the variables with the largest value are average Therms per square foot followed by average stories as seen below.



The results from this model show similar results as the previous one, indicating that in addition to average stories, average Therm usage per square foot provides most of the class discriminatory information for building subtype.

Both random forest models provide information that average energy usage per square foot provides the most information for classification of building subtype. The high classification accuracy of the models provide evidence that these variables not only provide the most information for classification, but also result in classifications with low errors. This effectively suggests that building types are distinct across energy usage.

Conclusion and Discussion

Two major goals were achieved through the various analyses. The first goal was to understand factors associated with two types of energy expenditure (electricity and gas). This was done through both parametric and non parametric models. The parametric linear regression identified features such as building occupancy size, building size, building age, building type and subtype, number of energy accounts, and community area to be significantly associated with energy usage. However, the level of significance and direction of the association varied between the two energy types. The non parametric regression aimed to assess the significance of months, and effectively weather, with energy expenditure. While this model performed worse than the parametric model, it shows that months have a significant association with energy usage, and thus usage varies during different times of the year.

The second goal of the analysis was to classify Building Subtypes and understand the factors that contribute most to this classification. This classification through a random forest model resulted in a low classification error and the most significant variable in classifying Building Subtype was found to be average energy expenditure per square foot. This indicates that energy usage per square foot provides discriminatory information on different building subtypes and provides further evidence that energy usage can be explained by building subtype and vice versa.

While these results provide insight into the factors that affect energy usage, there are limitations of the data and the analyses themselves.

From the linear regression predicting average kWh usage, an increase in building age showed a negative association with energy usage. This result is justified by understanding that the efficiency of appliances and electronics used in buildings is more indicative of energy efficiency and usage than the age of the building. Since this information is not available in this dataset, using data that provides information of electronic and appliances used in each building could provide more accurate information on energy efficiency and usage. Additional data limitations include having very few Industrial and Municipal building types and subtypes. This limitation prevented proper prediction of Municipal subtypes and Industrial building types in the analysis, which could have provided a potentially more accurate prediction. Furthermore, the data gives information only from 2010. More recent developments in housing and technology may also affect energy expenditure. Therefore, having more recent data would be beneficial in drawing an even stronger conclusion.

Although the linear model generally met the linearity and normal assumptions, there is still a slight trend observed in the residuals. A major limitation of linear models is that they can model linear trends. The spline model provided a non parametric approach to modeling the data, but the accuracy of the model did not improve compared to the linear model. Investigating further non parametric methods such as support vector machines could provide better accuracy and performance as well as represent the trend in the data better than the linear model. To improve the classification accuracy of the random forest model, more parameter values could be tuned to reach an even lower Out-of-Bag error. Additionally, doing partial permutations and unbiased trees so that categorical variables can be included could potentially lead to better accuracy.

By addressing these limitations in future research, further insight can be drawn about understanding factors affecting energy usage. By supplementing findings from this analysis with results from more accurate models or with data including more information, more concrete and specific information on energy usage can be understood.

Sources

- Ozoh, P., et al. (2014). A Comparative Analysis of Techniques for Forecasting Electricity Consumption. International Journal of Computer Applications 88(15), pp. 8-12.
- Ozoh, P., et al. (2016). A Predictive Framework for Electricity Consumption. Journal of IT in Asia 6(1).
- Amasyali, K. & El-Gohary, M.N. (2018). A review of data-driven building energy consumption prediction studies. Renewable and Sustainable Energy Reviews 81(1), pp.1192-1205.
- Tso, K.G. & Yao, K.K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy 32(9), pp.1761-1768.
- Rahman, A., et al. (2018). Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. Applied Energy 212, pp.372-385.