Modelling Building Energy Efficiency in New York
07/23/2017

1. Introduction

The following report will evaluate data from the IBM Data Science Experience: *"Modeling energy usage in New York City."* The purpose of the analysis will be to classify buildings based on the efficient consumption of energy and to create a model to identify inefficient energy buildings. Inefficient buildings identified may be able to take corrective actions which translate to annual energy savings.

2. Data

Data comes from two sources in csv files: BlocPower and public heating and cooling data for 103 buildings in New York City for 2016. BlocPower is a New York company that analyzes building data and connects investors to green building projects to save energy. The following tables describe the data coming from csv files. It is important to note that these buildings are not drawn from a random sample, but will be treated as such for the purpose of this analysis.

Source	CSV	Fields
BlocPower	`BlocPower_T.csv`	`UTSUM_Electricity_Usage, INFO_Year of Construction, INFO_Number of Stories, INFO_Total Square Feet, PLEI_1_Quantity, PLEI_3_Quantity`
BlocPower	`clusterEnergyLocation.csv`	`AddressID, property_name, Adress, Zipcode, Long, Lat, Annual Energy Bill (USD)`
Public Heating & Cooling	`CDD-HDD-Features.csv`	`Property Name, plug_load_consumption, ac_consumption, domestic_gas, heating_gas`

BlocPower_T.csv	
Column Name	Sample Data
UTSUM_Electricity_Usage	117,870 kWh
INFO_Year of Construction	1955
INFO_Number of Stories	4
INFO_Total Square Feet	14,600
PLEI_1_Quantity	1.0
PLEI_3_Quantity	2

clusterEnergyLocation.csv	
Column Name	Sample Data
AddressID	125 East 105th Street10029
property_name	ChurchofStCeciliaReport
Address	125 East 105th Street
Zipcode	10029
Long	-73.947326
Lat	40.791919
Annual Energy Bill (USD)	$21,216.60

CDD-HDD-Features.csv	
Column Name	Sample Data
Property Name	ChurchofStCeciliaReport
plug_load_consumption	11.651406
ac_consumption	0.983531
domestic_gas	0.096226
heating_gas	0.366193

3. Data Preparation

Due to missing values and wrong data types, the data was cleaned up and prepared. The table below describes the data transformation that took place before analysis.

BlocPower_T.csv Transformation to DataFrame			
Column Name	Info	Rename	Summary of Data Transformation
UTSUM_Electricity_Usage	98 non-null object	Energy Usage	Remove unwanted characters, change data type to float and filled NaN with mean values.
INFO_Year of Construction	100 non-null object	Year of Construction	Convert to float type.
INFO_Number of Stories	103 non-null int64	Number of Stories	Leave as is.
INFO_Total Square Feet	103 non-null object	Square Feet	Remove unwanted characters, change data type to float.
PLEI_1_Quantity	95 non-null float64	PLEI_1	NaN values, interpreted as 0 plugged in electrical equipment, so fill NaNs with 0.
PLEI_3_Quantity	88 non-null object	PLEI_3	Convert column to float type and fill NaNs with 0.

clusterEnergyLocation.csv Transformation to DataFrame			
Column Name	Info	Rename	Summary of Data Transformation
AddressID	103 non-null object	Address ID	Leave as is.
property_name	103 non-null object	Property Name	Leave as is.
Address	103 non-null object	Address	Leave as is.
Zipcode	103 non-null int64	Zipcode	Leave as is.
Long	103 non-null float64	Longitude	Leave as is.
Lat	103 non-null float64	Latitude	Leave as is.
Annual Energy Bill (USD)	103 non-null object	Annual Energy Bill (USD)	Remove unwanted characters and change data type to float.

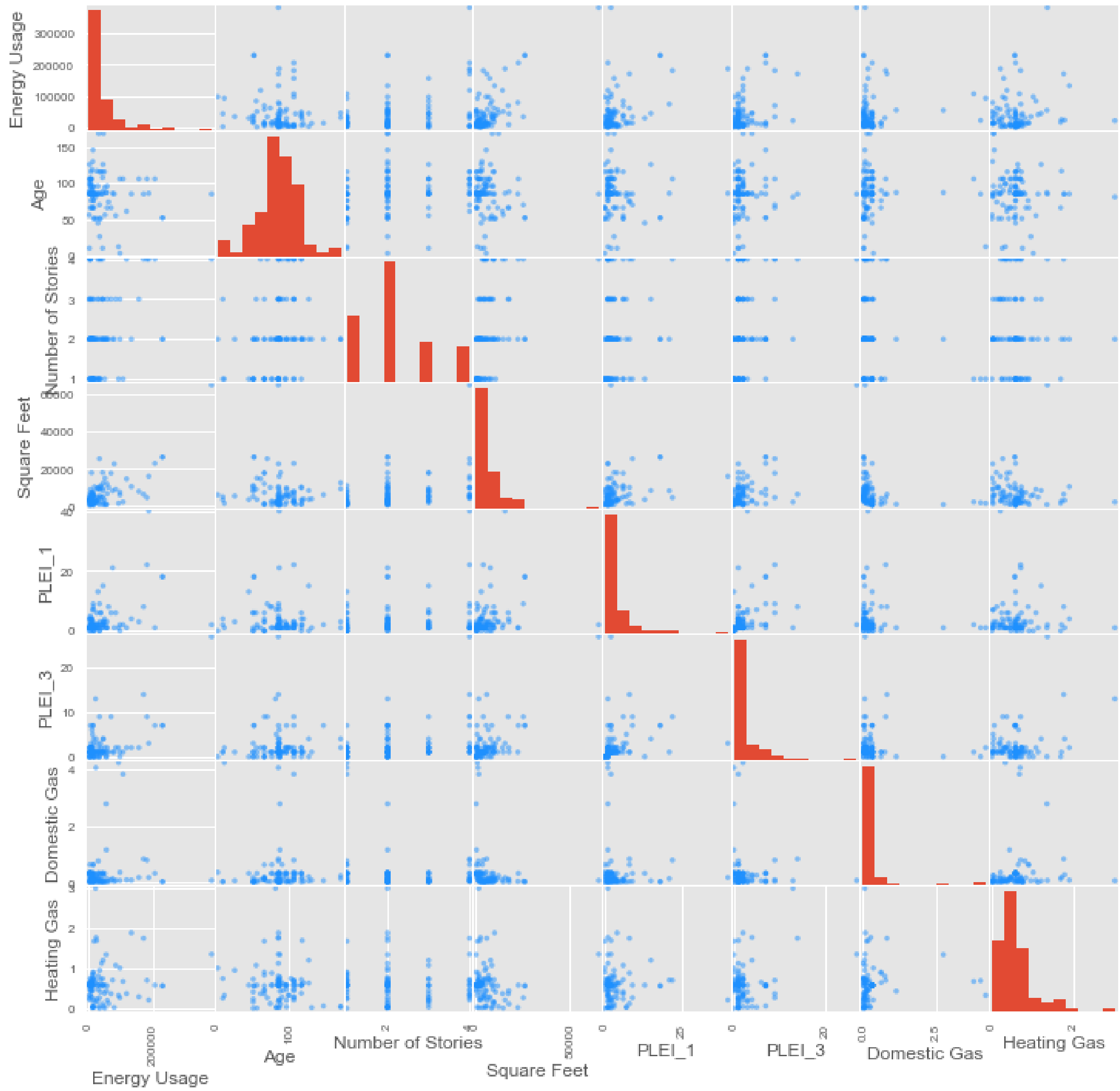
CDD-HDD-Features.csv Transformation to DataFrame			
Column Name	Info	Rename	Summary of Data Transformation
Property Name	103 non-null object	Property Name	Leave as is.
plug_load_consumption	103 non-null float64	Plug Load Consumption	Leave as is.
ac_consumption	103 non-null float64	AC Consumption	Leave as is.
domestic_gas	103 non-null float64	Domestic Gas	Leave as is.
heating_gas	103 non-null float64	Heating Gas	Leave as is.

Final Data Set:

bloc_df		
Column	Info	Sample
Property Name	103 non-null object	ChurchofStCeciliaReport
Energy Usage	103 non-null float64	117870
Age	103 non-null float64	61
Number of Stories	103 non-null int64	4
Square Feet	103 non-null float64	14600
PLEI_1	103 non-null float64	1
PLEI_3	103 non-null float64	2
Domestic Gas	103 non-null float64	0.096226
Heating Gas	103 non-null float64	0.366193
Plug Load Consumption	103 non-null float64	11.651406
AC Consumption	103 non-null float64	0.983531
Annual Energy Bill (USD)	103 non-null float64	21216.6
Year of Construction	96 non-null float64	1955
Address ID	103 non-null object	125 East 105th Street10029
Address	103 non-null object	125 East 105th Street
Zipcode	103 non-null int64	10029
Longitude	103 non-null float64	-73.947326
Latitude	103 non-null float64	40.791919

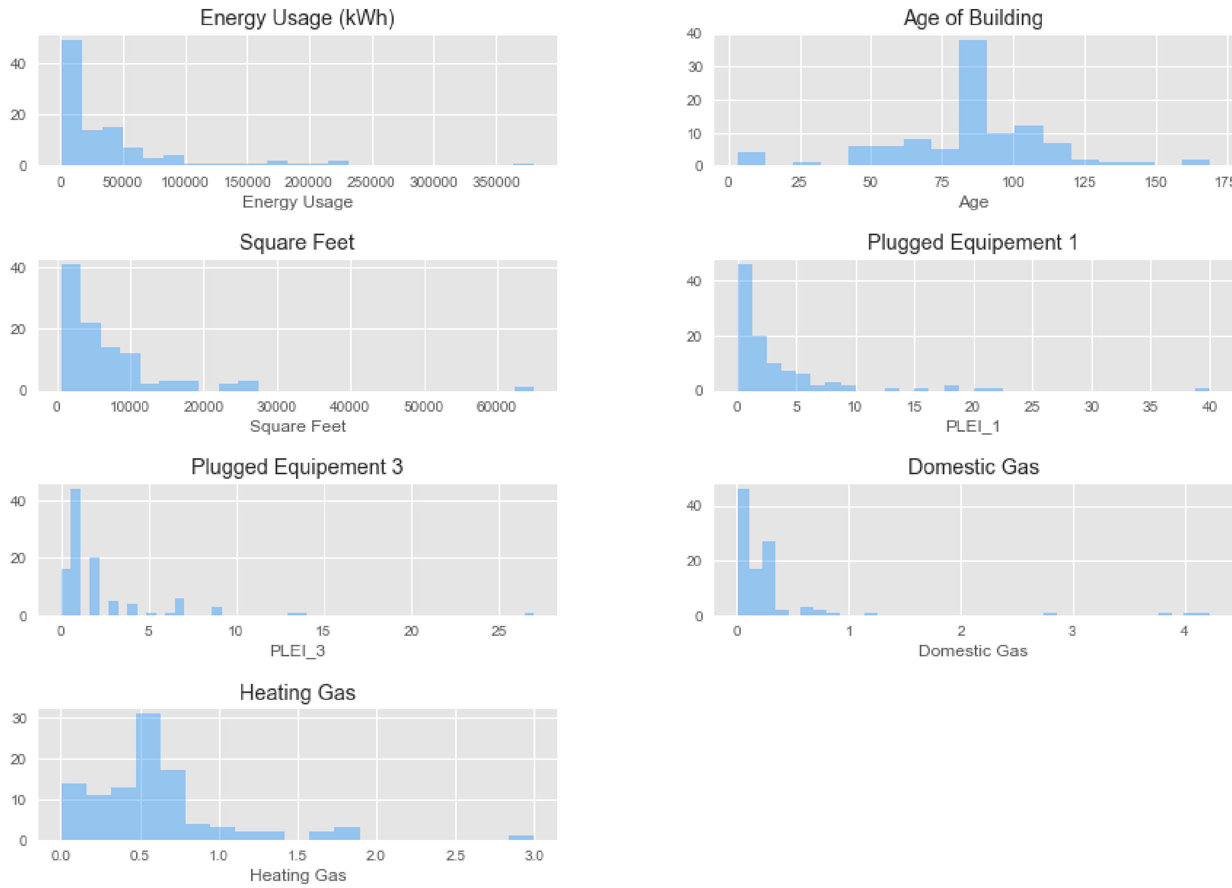
4. Exploratory Analysis

A correlation matrix using relevant parameters will allow to show variable relationship between energy usage and building characteristics.



Square feet appears to show the highest correlation with energy usage, and it does not seem there is a strong correlation between any of the parameters.

To look at a graphical distribution of parameters, histograms are plotted.

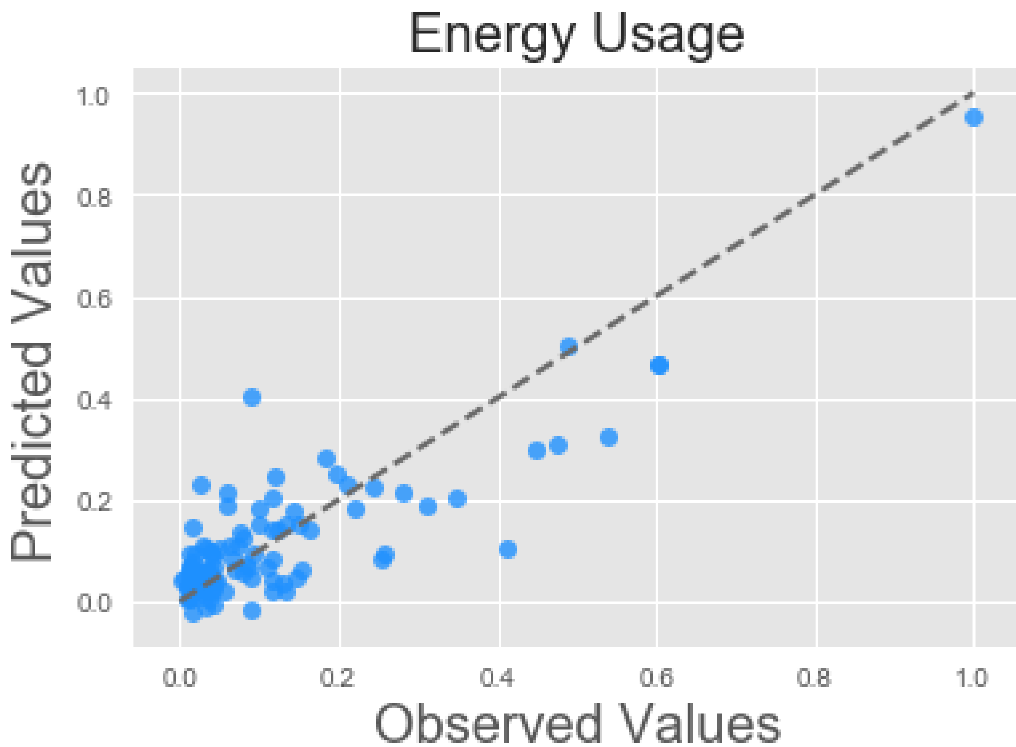


Age of Buildings shows somewhat of a normal distribution with an age building median close to 85 years. The other variables display a skewed right distribution. A regression analysis will help build a predictive model and show what variables help predict energy usage. To perform this analysis, first each parameters is scaled using sklearn's preprocessing.MaxAbsScaler(), so that the maximum value for each variable is 1. Next, a linear regression is run using linear_model.LinearRegression. The coefficients are:

Parameter	Coefficient
Intercept	-0.06714
Age	-0.0235
Number of Stories	0.048774
Square Feet	0.777122
PLEI_1	0.312308
PLEI_3	0.122954
Domestic Gas	0.229171
Heating Gas	0.143661

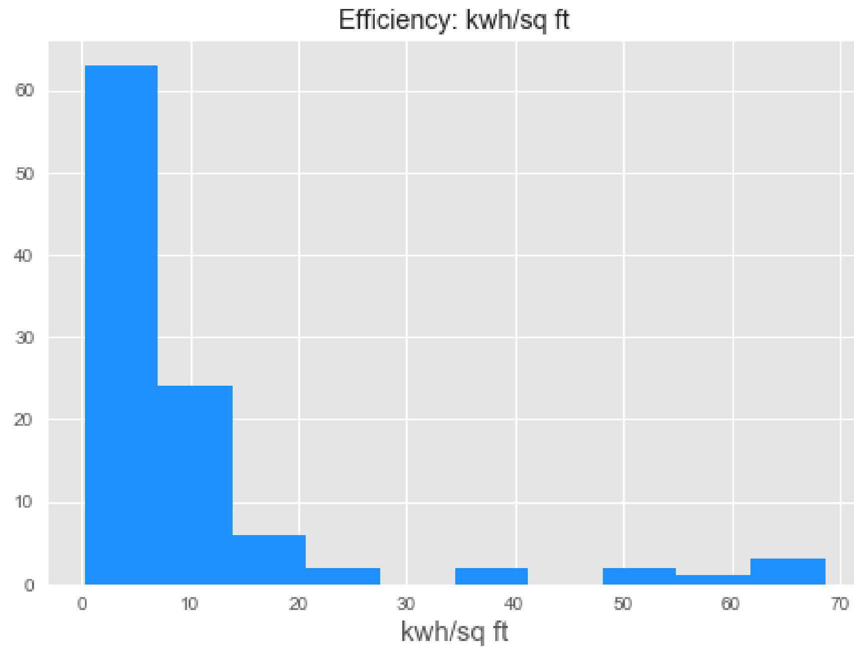
R-Squared: 0.71750454564

Square Feet is the best predictor of energy usage for a building, which is not all surprising. These parameters are then used to predict a building's energy usage. Predicted values are plotted against real energy usage value below to show accuracy of model. The dotted line represents a perfect model.



4. Labelling Inefficient Buildings

Since square footage appears to be the most relevant independent variable, energy usage /square feet is used to determine energy efficiency. The higher the ratio, the more inefficient a building is. To determine an efficiency threshold, a histogram of all ratio is plotted first.

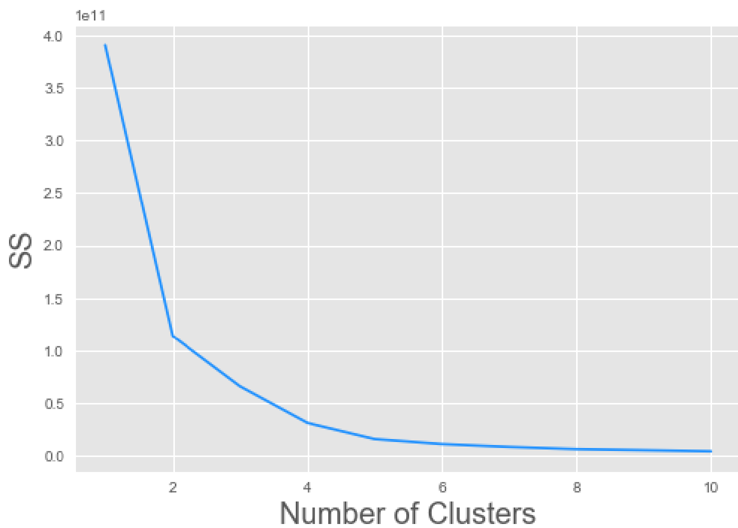


The skewed distribution shows there are a number of inefficient buildings, so buildings with an efficiency ratio>20 are labeled “True”. After labelling each building, there are 10 out of 103 building that consume energy inefficiently.

5. Modelling: Unsupervised Learning.

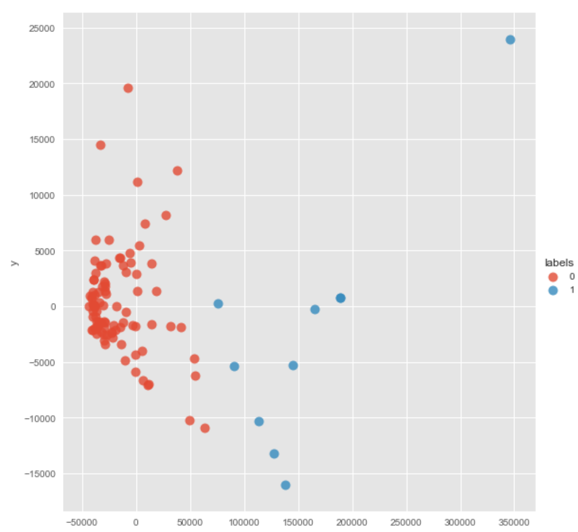
Even though the purpose of this analysis is to classify buildings into two groups, a support vector machine using sklearn's cluster KMeans classifies observation into k cluster. A graph with the sum of squares from each observation to the nearest cluster and the different k clusters is plotted. Using the elbow method, k is chosen where adding another cluster does not provide with better modelling.

Using all variables (Energy Usage, Age, Number of Stories, Square Feet, PLEI_1, PLEI_3, Domestic Gas, Heating Gas, Plug Load Consumption, AC Consumption, Annual Energy Bill (USD)):

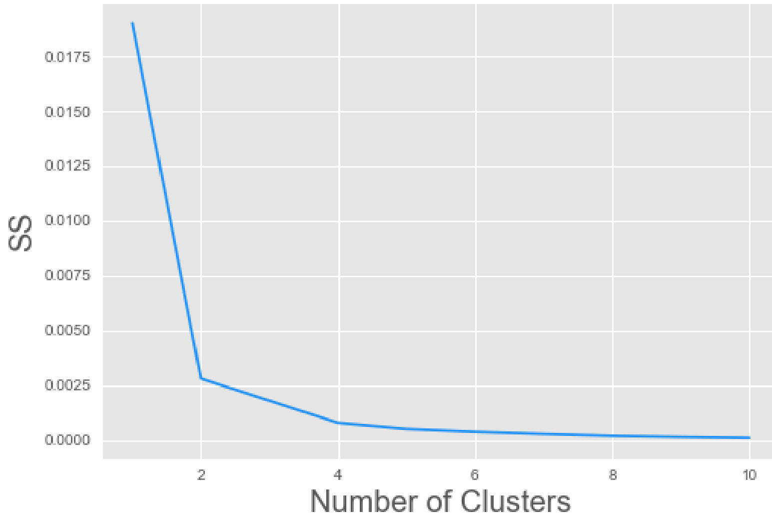


K=2.

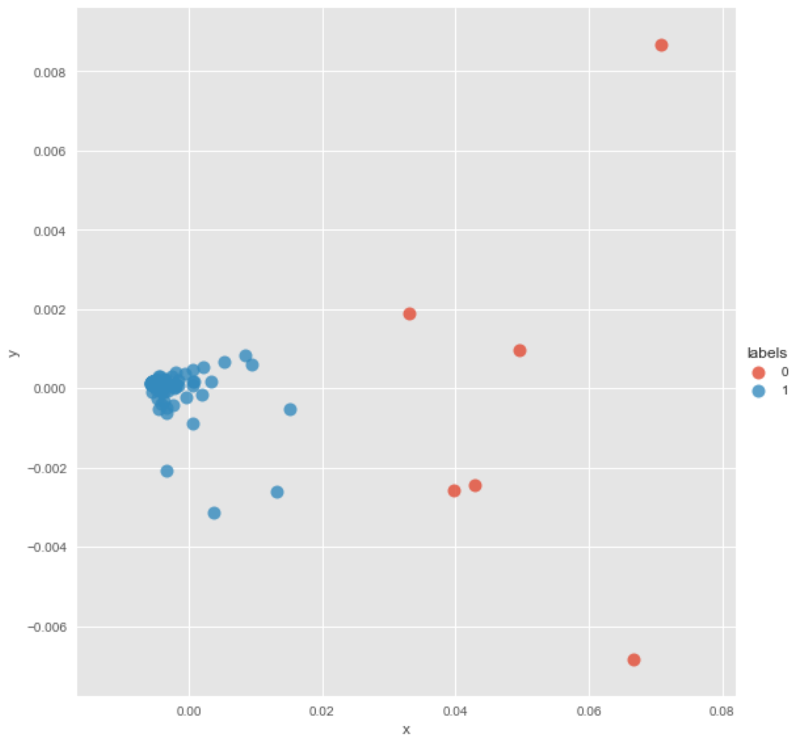
To display the clusters in a two dimensional graph, a principal components analysis is run on the variable and are plotted below.



The model accurately predicts 84.5% of the observations. Using only energy independent variables (Domestic Gas, Heating Gas, Plug Load Consumption, AC Consumption), each divided by the square footage should yield better results.



K =2

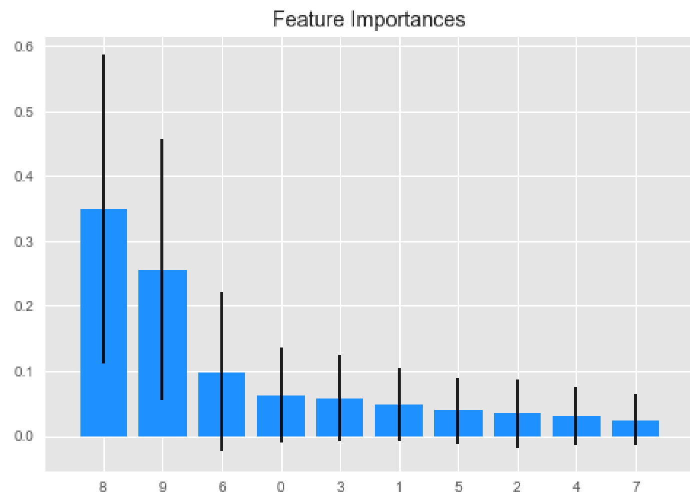


As expected, this model is more accurate, predicting 96% of the data.

6. Supervised Learning.

Since each building is labeled, an ensemble method, may obtain better predictive performance. First the data is split into training (80%) and testing (20%) testing data. Using the RandomForrestClassifier on the training data, predictions are then made with the parameters from the testing data. As it turns out, the model is able to accurately predict 100% of the testing data labels. To see which features helped the classifier best, feature importance is plotted below:

#	Feature
0	Energy Usage
1	Age
2	Number of Stories
3	Square Feet
4	PLEI_1
5	PLEI_3
6	Domestic Gas
7	Heating Gas
8	Plug Load Consumption
9	AC Consumption



Load consumption from electrical equipment, followed by air conditioning consumption, and gas used for domestic purposes are the most important leaves in the model. This seems intuitive, since new appliances and air conditioning units have become increasingly more energy efficient in the last decade. It makes sense that inefficient energy buildings probably have old electrical equipment that consume more energy.