ENERGY EFFICIENCY INDEX FOR RESIDENTIAL PROPERTIES IN BOSTON

Inform, Measure, and Discover

SPPUA 5262 Final Project

Vaishali Kushwaha

12/17/2015

Table of Contents

Introduction

In Boston, buildings account for about two-thirds of city's total greenhouse gas emissions[1]. Thus improving building energy efficiency is a significant component of Boston's Climate Action plan to reduce its greenhouse gas emissions 25 percent by 2020 and 80 percent by 2050. The city is developing and implementing initiatives to encourage residents and businesses to improve their energy use. A new initiative called "Renew Boston Whole Building Incentive" has been launched to significantly reduce the cost for energy efficiency improvements in two and three family homes[2]. There is twofold benefit of improving residential energy efficiency - offset the rising cost of energy and significantly reduce home's carbon footprint.

As the City of Boston is keen to promote residential energy efficiency, it is important to know how building stock energy efficiency is spread across the city. Identification of poorly performing areas can assist to promote the home improvement initiatives in a more effective and targeted manner. As utility data documenting energy use per household is often difficult to obtain, this analysis attempts to develop a proxy for structural energy efficiency using Tax Assessor's data. The aim is to develop an Energy Efficiency Index for residential buildings, which along with demographic and behavioral characteristic can act as indicator of energy use across Boston. The Energy Efficiency Index considers three variables – age of the building, type of heating system and type of air conditioning. The model can be used as preliminary tool to identify inefficient housing hotspots which can be then considered for a detailed energy efficiency assessment and home improvement projects.

[1] http://www.cityofboston.gov/eeos/reporting/about.asp
[2] http://www.cityofboston.gov/eeos/conservation/

Conceptual Framework

Residential energy use depends on numerous environmental, socioeconomic and structural factors. Santin et al. in their study of effects of occupancy and building characteristics on energy use in residential block found that age of the building is a very important factor; in general older houses tend to consume more energy than younger households, especially for space heating[3]. An air sealed and well insulated house needs less heating or cooling, hence it consumes less energy and allows residents to save money on operational and future upgrade costs[4]. The older buildings' design, material and structure allow easy dissipation of heat. Hence older buildings are considered energy inefficient from modern standards, unless they have undergone infiltration or retrofitting processes.

Space heating, in the US, accounts about 45 percent of household energy bills and is the largest energy expense in a residential unit[5]. Since, a less efficient heating system will use more energy to heat a given area therefore efficiency of heating system can be considered a good indicator of residential energy use. Most commonly used heating systems, in Boston, are furnaces (hot air) and boilers (hot water). Heat pumps are environment friendly alternative that meet both heating and cooling needs. Electric heating systems are the most energy efficient and most expensive as well.

Air conditioning, like space heating, is another factor that has significant contribution towards residential energy use. From an energy and environment perspective, cooling by natural means or fans are more energy efficient than air conditioning. Majority of residences in Boston do not have air conditioners, but as the climate is changing and summers are becoming hotter the use of air conditioners are likely to go up in future. The two most common type of air conditioning systems used are central and ductless.

Apart from structural characteristics, household size or number of occupants, household income, renter occupancy, and multi-family units with collective metering are

[3] Santin, O.G., Itard, L. and Vissher, H. 2009. The effect of occupancy and building characteristics on energy use for space and water heating in Dutch residential stock. *Energy and Buildings 41*. 1223-1232
[4] http://energy.gov/sites/prod/files/2014/01/f6/homeHeating.pdf
[5] http://energy.gov/sites/prod/files/2014/01/f6/homeHeating.pdf

other important household characteristics that have linear correlation with energy use. This analysis focuses on using the Tax assessor's database to study the residential building energy efficiency. The scope of this study is limited to analysis of structural influence on energy efficiency (and not all the household characteristics).

Measurement

Data Overview

City of Boston Assessing Department's centralized dataset consists of parcel-specific data for it's all the 168,146 unique parcels, for the year 2015. Boston's Assessing Department determines value of every land and building parcel, and releases it to public as part of open data initiative. The original dataset consists of 65 variables related to parcel ownership, composition, valuation, location and 2010 Census identifiers. In addition, two new variables – average land value per sq ft and average building value per sq ft, were added to normalize the values of land and buildings across the city. The dataset was also cleaned and modified for necessary corrections and improving analysis; e.g. replacing '0' with 'NA' in year built and remodeled, correcting the gross tax based on the prevailing tax rates.

The key variables that contributed to this analysis are: Land use (LU), Assessed value of land (AV_LAND), Assessed value of building (AV_BLDG), Total assessed value of the parcel (AV_TOTAL), Size of parcel (LAND_SF), Year in which property was built (YR_BUILT), Year in which property was last remodeled (YR_REMOD), living area (LIVING_AREA), Type of heating system in residential structure (R_HEAT_TYP), if residential structure has air conditioning (R_AC), and Geolocation and 2010 Census Variables.

Apart from Tax Assessor's database Census data was also incorporated in the analysis conducted at tract level. Demographics and socio-economic indicators like median income, proportion of black/white people, population density, properties occupied/vacant, unemployment rate etc. were used for correlation and regression analysis.

Model Development

The conceptual framework and key variables laid ground for developing a model for the residential energy efficiency index. The following flowchart provides an overview of the process.

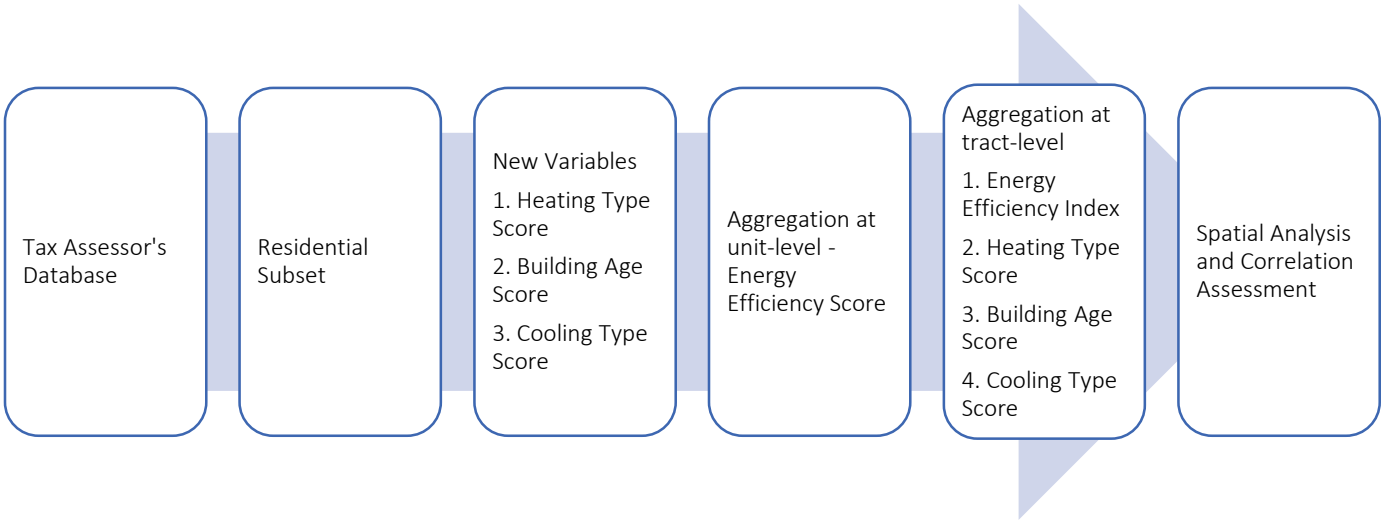


Figure 1. Process Flowchart

A residential subset was created from the larger database to select the parcels that were relevant to the objective of the study. It included: single-family, two-family, three-family, four or more family, apartments with seven or more units, and mixed use residential-commercial units.

The age of residential units was identified using the year built and year remodeled. It was assumed that remodeling can be considered a mechanism to overcome the shortcomings of aging buildings, hence for remodeled units the age was equivalent to the years passed since remodeling. A house efficiency score was developed assuming the age of the building is proportional to its heating needs. Another new variable indicating the house efficiency based on age (0-4 scale) was also created. Poor performance was represented with a lower score while higher score indicated better performance. Based on the performance, the various heating systems specified in the database were allocated an energy efficiency score. A new variable indicating the heating system efficiency score (0-4 scale) for residential

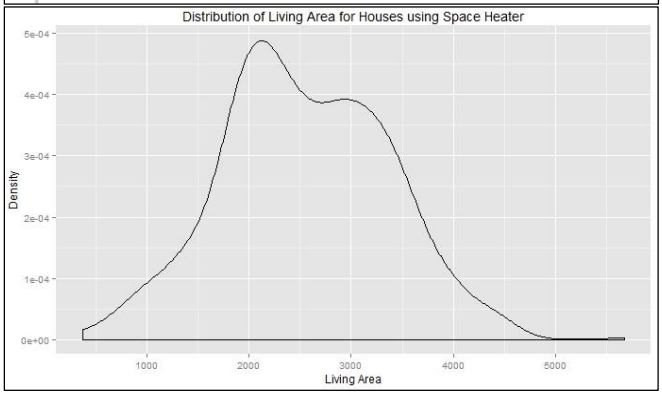
units was created where a lower score represented poor performance and higher score represented better performance. The residential units were also allocated a score based on presence and type of cooling system (1-3 scale) with lower score representing poorer energy performance.

The building age score, heating type score, cooling type and composite energy efficiency score were added at household level to form the residential energy efficiency indicator. All the four new variables were then aggregated at the tract level to facilitate the subsequent spatial analysis. The tract level energy efficiency scores were merged with the geographical data and Boston city's map to visually display the residential Energy Efficiency Index across city tracts. The tract level census variables were then incorporated in the analysis to conduct correlation and regression analysis between the Energy Efficiency Index and demographic variables.

Incorporating Living Area

Area of parcel can be important factor influencing to energy efficiency of the unit. Heating a smaller area requires lesser energy than heating a larger area. This correlation becomes significant in the special case of Space Heaters. Space heaters are recommended as a way to supplement inadequate heating, and they are efficient at heating small spaces. But use of space heaters as main heating system for large unit can be highly inefficient.

```
> summary(HTyp_S$LIVING_AREA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    378    1980    2500    2549    3155    5671
> nrow(HTyp_S)
[1] 864
```



The study of space heaters in Boston, showed that 864 parcels of single family, two-family and three-family units are using this system. A closer look at the living area of these parcels showed that only five parcels had living area under 500 square feet. Three-fourth of the parcels had living area greater than 1980 square feet, hence indicating that the parcels using space heaters are of medium to large size units. Hence for this study, use of space heaters was considered entirely inefficient.

Description of New Variables

- HEAT_SCORE This variable represents the R_HEAT_TYP in form of energy efficiency score. Each residential heating system type is allocated a numeric score based on its energy efficiency performance.

Heating Type	Label	Efficiency Range[1] [2]	Efficiency Classification	HEAT SCORE
Space Heater	S	Inefficient at house scale, fire hazard	Low	0
Hot Water	W	50-90%	Low - Medium	1
Heat Pump	P	6.8-10 HSP	Medium	2
Forced Air	F	59-98.5%	High	3
Electric	E	95-100%	Highest	4
Other	O	NA	NA	NA
None	N	NA	NA	NA

- BLDG_AGE This variable represents the effective age of residential units, using YR_BUILT and YR_REMOD. It is assumed that remodeling can be considered a mechanism to overcome the shortcomings of aging buildings, hence for remodeled units the age was equivalent to the years passed since remodeling.
- AGE_SCORE This variable represents residential unit energy efficiency based on the age of building. The scores were allocated on the assumption that older buildings are more energy inefficient.

Residential Unit Age	AGE SCORE
> 200 yrs	0
150 – 200 yrs	1
100 – 150 yrs	2
50 – 100 yrs	3
< 50 yrs	4

- COOL_SCORE This variable represents the R_AC in form of energy efficiency score. Each residential cooling system type is allocated a numeric score based on its energy efficiency performance. The tax assessor's data has three values for air conditioning: None, Ductless, and Central; they were allocated a score of 3, 2, and 1 respectively. The general idea being cooling by natural means or fans are more energy efficient than air conditioning.
- EE_SCORE This is an aggregate variable that combines the HEAT_SCORE, COOL_SCORE and AGE_SCORE in weighted sum. The age of the building was provided higher weightage because an inefficient structure will most likely result in energy wastage no matter how efficient the heating or cooling system is. The variable indicates the parcel specific composite energy efficiency index.

$$EE\_SCORE = AGE\_SCORE+0.75*HEAT\_SCORE+0.75*COOL\_SCORE$$

- AV_LAND_PER_SF expresses the assessed value of a parcel's land, divided by its area in square feet.
- AV_BLDG_PER_SF expresses the assessed value of a parcel's building, divided by its gross area in square feet.

Census Tract level Aggregated Variables

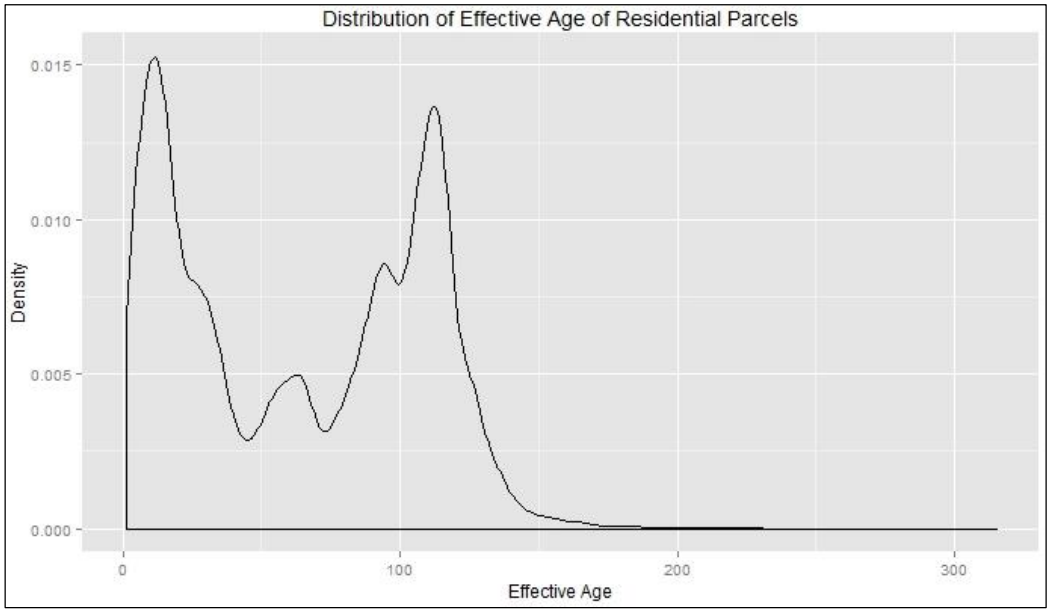
- *age.mean* The aggregation of parcel level energy efficiency score for building age component.
- *heat.mean* The aggregation of parcel level energy efficiency score for residential heating system component.
- *cool.min* The aggregation of parcel level energy efficiency score for residential air conditioning system component.
- *EE.mean* The aggregation of the parcel level composite energy efficiency index score.
- *bldg.mean* The aggregation of parcel level average building value per square feet.
- *value.mean* The aggregation of parcel level total building value (land value plus building value) per square feet.

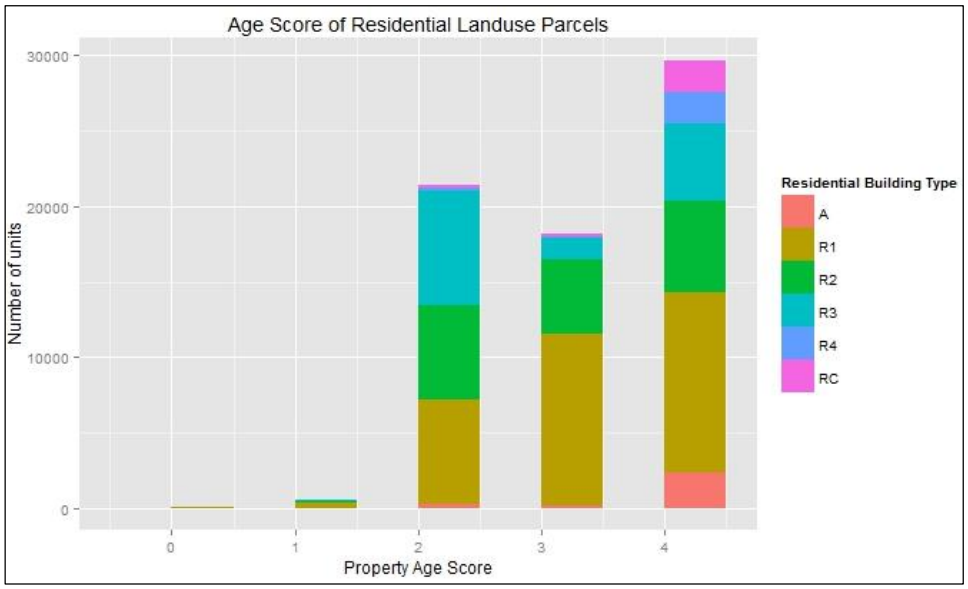
Data Analysis and Discovery

Building Age Score

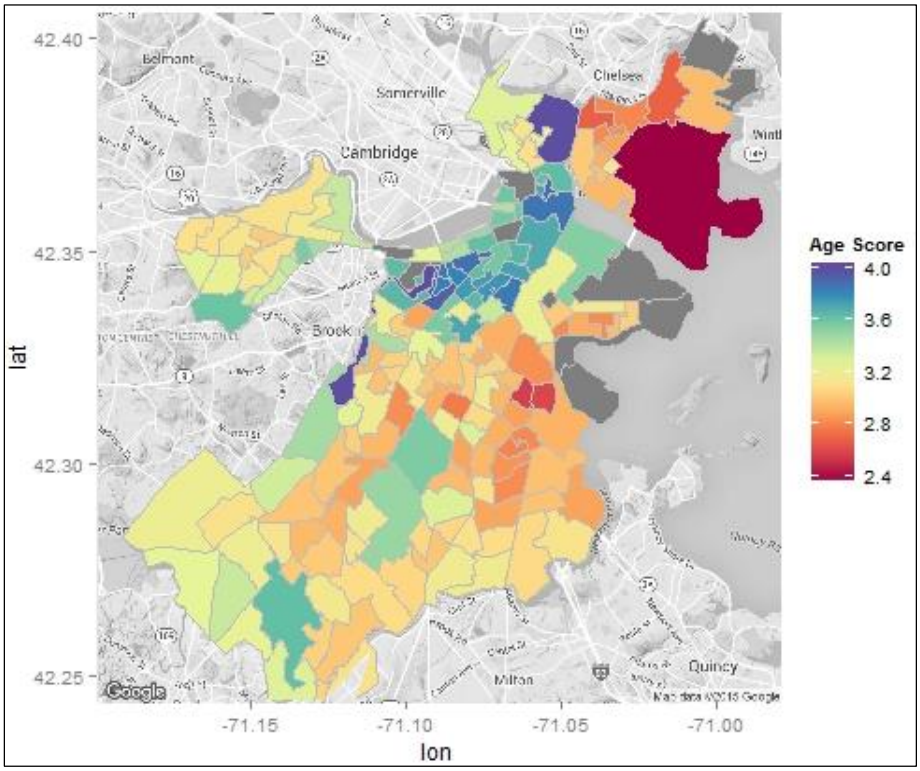
The effective age (based on when the property was built and most recently renovated) of residential properties show three-fourth of all the residential parcels is 18 years or more. Majority of the parcels are less than 100 years of age. The effective age distribution shows two peaks, one around 20years and another around 120 years. These might be the indication of booms in property construction and renovation. Accordingly the majority of buildings are new or renovated, with age score 4 (<50 years) and score 3 (50<age<100 years).

```
> summary(PTax_R_EE$BLDG_AGE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   1.00   18.00   65.00   64.31  105.00  315.00     517
> table(PTax_R_EE$AGE_SCORE,PTax_R_EE$LU)

        A    R1    R2    R3    R4    RC
  0     0    27    11     2     0     0
  1     0   303   133    82     1     0
  2   251  6965  6193  7593   248   189
  3   165 11334  5001  1385   134   180
  4  2291 11963  6098  5109  2092  2118
```





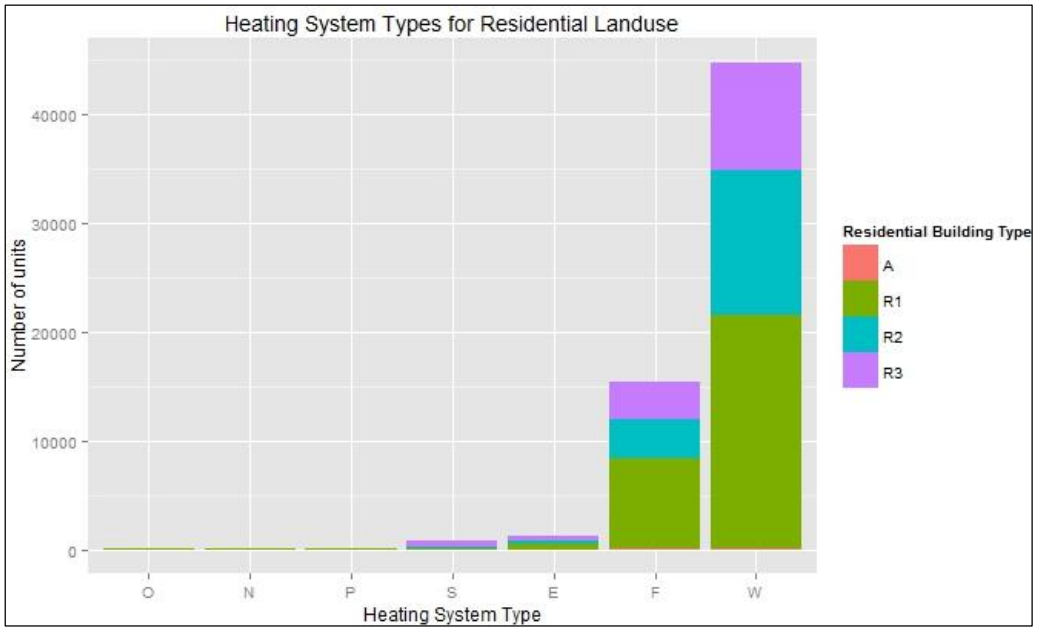
The spatial distribution of residential building age score, aggregated at Census tract level, show the mean score ranging from low of 2.4 to maximum of 4. It can be observed that newly constructed or renovated buildings are concentrated in the Downtown, North end, China town, Beacon Hill, Back Bay, South End, South Boston Waterfront; and few tracts in Allston/Brighton, Roxbury and Roslindale/Hyde park. The older buildings are concentrated in East Boston, South Boston, Dorchester, Roxbury and Jamaica Plain.



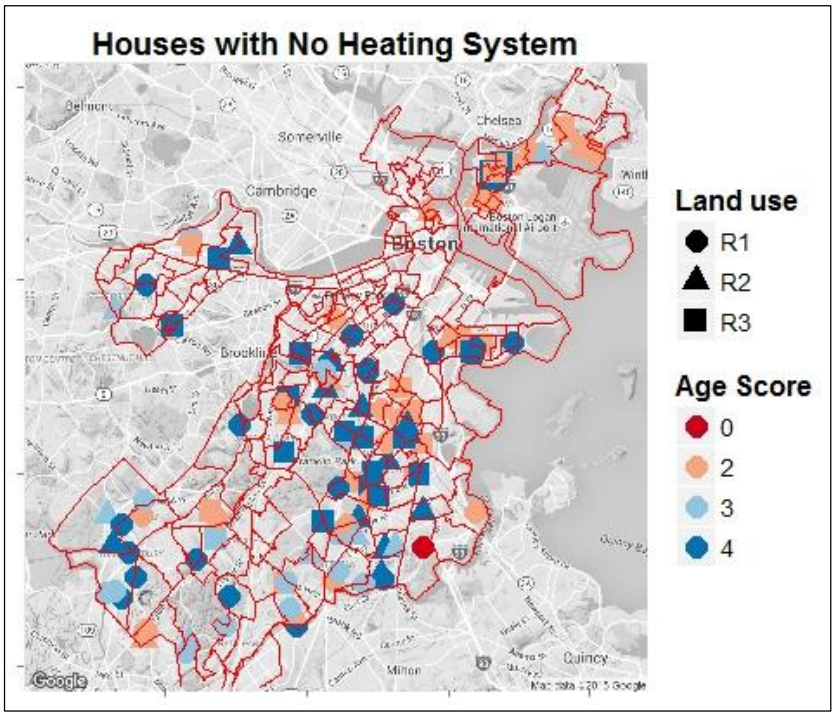
Heating System Type Score

The analysis shows that the heating system type has been described primarily for single-family (R1), two-family (R2), and three-family (R3) houses. The heating system type for majority of apartments (A) and all of four or more family houses (R4) and mixed residential use (RC) is unspecified. The most common type of heating system used is the Hot Water system (W) followed by the Forced Air system (F). The rest, Electric, Heat pump and Space heaters are used in rare cases.

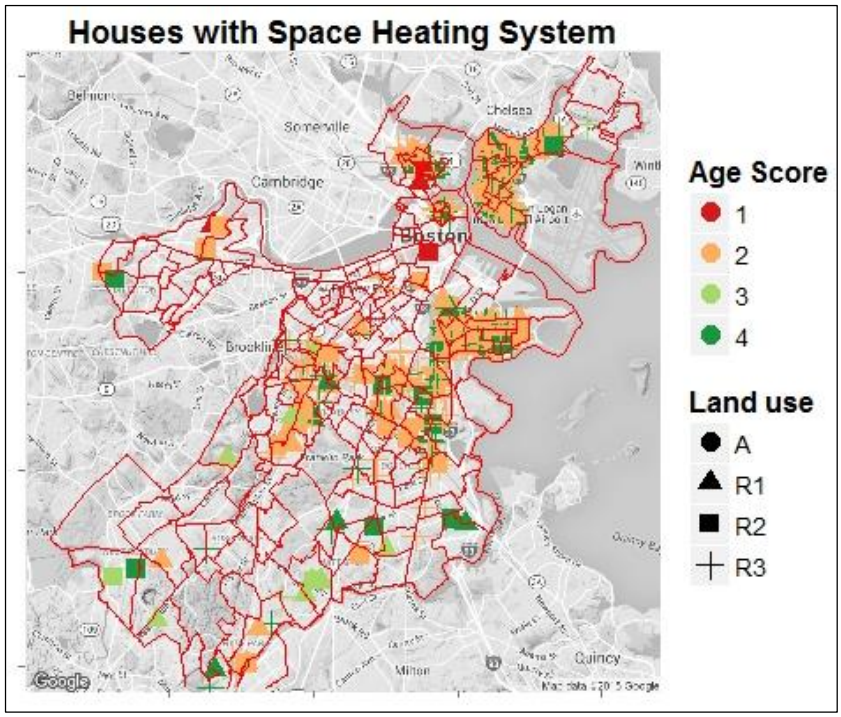
	A	R1	R2	R3	R4	RC
	2817	1	0	1	2565	2607
E	8	535	307	337	0	0
F	32	8401	3587	3312	0	0
N	0	51	34	36	0	0
O	0	15	8	8	0	0
P	0	88	19	14	0	0
S	4	88	151	621	0	0
W	148	21416	13330	9843	0	0



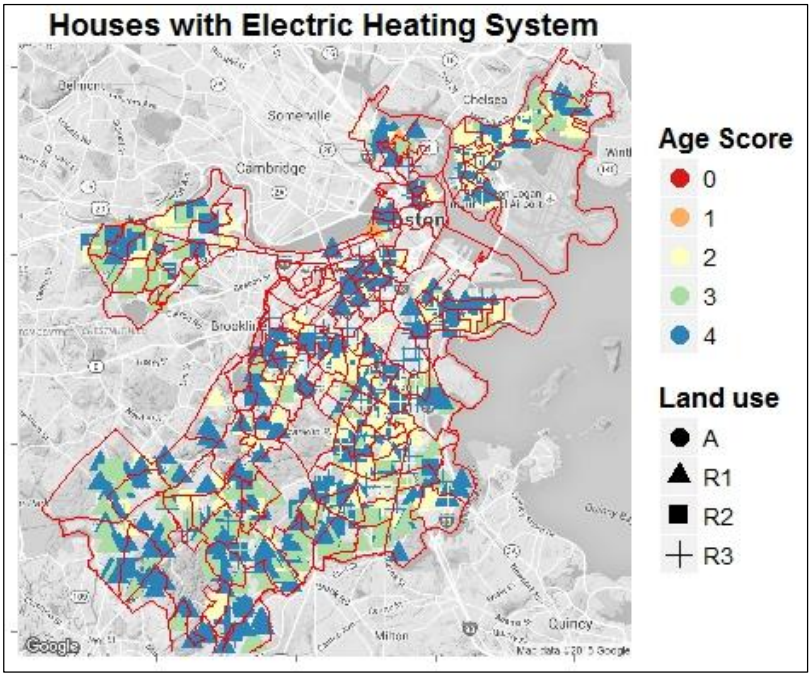
There are 121 parcels without any heating systems and they are concentrated in East Boston, South Boston, Dorchester, Roxbury and West Roxbury. There is no correlation between type of residential parcel (land use) and absence of heating system, but the parcels without heating system are primarily less than 50 yrs old (Age score 4) or between 100-150 years old (Age score 2).



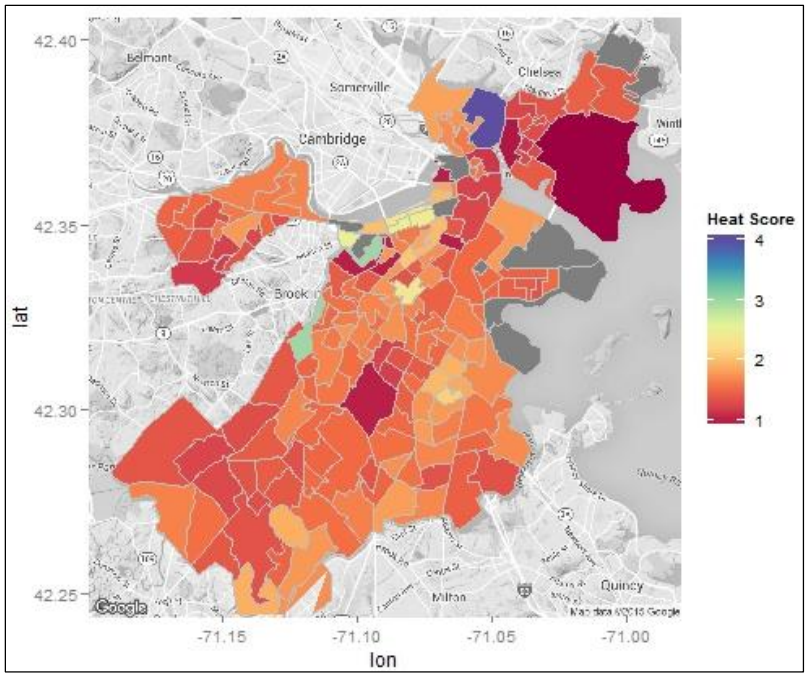
About 860 parcels use space heaters to fulfill their heating needs and majority of them are located in East Boston, South Boston, Dorchester, Roxbury and Jamaica Plain. Again, there is no clear relation between type of residential property (LU) and use of space heaters, but space heaters are excessively used in properties ageing 100-150 years (Age score 2).



In contrast to above findings, the electric heating system were spread all across the city and majority of them were used in new or newly renovated buildings (Score 4 and 3). The electric resistance heating is 100% energy efficient in the sense that all the incoming electric energy is converted to heat, but it has high operational cost.



The spatial distribution of residential building heating system type score, aggregated at Census tract level, show the mean score ranging from low of 1 to maximum of 4. It can be observed that the entire city performs below average (below Score 2.5) with only few tracts in Back Bay performing better. This is because the water heaters, scoring low on energy efficiency (Score 1), are the dominating system, and tends to skew the outcome. The tracts which are performing relatively poorer are located in East Boston, North

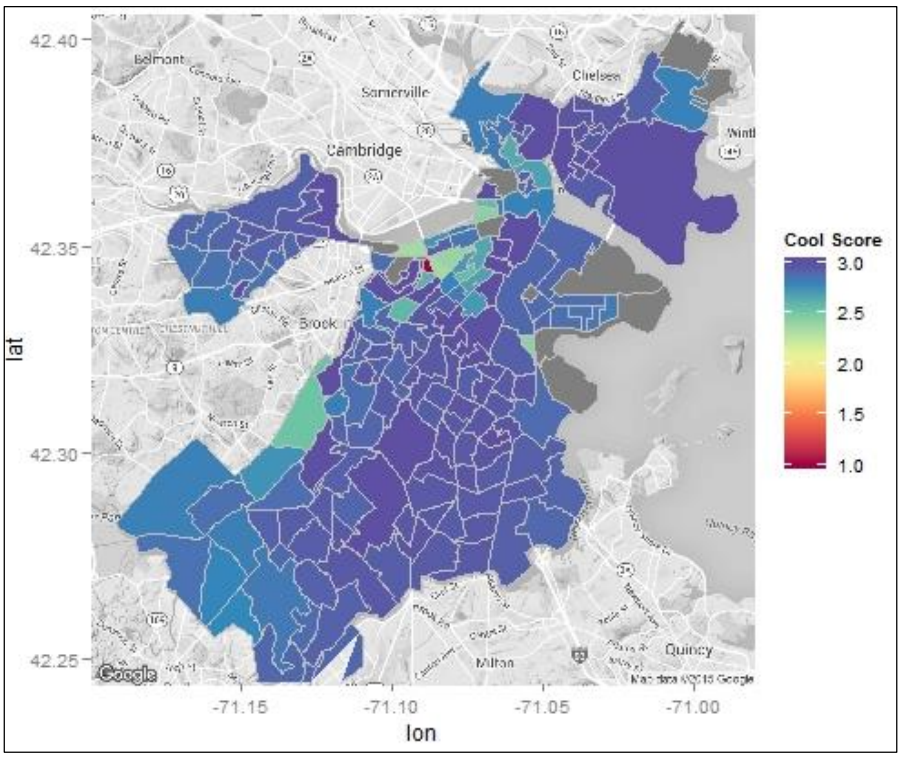


Boston, Downtown, Chinatown, Roxbury, South Dorchester, Jamaica Plain, West Roxbury, Roslindale, Hyde Park and Brighton.

Cooling System Type Score

Similar to heating system, cooling system is also not defined for majority of apartments (A) and all of four or more family houses (R4) and mixed residential use (RC). Moreover only a limited number of residential parcels are using air conditioning, and the ones that do have a centralized air conditioning system (C). The majority of residential parcels do not use cooling systems (N). Hence the cooling system type score is heavily skewed towards Score 3, as shown in the map below. The areas which perform relatively poorer, because of the heavy use of centralized air conditioning, are concentrated in Back Bay, South End and Jamaica Plain neighborhoods. This variable is likely to make insignificant impact on the composite index of residential building energy efficiency.

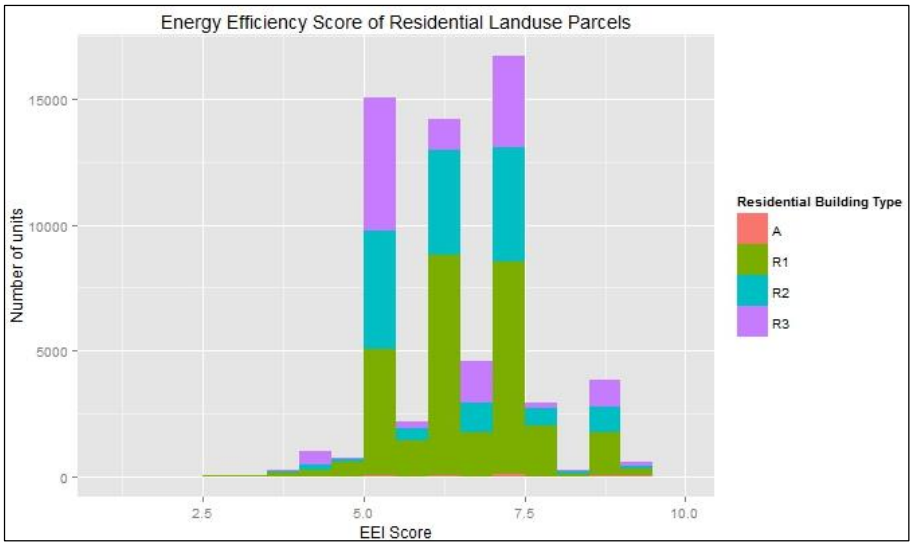
```
> table(PTax_R_EE$R_AC,PTax_R_EE$LU)

      A    R1    R2    R3    R4    RC
  2817     1     0     1  2565  2607
C    1  5106  1383   651     0     0
D    0    44    16     6     0     0
N  191 25444 16037 13514     0     0
```



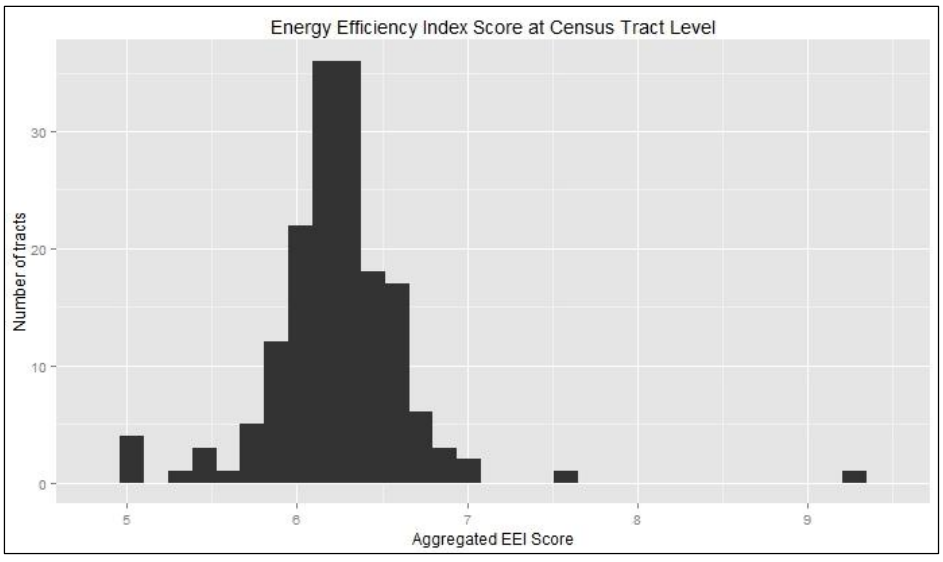
Energy Efficiency Index Score

The Energy Efficiency Index for residential parcels ranged from minimum 1.5 to maximum 9.25 (with possibility of minimum 0.75 to maximum 9.25). Majority of the properties scored between 5 and 7. This mediocre performance was magnified when the parcel level scores were aggregated Census tract level. The tract-level Energy Efficiency Index indicates that residential housing energy efficiency ranges from minimum 5 to 9.25, with a huge majority tracts scoring between 6 and 6.5.

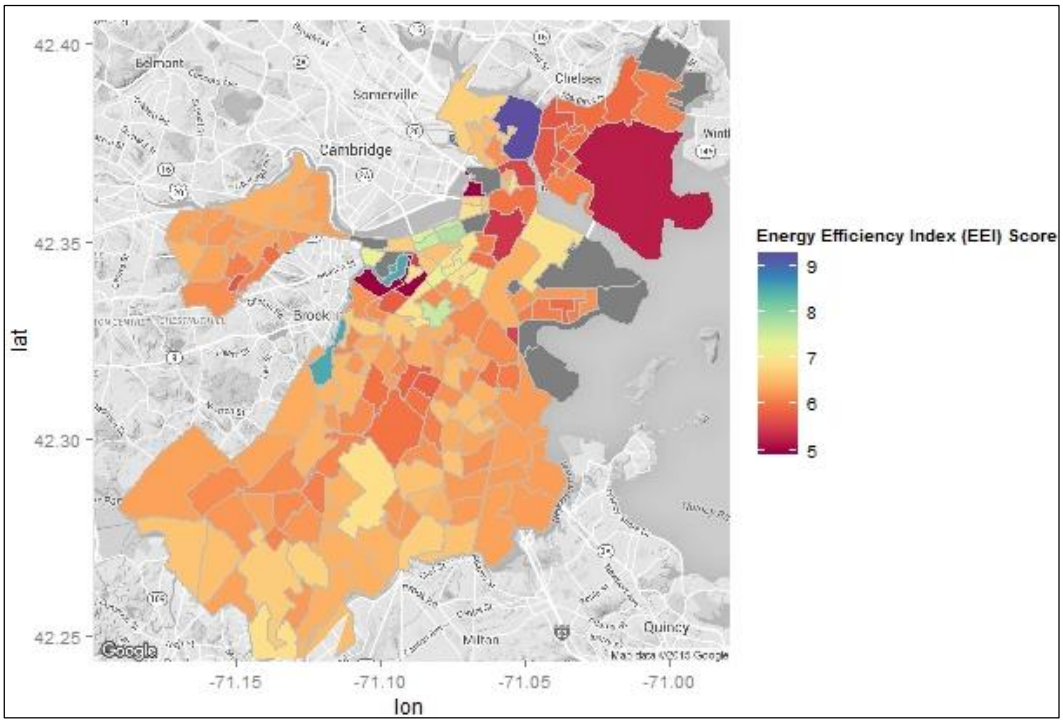
```
> summary(PTax_R_EE$EE_SCORE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  1.500   5.000   6.000   6.252   7.000   9.250    8146
```




```
summary(EE.mean$EE_SCORE)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
5.000   6.055   6.233   6.234   6.394   9.250
```



With a majority of houses using similar type of heating and cooling system, the composite index was expected to be skewed. But for this study the objective was to identify the areas where the residential building stock is performing poorly on the Energy Efficiency Index. And the tract level aggregation of Energy Efficiency Index points to such areas. All the tracts in East Boston, North end, Downtown, and Chinatown area are performing poorly as compared to the rest of the city. Some tracts in South Boston, Roxbury, North Dorchester, Allston/Brighton and Jamaica Plains are also performing poorly. The areas doing better than rest are in Back Bay, South End, Waterfront, Roxbury South neighborhoods.



Correlation and Regression Analysis

In order to study the relation between Energy Efficiency Index and other key economic and demographic variables, a correlation and regression analysis was conducted. A correlation study was conducted on property occupied, median income, proportion of white population, population density, average building value per square feet and Energy Efficiency Index. As the Energy Efficiency Index is applicable to only residential units, the correlation and regression analysis was spatially limited to residential census tracts only. Moreover there was one single outlier tract, with extreme average building value per square feet that was skewing the analysis. Hence the outlier tract was also removed from the analysis.

The correlation analysis showed that Energy Efficiency Index was positively related to properties occupied (r=0.003, p <0.97), median income(r=0.237, p <0.009), proportion of black population (r=0.004, p <0.247), population density (r=0.106, p <0.247), average building value per square feet (r=0.584, p <2.51e-12). The null hypothesis can be rejected for median incomes and average building value per square feet.

	r.Poccupied	r.medincome	r.propblack	r.popdens	r.EE_SCORE	r.AV_BLDG_PER_SF
Poccupied	1.000000000	0.14441529	-0.360364586	-0.005631976	0.003379987	0.1315608
medincome	0.144415289	1.00000000	-0.508904934	0.020290244	0.237490937	0.4687958
propblack	-0.360364586	-0.50890493	1.000000000	-0.252325207	0.003823566	-0.4718514
popdens	-0.005631976	0.02029024	-0.252325207	1.000000000	0.106474750	0.5304883
EE_SCORE	0.003379987	0.23749094	0.003823566	0.106474750	1.000000000	0.5841042
AV_BLDG_PER_SF	0.131560788	0.46879581	-0.471851438	0.530488253	0.584104180	1.0000000

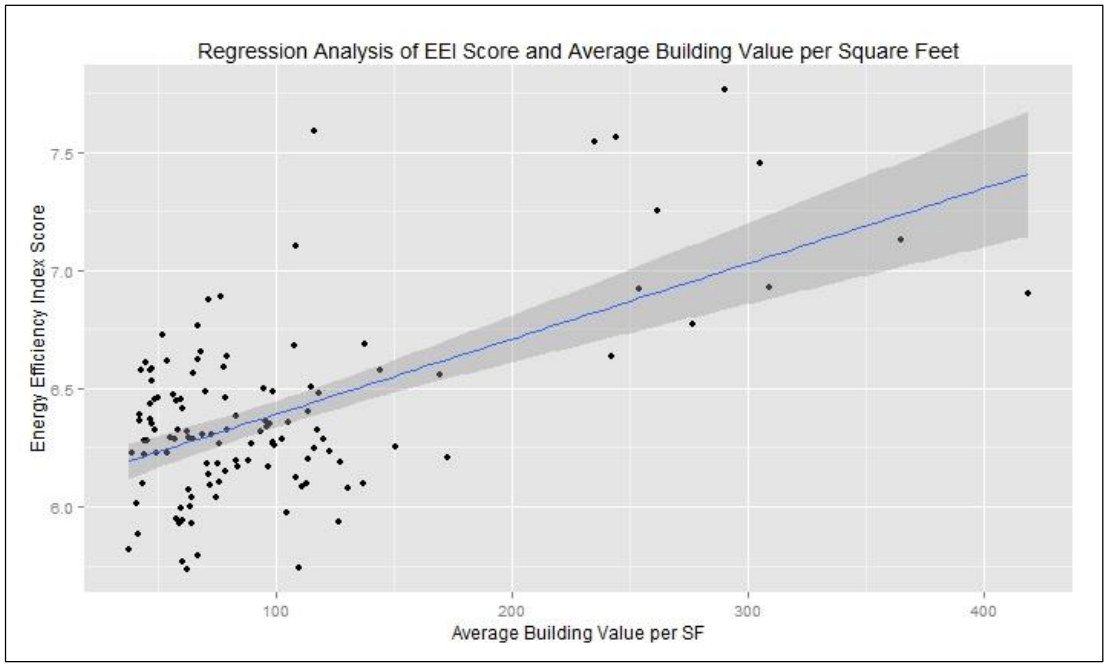
	P.Poccupied	P.medincome	P.propblack	P.popdens	P.EE_SCORE	P.AV_BLDG_PER_SF
Poccupied	NA	1.155584e-01	5.277588e-05	9.513194e-01	9.707733e-01	1.520508e-01
medincome	1.155584e-01	NA	2.940711e-09	8.258987e-01	9.005863e-03	6.656109e-08
propblack	5.277588e-05	2.940711e-09	NA	5.431050e-03	9.669397e-01	5.321308e-08
popdens	9.513194e-01	8.258987e-01	5.431050e-03	NA	2.470854e-01	4.610121e-10
EE_SCORE	9.707733e-01	9.005863e-03	9.669397e-01	2.470854e-01	NA	2.507328e-12
AV_BLDG_PER_SF	1.520508e-01	6.656109e-08	5.321308e-08	4.610121e-10	2.507328e-12	NA

The regression analysis was conducted keeping Energy Efficiency Index constant and the rest of the selected economic and demographic indicators as variables. The model justifies about 48% (R-squared = 0.477, p-value 9.4e-16) of variation in the Energy Efficiency Index.

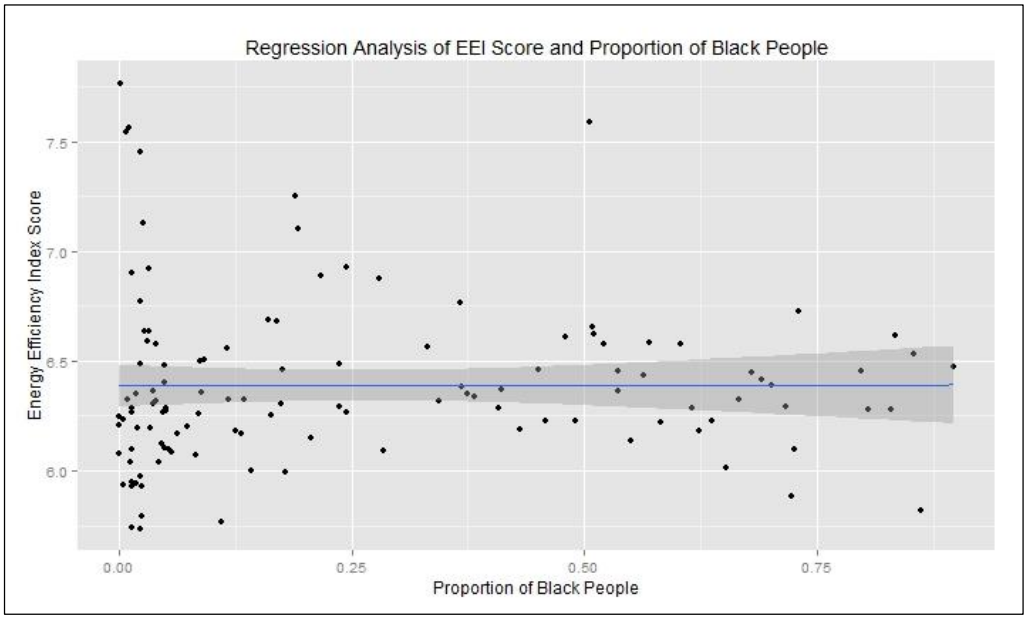
```
Call:
lm(formula = EE_SCORE ~ Poccupied + medincome + propblack + popdens
    AV_BLDG_PER_SF, data = tracts[!dataout, ])

Residuals:
     Min       1Q   Median       3Q      Max
-0.74663 -0.15018 -0.02895  0.13196  0.92346

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.831e+00  5.721e-01  10.192  < 2e-16 ***
Poccupied       1.215e-01  5.969e-01   0.204  0.83907
medincome       6.908e-08  1.106e-06   0.062  0.95032
propblack       5.152e-01  1.222e-01   4.216 5.00e-05 ***
popdens        -7.435e-06  2.224e-06  -3.344  0.00112 **
AV_BLDG_PER_SF  4.918e-03  5.166e-04   9.520 3.66e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2755 on 114 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.499,     Adjusted R-squared:  0.477
F-statistic: 22.71 on 5 and 114 DF,  p-value: 9.433e-16
```

The multiple regression analysis shows Energy Efficiency Index is

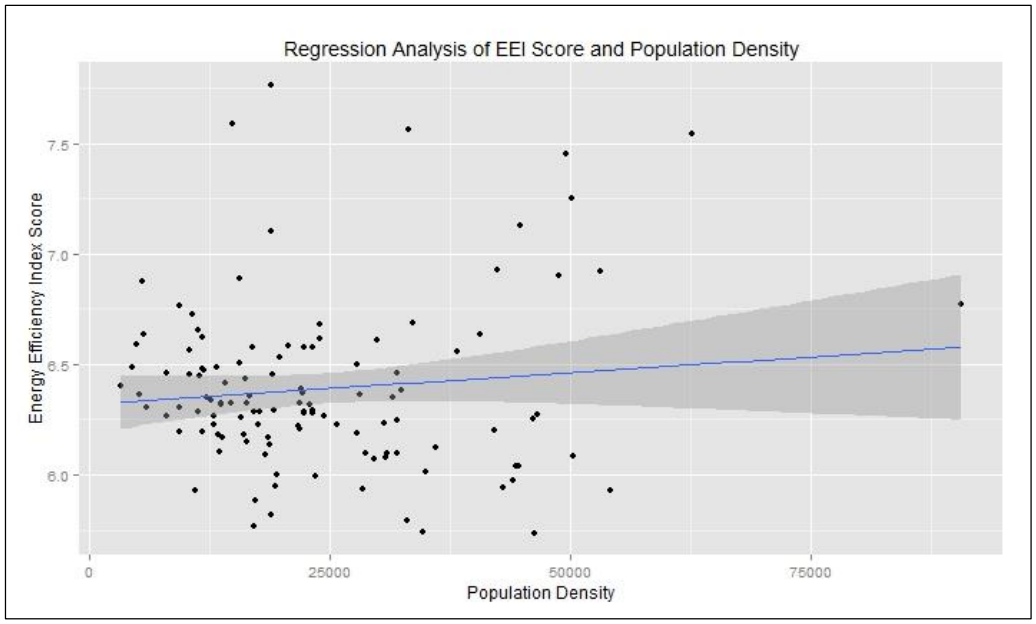
- Positively and strongly correlated to average building value per square feet (coefficient = 0.0049, p-value =0). Therefore it can be concluded that the energy efficiency increases as the building value per square feet increases, that is parcels with higher values are also likely to me more energy efficient.



- Strongly correlated to proportion of black population (coefficient = 0.5, p-value =0). The energy efficiency score seems to be independent of proportion of black population.



- Strongly correlated to population density (coefficient = -7.4e-06, p-value =0.001). The energy efficiency score seems to be somewhat positively correlated with population density, indicating small improvement in energy efficiencies in dense tracts.



Conclusion and Discussion

The objective of this project was to use Tax Assessor's 2015 data to develop a measurement that can detect energy inefficient residential building stock in Boston. In order to achieve the objective an Energy Efficiency Index consisting of three components: effective age of the parcel, type of heating system used and type of cooling system used, was developed. The composite index especially points towards residential parcels in East Boston, and this is also the neighborhood that has higher concentration of space heater use and parcels without heating. Some other poorly performing areas are in Roxbury and city's central area (North End, Downtown, Chinatown, and South End). The central area performs well on age score, indicating newer or recently renovated buildings but the use of low scoring heating system type pulls the overall energy efficiency down. On the other hand, tracts performing poorly in Roxbury house older parcels as well heating systems of low efficiency.

In conclusion, the energy efficiency measurement has potential to become a useful tool to support the ongoing initiative -- Renew Boston Whole Building Incentive, for improving energy efficiency in Duplexes and Triple-Deckers. It can indicate towards tracts and blocks that are likely to be in need of home weatherization or retrofitting, hence guiding city officials for directed outreach. But the measurement is only as good as the information. This measurement is not full-proof because it measures efficiency based on type of heating and cooling system, but in practice their age/condition of such systems and type of fuel used dominate their efficiency.

The herein mentioned Energy Efficiency Index provides a rudimentary idea of the structural efficiency of residential parcels. Combining this with other household and behavioral factors, like household size, income, and renter/owner occupied, can produce a more advanced and comprehensive indicator of residential energy use. Because in addition to the type and condition of the infrastructure, the behavior of people living in these buildings also influences the total energy used by a property.

Appendix –R Syntax

```
PTax_EE <- read.csv('V:/PPUA 5262 Big Data/Data/Tax Assessor 2015 - Data.csv', stringsAsFactors = FALSE)
## Residential subset
PTax_R_EE<- PTax_EE[PTax_EE$LU =='R1'| PTax_EE$LU == 'R2'| PTax_EE$LU == 'R3'| PTax_EE$LU == 'R4'| PTax_EE$LU
== 'A'| PTax_EE$LU == 'RC',]
View(PTax_R_EE)

##Heat Type and LU plot
HType_LU_Tab <- table(PTax_R_EE$R_HEAT_TYP,PTax_R_EE$LU)
View(HType_LU_Tab)

HType <- PTax_R_EE[PTax_R_EE$R_HEAT_TYP == 'E'|PTax_R_EE$R_HEAT_TYP == 'F'|PTax_R_EE$R_HEAT_TYP ==
'N'|PTax_R_EE$R_HEAT_TYP == 'O'|PTax_R_EE$R_HEAT_TYP == 'P'|PTax_R_EE$R_HEAT_TYP ==
'S'|PTax_R_EE$R_HEAT_TYP == 'W',]
HTypeTab <- table(HType$R_HEAT_TYP)
HTypeorder <- names(HTypeTab)[order(HTypeTab)]
HType$R_HEAT_TYP2 <- factor(HType$R_HEAT_TYP, levels = HTypeorder)
ggplot(HType,aes(R_HEAT_TYP2))+geom_histogram(aes(fill=LU))+ labs(title = 'Heating System Types for Residential
Landuse',x = 'Heating System Type', y = 'Number of units', fill = 'Residential Building Type')

## Space Heater and Living Area Distribution
HTyp_S <- PTax_R_EE[PTax_R_EE$R_HEAT_TYP == 'S',]
nrow(HTyp_S)
summary(HTyp_S$LIVING_AREA)
ggplot(HTyp_S,aes(LIVING_AREA)) + geom_density() + labs(title = 'Distribution of Living Area for Houses using Space
Heater', x = 'Living Area', y = 'Density')

## Allocating Residential Heat Type Energy Efficiency Score
PTax_R_EE$HEAT_SCORE <- NA
PTax_R_EE$HEAT_SCORE <- ifelse(PTax_R_EE$R_HEAT_TYP == 'S', 0, PTax_R_EE$HEAT_SCORE)
PTax_R_EE$HEAT_SCORE <- ifelse(PTax_R_EE$R_HEAT_TYP == 'W', 1, PTax_R_EE$HEAT_SCORE)
PTax_R_EE$HEAT_SCORE <- ifelse(PTax_R_EE$R_HEAT_TYP == 'P', 2, PTax_R_EE$HEAT_SCORE)
PTax_R_EE$HEAT_SCORE <- ifelse(PTax_R_EE$R_HEAT_TYP == 'F', 3, PTax_R_EE$HEAT_SCORE)
PTax_R_EE$HEAT_SCORE <- ifelse(PTax_R_EE$R_HEAT_TYP == 'E', 4, PTax_R_EE$HEAT_SCORE)

## Allocating age to Residential buildings
PTax_R_EE$YR_BUILT <- ifelse(PTax_R_EE$YR_BUILT == '',NA,PTax_R_EE$YR_BUILT)
PTax_R_EE$YR_BUILT <- ifelse(PTax_R_EE$YR_BUILT == 0,NA,PTax_R_EE$YR_BUILT)
PTax_R_EE$YR_REMOD <- ifelse(PTax_R_EE$YR_REMOD == '',NA,PTax_R_EE$YR_REMOD)
PTax_R_EE$YR_REMOD <- ifelse(PTax_R_EE$YR_REMOD == 0,NA,PTax_R_EE$YR_REMOD)
PTax_R_EE$BLDG_AGE <- ifelse(is.na(PTax_R_EE$YR_REMOD), (2015 - PTax_R_EE$YR_BUILT), (2015 -
PTax_R_EE$YR_REMOD))
PTax_R_EE$BLDG_AGE <- ifelse(PTax_R_EE$BLDG_AGE <=0,NA,PTax_R_EE$BLDG_AGE)
PTax_R_EE$BLDG_AGE <- ifelse(PTax_R_EE$BLDG_AGE == 1020,NA,PTax_R_EE$BLDG_AGE)

##Allocating Building Age Score
PTax_R_EE$AGE_SCORE <- NA
PTax_R_EE$AGE_SCORE <- ifelse(PTax_R_EE$BLDG_AGE < 50,4,PTax_R_EE$AGE_SCORE)
PTax_R_EE$AGE_SCORE <- ifelse(PTax_R_EE$BLDG_AGE >= 50,3,PTax_R_EE$AGE_SCORE)
PTax_R_EE$AGE_SCORE <- ifelse(PTax_R_EE$BLDG_AGE >= 100,2,PTax_R_EE$AGE_SCORE)
PTax_R_EE$AGE_SCORE <- ifelse(PTax_R_EE$BLDG_AGE >= 150,1,PTax_R_EE$AGE_SCORE)
PTax_R_EE$AGE_SCORE <- ifelse(PTax_R_EE$BLDG_AGE >= 200,0,PTax_R_EE$AGE_SCORE)

## Allocating Residential Air Conditioner Energy Efficiency Score
PTax_R_EE$COOL_SCORE <- NA
PTax_R_EE$COOL_SCORE <- ifelse(PTax_R_EE$R_AC == 'C', 1, PTax_R_EE$COOL_SCORE)
PTax_R_EE$COOL_SCORE <- ifelse(PTax_R_EE$R_AC == 'D', 2, PTax_R_EE$COOL_SCORE)
PTax_R_EE$COOL_SCORE <- ifelse(PTax_R_EE$R_AC == 'N', 3, PTax_R_EE$COOL_SCORE)
```

```
View(PTax_R_EE)

##Summary of Cooling Systems
table(PTax_R_EE$R_AC,PTax_R_EE$LU)
ggplot(PTax_R_EE,aes(COOL_SCORE))+geom_bar(aes(fill=LU), binwidth = 0.5, xlim=(4))+ labs(title = 'Age Score of Residential Landuse Parcels',x = 'Property Age Score', y = 'Number of units', fill = 'Residential Building Type')

##Summary of Age
summary(PTax_R_EE$BLDG_AGE)
ggplot(PTax_R_EE,aes(BLDG_AGE))+geom_density()+labs(title = 'Distribution of Effective Age of Residential Parcels', x='Effective Age', y='Density')
table(PTax_R_EE$AGE_SCORE,PTax_R_EE$LU)
ggplot(PTax_R_EE,aes(AGE_SCORE))+geom_bar(aes(fill=LU), binwidth = 0.5, xlim=(4))+ labs(title = 'Age Score of Residential Landuse Parcels',x = 'Property Age Score', y = 'Number of units', fill = 'Residential Building Type')

## Aggregate Energy Efficiency score at building level
PTax_R_EE$EE_SCORE <- PTax_R_EE$AGE_SCORE+0.75*PTax_R_EE$HEAT_SCORE+0.75*PTax_R_EE$COOL_SCORE
summary(PTax_R_EE$EE_SCORE)
table(PTax_R_EE$EE_SCORE,PTax_R_EE$LU)
ggplot(PTax_R_EE,aes(EE_SCORE))+geom_bar(aes(fill=LU), binwidth = 0.5, xlim=(4))+ labs(title = 'Energy Efficiency Score of Residential Landuse Parcels',x = 'EEI Score', y = 'Number of units', fill = 'Residential Building Type')

## Aggregate Heat, Age and EE Score at Tract Level
heat.mean <- aggregate(HEAT_SCORE~CT_ID_10, data = PTax_R_EE,mean, na.rm = TRUE)
age.mean <- aggregate(AGE_SCORE~CT_ID_10, data = PTax_R_EE,mean, na.rm = TRUE)
cool.mean <- aggregate(COOL_SCORE~CT_ID_10, data = PTax_R_EE,mean, na.rm = TRUE)
EE.mean <- aggregate(EE_SCORE~CT_ID_10, data = PTax_R_EE,mean, na.rm = TRUE)
View(EE.mean)
summary(EE.mean$EE_SCORE)
ggplot(EE.mean,aes(EE_SCORE))+ geom_histogram()+ labs(title = 'Energy Efficiency Index Score at Census Tract Level',x = 'Aggregated EEI Score', y = 'Number of tracts', fill = 'Residential Building Type')

#Adding the Normalized land and Property value Variables
PTax_R_EE$GROSS_AREA <- ifelse(PTax_R_EE$GROSS_AREA < 100, NA, PTax_R_EE$GROSS_AREA)
PTax_R_EE$LAND_SF <- ifelse(PTax_R_EE$LAND_SF < 100, NA, PTax_R_EE$LAND_SF)
PTax_R_EE <- transform(PTax_R_EE, AV_LAND_PER_SF = PTax_R_EE$AV_LAND/PTax_R_EE$LAND_SF, AV_BLDG_PER_SF = PTax_R_EE$AV_BLDG/PTax_R_EE$GROSS_AREA)
PTax_R_EE$AV_LAND_PER_SF <- ifelse(PTax_R_EE$AV_LAND_PER_SF == 0, NA, PTax_R_EE$AV_LAND_PER_SF)
PTax_R_EE$AV_BLDG_PER_SF <- ifelse(PTax_R_EE$AV_BLDG_PER_SF == 0, NA, PTax_R_EE$AV_BLDG_PER_SF)
PTax_R_EE$TOTAL_PER_SF <- PTax_R_EE$AV_LAND_PER_SF+PTax_R_EE$AV_BLDG_PER_SF


## Aggregate building and total value per gross area
bldg.mean <- aggregate(AV_BLDG_PER_SF~CT_ID_10, data = PTax_R_EE,mean, na.rm = TRUE)
value.mean <- aggregate(TOTAL_PER_SF~CT_ID_10, data = PTax_R_EE,mean, na.rm = TRUE)

##merge tract level scores
tract.score1 <- merge(heat.mean, age.mean, by='CT_ID_10',all.x=TRUE)
tract.score2 <- merge(tract.score1,cool.mean,by= 'CT_ID_10',all.x=TRUE)
tract.score3 <- merge(tract.score2,EE.mean,by= 'CT_ID_10',all.x=TRUE)
tract.score4 <- merge(tract.score3,bldg.mean,by= 'CT_ID_10', all.x = TRUE)
tract.score5 <- merge(tract.score4,value.mean,by= 'CT_ID_10', all.x = TRUE)
View(tract.score5)


####GIS MAPPING ####
require(rgdal)
require(sp)
require(ggplot2)
require(ggmap)
```

```
require(rgeos)

### import shape file (.shp)###
tracts_geo<-readOGR(dsn="V:/PPUA 5262 Big Data/Data/GIS Data", layer ="Tracts_Boston BARI")
plot(tracts_geo)
proj4string(tracts_geo)

### make tracts_geo accessible to ggplot
tracts_geo<-fortify(tracts_geo, region = "CT_ID_10")
View(tracts_geo)

###merge tracts_geo with additional attributes
## tracts_EE<-read.csv("V:/PPUA 5262 Big Data/R Assignments/Tract EE Data.csv")
tracts_geo<-merge(tracts_geo,tract.score5,by.x='id',by.y='CT_ID_10',all.x=TRUE)
tracts_geo<-tracts_geo[order(tracts_geo$order),]
View(tracts_geo)

###get google map of Boston
Boston<-get_map(location = c(left=-71.193799,bottom=42.15,right=-70.985746,top=42.5), color = "bw")
base<-ggmap(Boston)
base

my.map.theme <- theme(plot.title = element_text(size = 18, face="bold"),
                      axis.title.x = element_blank(),
                      axis.title.y = element_blank(),
                      axis.text.x = element_blank(),
                      axis.text.y = element_blank(),
                      legend.title = element_text(size = 16, face = "bold"),
                      legend.text = element_text(size = 14))

myPalette <- colorRampPalette(brewer.pal(10, "Spectral"))

###Houses with No Heat on map
NoHeat <- PTax_R_EE[PTax_R_EE$R_HEAT_TYP == 'N',]
View(NoHeat)

NoHeat_map <- base + geom_point(data=NoHeat, aes(x = X, y = Y, color = factor(AGE_SCORE), shape = factor(LU)), size =
4) +
  geom_path(aes(x=long, y=lat, group=group), color = 'red', data=tracts_geo) +
  labs(title = "Houses with No Heating System", color = "Age Score", shape = "Land use") + guides(size = FALSE) +
  scale_colour_brewer(type = "div", palette = 5) + my.map.theme

NoHeat_map

###Houses with Space Heater on map
SHeat <- PTax_R_EE[PTax_R_EE$R_HEAT_TYP == 'S',]
View(SHeat)

SHeat_map <- base + geom_point(data=SHeat, aes(x = X, y = Y, color = factor(AGE_SCORE), shape = factor(LU)), size =5) +
  geom_path(aes(x=long, y=lat, group=group), color = 'red', data=tracts_geo) +
  labs(title = "Houses with Space Heating System", color = "Age Score", shape = "Land use") + guides(size = FALSE) +
  scale_colour_brewer(type = "div", palette =8) + my.map.theme
SHeat_map

###Houses with Electric Heater on map
EHeat <- PTax_R_EE[PTax_R_EE$R_HEAT_TYP == 'E',]
View(EHeat)

EHeat_map <- base + geom_point(data=EHeat, aes(x = X, y = Y, color = factor(AGE_SCORE), shape = factor(LU)), size =5) +
  geom_path(aes(x=long, y=lat, group=group), color = 'red', data=tracts_geo) +
  labs(title = "Houses with Electric Heating System", color = "Age Score", shape = "Land use") + guides(size = FALSE) +
```

```
  scale_colour_brewer(type = "div", palette =9) + my.map.theme
EHeat_map

## EEI Score on Map
base+ geom_polygon(aes(x=long, y=lat, group=group, fill=EE_SCORE), data=tracts_geo) + geom_path(aes(x=long, y=lat,
group=group), color = 'gray', data=tracts_geo)+ scale_fill_gradientn(colours = myPalette(100),
limits=range(tracts_geo$EE_SCORE)) + labs(fill='Energy Efficiency Index (EEI) Score')

base+ geom_polygon(aes(x=long, y=lat, group=group, fill=HEAT_SCORE), data=tracts_geo) + geom_path(aes(x=long, y=lat,
group=group), color = 'gray', data=tracts_geo)+ scale_fill_gradientn(colours = myPalette(100),
limits=range(tracts_geo$HEAT_SCORE)) + labs(fill='Heat Score')

base+ geom_polygon(aes(x=long, y=lat, group=group, fill=AGE_SCORE), data=tracts_geo) + geom_path(aes(x=long, y=lat,
group=group), color = 'gray', data=tracts_geo)+ scale_fill_gradientn(colours = myPalette(100),
limits=range(tracts_geo$AGE_SCORE)) + labs(fill='Age Score')

base+ geom_polygon(aes(x=long, y=lat, group=group, fill=COOL_SCORE), data=tracts_geo) + geom_path(aes(x=long, y=lat,
group=group), color = 'gray', data=tracts_geo)+ scale_fill_gradientn(colours = myPalette(10),
limits=range(tracts_geo$COOL_SCORE)) + labs(fill='Cool Score')


####Correlation and Regression Analysis ####
## Merge tract level data of EEI Project with Deomographics data
demog <- read.csv('V:/PPUA 5262 Big Data/Data/Tract Census Data.csv', stringsAsFactors = FALSE)
View(demog)
tracts<-merge(demog,tract.score5,by='CT_ID_10',all.x=TRUE)
View(tracts)

##subset of only residential tracts and omitting outlier of Average building and total value
dataout=tracts$AV_BLDG_PER_SF>450 | tracts$Res == 0 | tracts$Type == 'I' | tracts$Type == 'P'
View(dataout)
View(tracts[!dataout,])

## Correlation - EE Score (50), Bldg Value per SF (51), homeownership(18), mdeium income(22), population density(38),
proportion of white population (28)
require(Hmisc)
names(tracts)
correlations<-rcorr(as.matrix(tracts[!dataout,][c(15,22,29,38,50,51)]))
summary(correlations)

View(correlations[1])
View(correlations[2])
View(correlations[3])

####Regression - EE Score, Bldg Value per SF, Property Density, Proportion of White at tract Level
regression<-lm(EE_SCORE~Poccupied+medincome+propblack+popdens+AV_BLDG_PER_SF, data=tracts[!dataout,])
class(regression)
summary(regression)

require(ggplot2)
base1<-ggplot(data=tracts[!dataout,], aes(x=AV_BLDG_PER_SF, y=EE_SCORE)) + geom_point() + labs(title='Regression
Analysis of EEI Score and Average Building Value per Square Feet', x='Average Building Value per SF', y= 'Energy
Efficiency Index Score')
base1 + geom_smooth(method=lm)

base2<-ggplot(data=tracts[!dataout,], aes(x=propblack, y=EE_SCORE)) + geom_point() + labs(title='Regression Analysis
of EEI Score and Proportion of Black People', x='Proportion of Black People', y= 'Energy Efficiency Index Score')
base2 + geom_smooth(method=lm)

base3<-ggplot(data=tracts[!dataout,], aes(x=popdens, y=EE_SCORE)) + geom_point() + labs(title='Regression Analysis of
EEI Score and Population Density', x='Population Density', y= 'Energy Efficiency Index Score')
```

```
base3 + geom_smooth(method=lm)

##GGally plots##
require(GGally)
ggpairs(data=tracts[!dataout,], columns=c(18,22,28,38,50,51))

##Extract residuals##
View(data.frame(regression$residuals))
tracts<-merge(tracts,data.frame(regression$residuals),by='row.names',all.x=TRUE)

tracts_geo<-fortify(tracts_geo, region = "CT_ID_10")
tracts_geo<-merge(tracts_geo,tracts,by.x='id',by.y='CT_ID_10',all.x=TRUE)
tracts_geo<-tracts_geo[order(tracts_geo$order),]

Boston<-get_map(location=c(left = -71.193799, bottom = 42.15, right = -70.985746, top = 42.5))
Bostonmap<-ggmap(Boston)
Bostonmap+ geom_polygon(aes(x=long, y=lat, group=group, fill=regression.residuals), data=tracts_geo)+
geom_path(aes(x=long, y=lat, group=group), color = 'gray', data=tracts_geo)
```