Using Permit, Parcel Data to Investigate Residence Energy Systems

Jason Lu: jasonlu6@bu.edu
Zixin "Cindy" Ding: cindydzx@bu.edu
Fang "Hugh" Qu: hughqu@bu.edu

1. Abstract

BU's Institute for Sustainable Energy (ISE) is a University-wide institute that promotes faculty research. We have partnered with Michael Walsh and Adam Pollack to design and analyze a database of tax parcels and building permits for populating a building energy systems inventory and investigating the trends and distribution of heating and energy efficient systems. The goals of the project are: 1) determine which communities are more likely to have heat pumps, based on various demographics, 2) determine which communities are benefitting from energy efficient systems or not, and 3) implement a predictive model which determines which parcel building has a particular type of heating system. 4) Visualize the distribution of heating systems in Boston. We incorporate the permit and parcel database, and the parcel data key dataset to help BU ISE determine which buildings are most likely to have heat pumps.

2. Data Sources

The data is downloaded from the Analyze Boston website, which contains the majority of the CSV files that serves as a basis for our data analysis. All data is attributed to the Department of Technology and Innovation for the Analyze Boston website.

2.1.1 Boston Property Assessments: this dataset is Boston's Assessor's database of tax parcels which contains fields for building heating and cooling systems. We download five datasets for the final project, with the fiscal years from 2014 to 2018 to do exploratory data analysis.

2.1.2 Boston Building Permits: this dataset contains text descriptions about building renovations which includes changes to building energy systems.

2.1.3 Parcel Data Key by Boston Assessing Department: this data source is a legend for determining the descriptions of the Boston Property Assessments attributes.

3. Cleaning / Processing / Collection of Data

3.1 Boston Property Assessments: For the parcel database, we extract the parcel ID (unique data key in two dataset), zip code, type of land use, average total assessed property valuation (housing and property cost), living area, type of heating systems(dependent variable), owner occupancy (whether owner occupied), air conditioning and cooling systems for normal residential and condo unit(R_AC and U_AC), and heating systems for residential and condo unit

(R_HEAT_TYP and U_HEAT_TYP). For cleaning part, we melt four columns above to three columns, indicating each combination of air conditioning, type of heating system, and residential house or condo unit, (R_AC, U_AC, R_HEAT_TYP, U_HEAT_TYP) to (R or U, HEAT_TYP, AC) and calculate the average assessed value per sq ft, by dividing average total assessed value by the living area for each parcel ID. Since Michael focuses on residence heat types, we filter out commercial, agricultural, industrial type property data. We aggregate five years (2014-2018) Boston Property Assessment Data and ignore all NAN rows in selected columns above, which significantly reduce nearly 50% of original data. (660,000 to 316,300)

3.2 Boston Building Permits: we use regular expression parsing techniques to extract the parcel ID, comments, and location (longitude/latitude). We ignore all NaN rows.

To get the top 10 most frequent words in the comments, we implement the TF IDF, combined with SVD (term frequency) method with a corpus of 17253 terms for 363754 rows(Unique Parcel ID). Since the matrix is big and sparse, we use truncated SVD to select the top 100 frequent words. We then sort and extract the most frequent words from the sorted corpus.

The top 10 most frequent terms in the corpus is as follows: water, heater, new, install, replace, gas, kitchen, boiler, existing, replacement. These words are relatable to permit dataset, which records every renovation in residence in Boston area.

4. Approach

As for answering the questions, we first applied statistical modeling methods to predict the class (P) heat pumps. However, four common classifiers are not able to provide enough predictions for the class (P). The reason is that the portion of class (P) is so small as compared to other classes. Even though we set the class weight to balanced, the result is not ideal, then we concluded that the prediction model is unable to provide enough information for class (P). More details of Modeling can be found in section 5.1.1.

Next, we first figured out the rate of change of each heat type for both residential and condo buildings. The details of rate of change can be found in section 5.2. We then visualizate heat systems distribution, average assessed value per square feet, owner occupancy type by years in map visualization part to answer the question of which type of communities is benefit from the energy efficient system(solar panels) or has heat pumps installed.

5. Experimental Results & Visualization

5.1 Statistical Models

5.1.1 Modeling

We employ four common classifiers in order to identify the model that yields the best accuracy

for our dataset. Namely, we employ logistic regression (LR), Gaussian Naive Bayes (GNB), support vector machines (SVM), and a decision tree classifier (DTC). A breakdown of the accuracy of each classifier model is shown below:

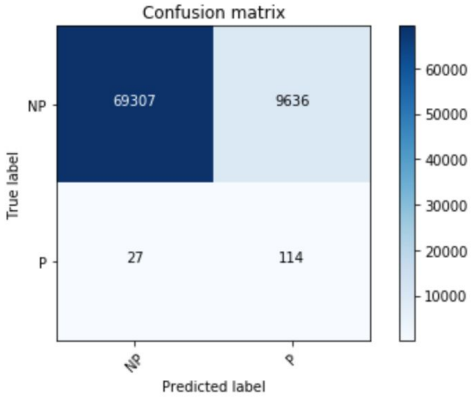
Model	**Accuracy score (Training / Testing)**
SVM	**0.72178**
Naive Bayes	**0.91773**
Decision Tree	**0.9928/0.9993**
Logistic Regression(P,NP)	**0.8765/0.8764**
Logistic Regression (7 categories)	**0.6513/0.6529**

We will mainly discuss the results of our logistic regression and decision tree classifiers below.

5.1.2 Logistic Regression and Decision Tree Classifier

Logistic Regression for P (heat pumps) and NP (not heat pumps) classes:

In order to develop the trained predictive models for the historical model, we split the data into training and testing. Since our main concern is the heat pumps, so we divide the categorical heat types into two main categories: P and NP. The Logistic Regression model based on P and NP results in an accuracy score of 87.65% on training data and 87.64% on testing data, Since the portion of class (P) category is too small, we implemented L1 regularization method to prevent overfitting and "balanced weighting" for each class in order to allow our model make more accurate prediction on the class (P). The resulting confusion matrix for class (P) and class (NP) is provided below with p value for each column.



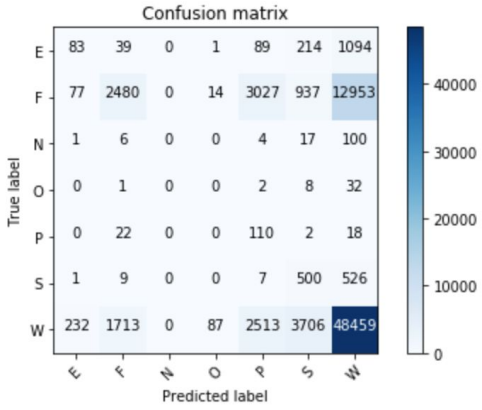
	Column name	P_Value
0	YR_BUILT	0.000000
1	LIVING_AREA	0.000000
2	LU_A	0.266247
3	LU_CD	0.274792
4	LU_E	0.000000
5	LU_EA	0.503454
6	LU_R1	0.000000
7	LU_R2	0.000000
8	LU_R3	0.000000
9	LU_RC	0.685136
10	STRUCTURE_CLASS_1	0.938555
11	STRUCTURE_CLASS_A	0.392594
12	STRUCTURE_CLASS_B	0.792321
13	STRUCTURE_CLASS_C	0.726990
14	STRUCTURE_CLASS_D	0.274845
15	STRUCTURE_CLASS_R	0.949766
16	OWN_OCC_N	0.000011
17	OWN_OCC_Y	0.003143
18	RorU_R	0.880392
19	RorU_U	0.275836
20	AC_C	0.000000
21	AC_D	0.000165
22	AC_E	0.964501
23	AC_F	0.949814
24	AC_N	0.000000

Figure 5.1 for Confusion Matrix and p_value (Logistic Regression on 2 classes)

Based on the confusion matrix above, our prediction model has slight improvement when it predicts the class (P) before we add weight and regularization arguments. However, it is still unable to have more classifications for the class (P). Therefore, we concluded that there is no enough information that our model could predict for the class (P).

Logistic Regression for 7 classes:

As for the prediction for on the heating system, our categories of electric (E), forced air (F), none (N), other (O), heat pumps (P), space heater (S), and hot water (W) results in an accuracy score of 65.13% on training dataset and 65.29% on testing dataset. One thing to note is that our prediction model is unable to differentiate the difference between class (F: Forced water) and class (W: hot water), more explanation for this reason can be found in the visualization part. The resulting confusion matrix for 7 classes is provided below with p value for each column.



	Column name	P_Value
0	YR_BUILT	0.000000
1	LIVING_AREA	0.000000
2	LU_A	0.000018
3	LU_CD	0.000000
4	LU_E	0.000000
5	LU_EA	0.000000
6	LU_R1	0.000000
7	LU_R2	0.000000
8	LU_R3	0.000000
9	LU_RC	0.017728
10	STRUCTURE_CLASS_1	0.000000
11	STRUCTURE_CLASS_A	0.000000
12	STRUCTURE_CLASS_B	0.655314
13	STRUCTURE_CLASS_C	0.000000
14	STRUCTURE_CLASS_D	0.000000
15	STRUCTURE_CLASS_R	0.064456
16	OWN_OCC_N	0.000000
17	OWN_OCC_Y	0.000000
18	RorU_R	0.031865
19	RorU_U	0.000000
20	AC_C	0.000000
21	AC_D	0.000000
22	AC_E	0.418555
23	AC_F	0.000000
24	AC_N	0.000000

Figure 5.2 for Confusion Matrix and p_value (Logistic Regression on 7 classes)

Based upon the confusion matrix above, after adjusting class weight to “balanced”, applying L1 regularization method and delete not statistically important features according to p-value make our model better now when predicting class (P). The accuracy has about 2% increase. Nevertheless, we still believe that there is not enough information that our prediction model could predict for the accurately class(P).

Decision Tree Classifier:

Our model predicts relatively accurate for the class (P) with a higher True Positive rate and False Negative rate. The results has an accuracy score of around 99.28% on the training dataset, and has an accuracy score of around 99.93% on the testing dataset. However, the dataset also shows that heating pumps are not very frequent in the heating system types, which says that heat pumps are too small of a dataset to make a conclusive prediction.

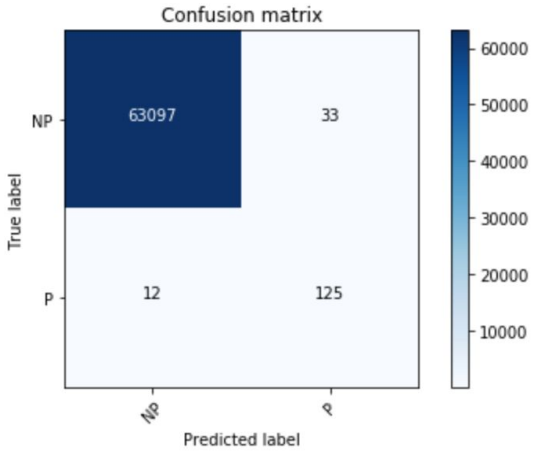


Figure 5.3 for Confusion Matrix (Decision Tree on 2 classes)

5.1.3 Map Visualization of Property Assessment Data

R_HEAT_TYP	Structure heat type:				
	E	Electric	O	Other	W Hot Water
	F	Forced Air	P	Heat Pump	
	N	None	S	Space Heater	

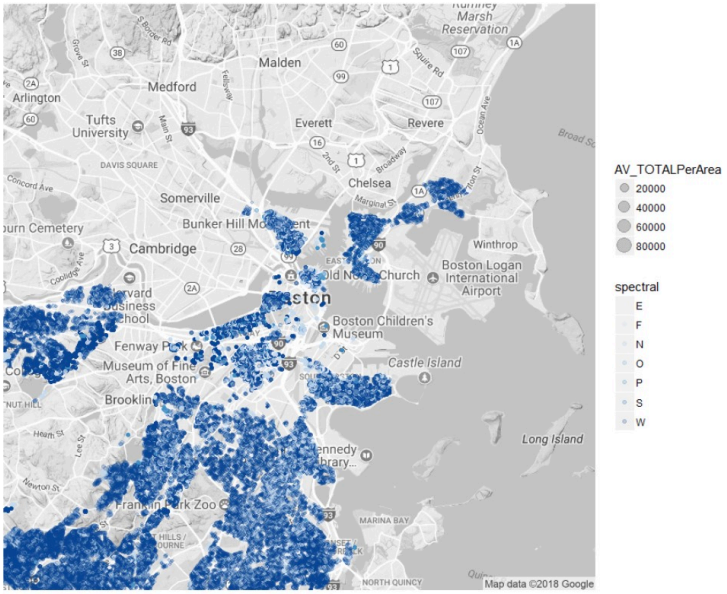


Figure 5.4 Type of Heating Systems Distribution by Average Assessed Value per sq ft

Given the geographical location(longitude, latitude), and whether the house is owner occupied, plus the type of heating system, we graph the distribution of heating system among the Boston demographic communities. The size of data points represents the average assessed value per square feet (sq ft). In Figure 5.4, the data points of heating systems (hot water and forced air) are spread around the urban areas of Boston, which are occupied by owner. Typically, the average price per sq ft is higher near Charles River Area, and area near Franklin Park Zoo(located near the bottom left of map), implied by the deeper blue area. However, due to the constraint of color, to visualize better of the heating system, we only select forced air and hot water heat type and make Figure 5.5, which does not change the overall spread. The light blue in Figure 5.5 is hot water and the white is forced air. The majority of heating system is hot water.

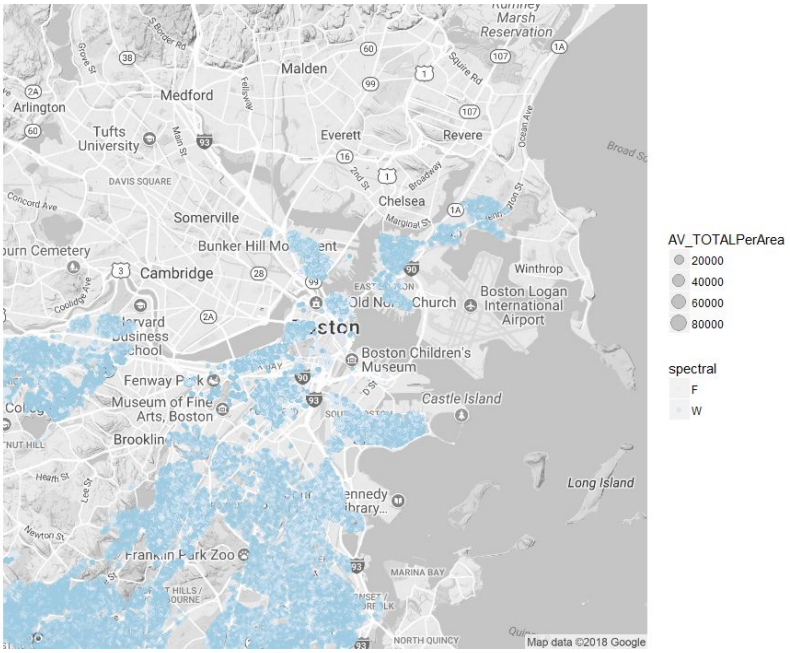


Figure 5.5 Forced Air and Hot Water Heating System Distribution by Average Assessed Value per sq ft

Since Michael is particularly interested in the spread of heat pumps as they are the most energy efficient, we create a subset of the heat pumps system data. It's worth to note that because of the rare quantity of heat pumps in cleaned dataset (approximately 600 points out of 300,000 data points in cleaned parcel data), so we only select non-empty geolocation (longitude and latitude) for heat pumps in original Boston Property Assessment data for 2014 and 2018, respectively.

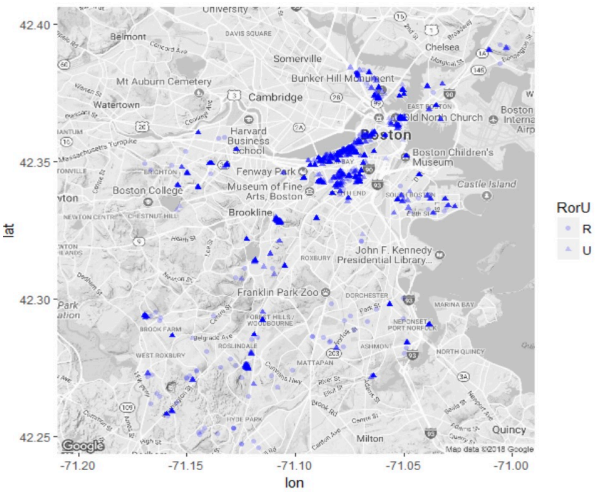


Figure 5.6 Heat Pumps Distribution for 2014 by Residential House(Circle) and Condo Unit(Triangle)

Figure 5.7 Heat Pumps Distribution for 2018 by Residential House(Circle) and Condo Unit(Triangle)

By comparing Figure 5.6 and 5.7, there is approximately no change in heat pumps installed in Boston area (There are 4223 data in 2014 and 4217 in 2018). It's worth to note that there is a large cluster of heat pumps installed near Charles River Area, which might be geothermal (water source) heat pumps that help transfer heat between the residence and a nearby water source.

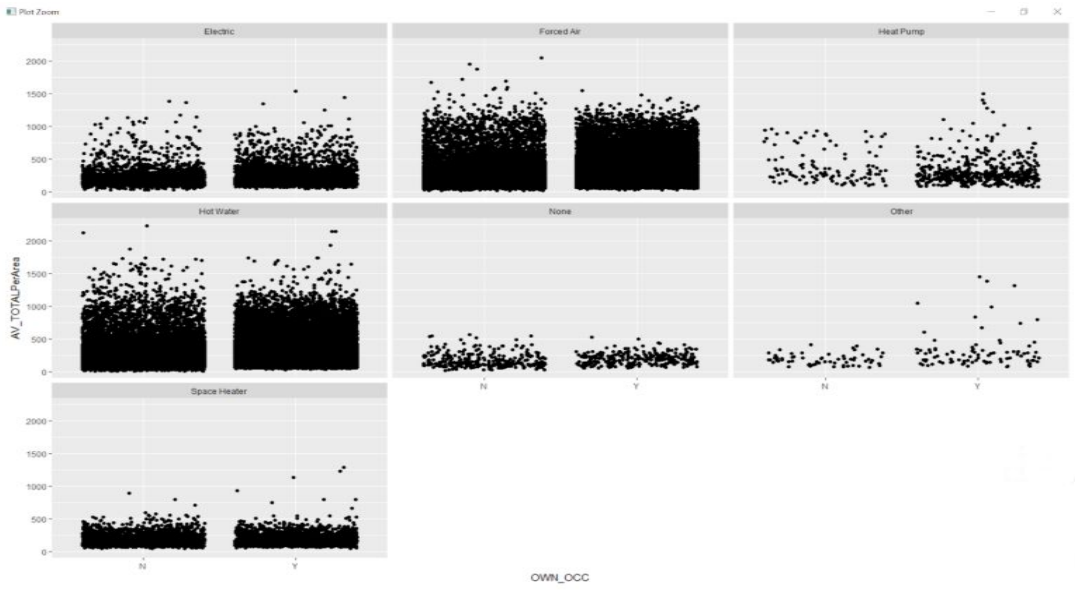


Figure 5.8 Owner Occupancy(x:N/Y), Average Total Assessed Value per sqft(y) and Heat Type(facet)

In Figure 5.8, we still use the clean dataset to compare the type of heating systems under assessed value per sq ft and owner occupancy. We use jitter plot to add a small amount of random variation to the location of each point to handle overlapping in traditional plots caused by discreteness in datasets. The right column in every facet suggests the house is owner occupied, and the left column implies that the house is rented. There are clearly more forced air and hot water among all heating systems, with higher assessed value, equally distributed among N/Y for owner occupancy. Given that factor, a common guess would be: Since the years where all recorded buildings are built span from 640 and 2016, with median built in 1910, it is possible that those ancient buildings may use traditional forced air or hot water heating system, instead of recently innovated heat pumps. To investigate this guess, we have to look at YR_BUILT(Years when residence built) factor to determine the difference of heating systems for various building years.

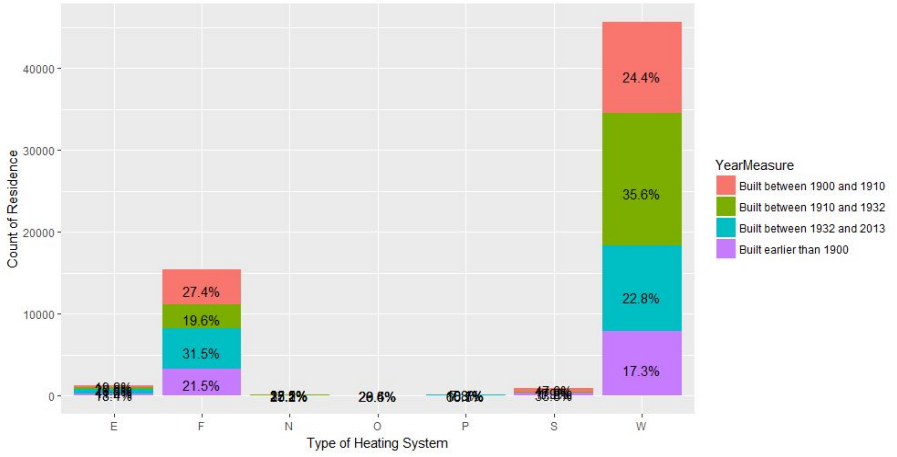


Figure 5.9 Residence distribution Grouped by Type of Heating System and Years Built in 2014 Boston Property Assessment Data

Figure 5.9 validates our assumption: Forced Air and Hot Water bins contain 70% - 80% residence built before 1932, while only 20% and 30% of building use above two heating systems were built between 1932 and 2013. To look clearly what other types of heating system besides forced air and hot water, we rescale the graph with five other types heating system (Figure 5.10). For all installed heat pumps in residence, 66.7% of them are built between 1932 and 2013, which are significant(large) compared to the percentage of buildings built between 1932 and 2013, using forced air and hot water heating system.

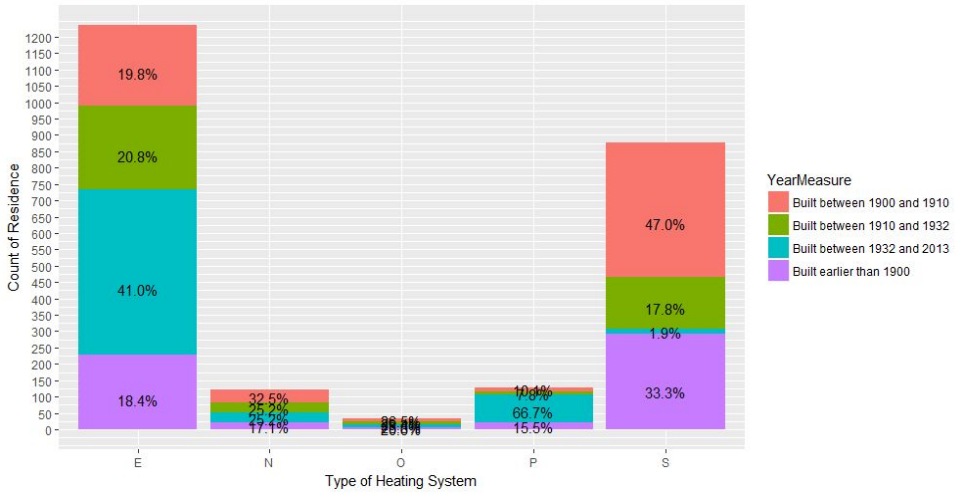


Figure 5.10 Residence Distribution by Type of Heating System and Years Built in 2014 Boston Property Assessment Data Except Forced Air and Hot Water

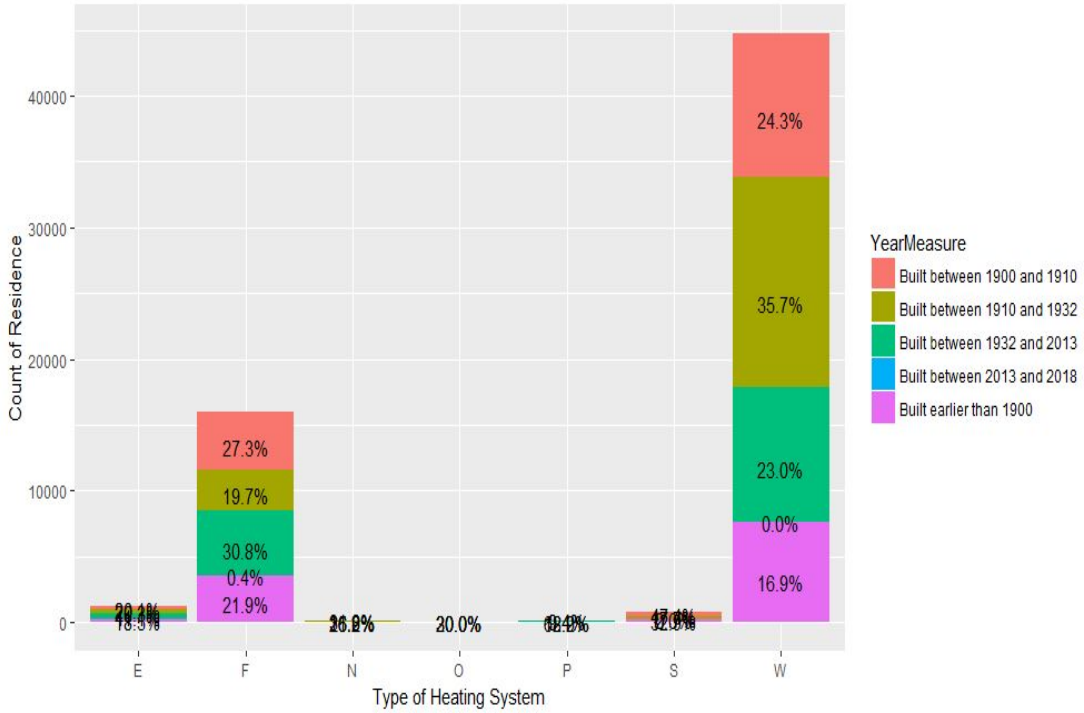


Figure 5.11 Residence Distribution by Type of Heating System and Years Built in 2018 Boston Property Assessment Data

To compare the trend of 2018 with 2014 in our five years dataset, we create Figure 5.11 to determine the residence distribution of heating systems and years built in 2018. There is slightly change in the number of forced air and hot water heating system. To visualize it better, we filter the data with only buildings that are constructed between 2013 and 2018 in Figure 5.12

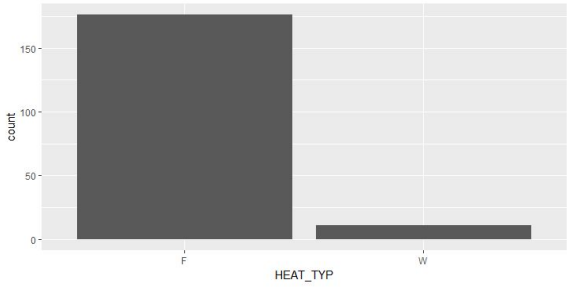


Figure 5.12 Residence Built between 2013 and 2018 by Type of Heating System

To sum up, the general distribution of heating systems does not change within 5 years period. The majority of heating systems are still forced air and hot water with average assessed value equally distributed up to 1250 dollars per sq ft (Figure 5.8). As we all know, Boston is an old city where there are a large number of ancient buildings that use these two traditional heat systems. Before 1932, these two types of heating systems are mainstream. But, after 1932, some of residence resort to heat pumps, although still much less than residence using forced air and hot

water heating systems. Besides, the difference between hot water and forced air heating systems installation is very hard to distinguish: the average assessed value per sq ft, the geographical location, the total number of installation, whether installed in residence or condo unit are very much similar for two systems. This could explain why our prediction model(logistic regression and classification tree) tend to misclassify forced air and hot water.

By talking to Michael, we understand that replacing conventional building boilers with heat pumps in city buildings could substantially increase the viability of renewable energy use in the city, according to a recent study from researchers from the Earth Institute's Sustainable Engineering Lab. But this theory has some limitations in Boston in practice.

As shown in Figure 5.6 and 5.7, the residence using heat pumps are centered on downtown area, part of reason would be in downtown area. Part of reasons would be that heat pumps tend to be somewhat ineffective in any climate where the outdoor air temperature falls near or below freezing on a regular basis. Heat pumps move move heat from outdoors to indoor in winter, and move heat from indoors to outdoors in summer. However, during winter time, in Boston rural area, where temperature always falls below freezing point, it would be very energy inefficient to move heat from a very cold area to a hotter one compared to moving heat between two areas with a more moderate temperature difference. In Boston urban area, under the effect of urban heat island[1] effect, the temperature in city of Boston would be much higher than rural area, thus residence in city of Boston tend to choose heat pumps. There's also more heat available outside in a moderate climate than in a cold climate. It's important to note that even in a cold climate, there's still heat in the outside air to be pumped indoors, but the unit needs to work harder to extract the heat that's available.

Another reason would be the heat produced by heat pumps isn't as intense as the heat produced by forced air and hot water boiling heat systems. People who are used to traditional furnaces can be uncomfortable with the milder heat produced by these systems. So maybe under the severe weather condition, it would be more energy efficient to stick to traditional heating systems, unless we could solve the problem of energy efficiency.

5.2 Heat Type Rate of Change

After talking to our project partners Michael and Adam, we focused on heat type of residential building in all structures (including buildings for agricultural, Tax-exempt, and commercial use in Type of Property column) . In Property Assessment Data (Parcel Data), the overall residential heat types are separated into R_HEAT_TYP (residential heat type) and U_HEAT_TYP(condo heat type, recorded based on each unit of condo), and the overall trend of heat types for condo and residential building per year.

[1] An **urban heat island** (**UHI**) is an urban area or metropolitan area that is significantly warmer than its surrounding rural areas due to human activities.

Residential Heating Type

Residential Heat Type	2014-2015	2015-2016	2016-2017	2017-2018	2014-2018
Space Heater	-1.46%	-2.50%	-5.02%	-3.44%	-11.88%
Hot Water	-0.13%	-0.67%	-0.43%	-0.55%	-1.77%
Electric	-0.49%	-1.15%	-1.49%	-1.85%	-4.89%
Heat Pump	-3.08%	-0.79%	0.00%	1.60%	-1.31%
Forced Air	1.54%	0.89%	1.91%	1.03%	5.48%
Total Unique Parcel ID	1.17%	0.64%	1.03%	1.16%	4.06%

Condominium Heating Type

Condo Heat type	2014-2015	2015-2016	2016-2017	2017-2018	2014-2018
Space Heater	-0.97%	-0.98%	-0.99%	-4.00%	-6.80%
Hot Water	-1.18%	-1.12%	-0.75%	-0.49%	-3.49%
Electric	-0.97%	-0.08%	-0.25%	-0.23%	-1.53%
Heat Pump	-0.27%	0.32%	0.34%	-0.58%	-0.20%
Forced Air	5.14%	3.24%	5.22%	6.95%	22.15%
Total Unique Parcel ID	1.17%	0.64%	1.03%	1.16%	4.06%

Figure 5.13 Heat Type Rate of Change

We used the Boston Property Assessment dataset and calculate the rate of change based on the formula:

Rate of change for heat type i= (Number of heat type i in Year 2 - Number of one heat i in Year 1)/(Number of heat type i in Year 1) × 100%

Rate of change for unique Parcel ID = (Number of unique Parcel ID in Year 2 - Number of unique Parcel ID in Year 1)/(Number of unique Parcel ID in Year 1) × 100%

Forced air heating system has a significant increasing rate for both residential and condo unit, 5.48% and 22.15%. As indicated in section 5.1.3, the majority of the buildings constructed between 2013 and 2018 use forced air heating systems. The rate of change table is consistent with plots in map visualization section.

5.3 Histogram of Solar Panels

Under the guidance of Michael, we use stack bar chart to visualize the amount of kilowatts of solar panel energy used by zip code. We are interested in knowing if there is a relationship between the occupancy type of the household with respect to the capacity of solar panel energy used in each household.

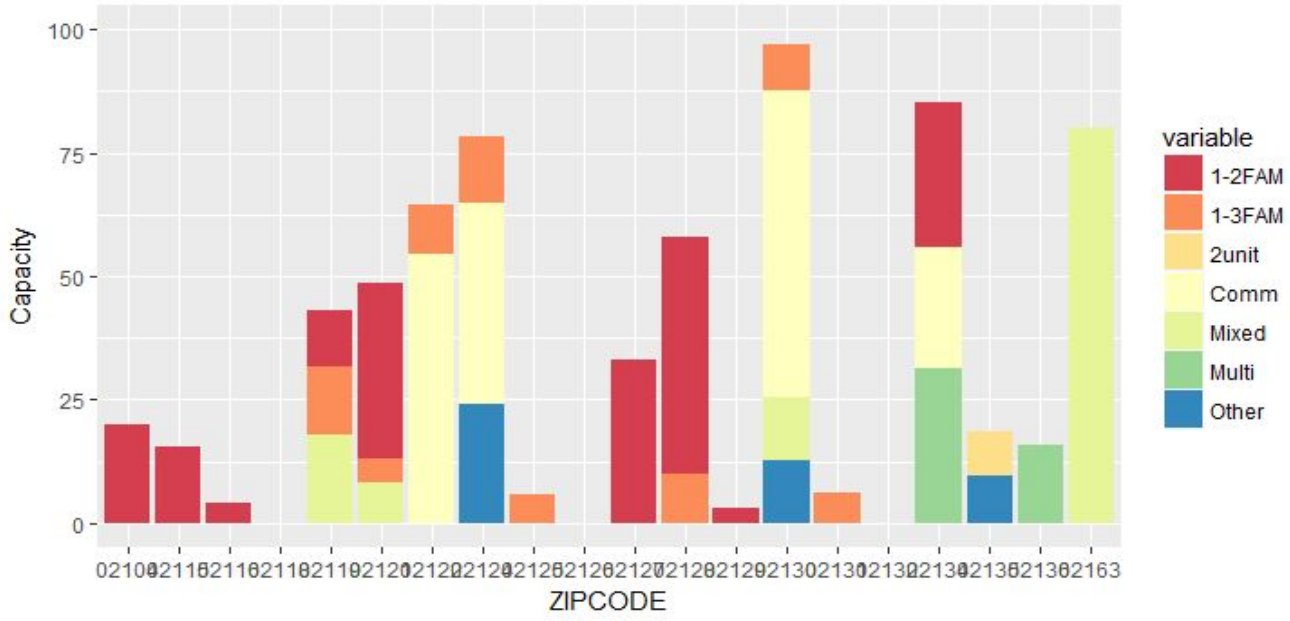


Figure 5.14 Solar Panel Usage Stack Bar Histogram

We are particularly interested in the numbers before the 'kW' (kilowatt) in the comments column of the Boston Permits dataset in order to estimate solar capacity usage per household. It's worth to note that this plot might have "selection bias": The permit dataset only measures the renovation required in 2018, and not all houses will report solar capacity in comments column(in fact, 638 out of 300,000 rows). Additionally, Zip code might not be a good measure as area within same Zip code might have various housing price level.

The histogram has the x-axis of the Zip code of the permit buildings (range from 02104 to 02163), ordered in ascending assessed value per sq ft order, and has the y-axis of the summation of solar capacity usage (in kilowatts) per Zip code. The histogram is organized in a stacked bar format, with each of the building occupancy type (1-2 family, 1-3 family, 2 unit, community, mixed use, multi family, and other) designated with a distinct color. In terms of interesting findings, we found out that there are only 638 solar panels installed in the entire zip code range of 02104 to 02163, from Boston Permit Dataset. This particular finding is beneficial for BU ISE to know, since solar panel energy, although very efficient, is a fairly expensive means of making Boston more sustainable, since many of the households may not be able to afford it.

Because of shading, insufficient space and ownership issues, 1/5 Boston homes are simply unfit for solar panels. The highest bin's zip code in the above graph is 02130. The area in 02130 is

Arnold Arboretum of Harvard University, where each residence has sufficient space and the housing price is relatively high that most family could afford solar panels. However, with the introduction of shared solar, homeowners can subscribe to “community solar gardens”, and generate solar electricity without actually having solar panels on their own rooftops. There is a large percentage of community, mixed, multi, and other type of residence using solar panels, instead of 1-2, 1-3, 2 unit family, which might share solar panels together. If we have more data sources, such as race, income level, type of solar panels in every residence, this finding could be better explained.

The histogram is important because it shows that capacity for area in similar geolocation and each individual type of occupancy, which we will able to explain why larger dwellings require more solar capacity to generate their energy.

6. Conclusion and Summary

After analyzing the rate of change of each type of heating system per year, we found that our models (Logistic Regression, Naive Gaussian Bayes, SVM, Decision Tree) are not able to get enough information for our interested class P.

From our data, we can also calculate the heat pump usage rate of change for each housing neighborhood. Over the span of the past five fiscal years, we noticed that the usage rate of the heat pumps have been steady for the condominium housing, but with great variation for the residential housing. We can also calculate the solar capacity usage of each household with a stacked bar histogram, which shows both the averaged assessed value per zip code and land prices increasing. The decision tree shows that the majority of the Boston communities with heat pumps are mostly the 1-Residential, 2-Residential, and 3-Residential family occupancy type.

For visualization part, the trend of heating systems does not change much from 2014-2018, and the majority is still forced air and hot water heating system. The reason why heat pumps are not popular in Boston is severely cold weather. It would be energy inefficient to use heat pumps in severely cold (below freezing point) weather condition to extract heat from outdoors, and people may not get used to the comparatively moderate heat generated by heat pumps. However, there are relatively large number of heat pumps in city of Boston, particularly near Charles River, compared to suburbs, due to the higher temperature of urban heat island effect and it would be much easier to transfer heat from water to residence (Geothermal heat pumps).

There is very small distinction between forced air and hot water installation situation. The average assessed value per sq ft, the geographical location, the total number of installation, whether installed in residence or condo unit are very much similar for two systems. This could explain why our prediction model(logistic regression and classification tree) tend to misclassify forced air and hot water.

For energy efficient system, particularly solar energy, from our graph (Figure 5.14), the area where average assessed per sq ft is high and sufficient large space is available for solar panels installation, has higher use of solar panels. An example would be Arnold Arboretum of Harvard University. But there is still ways to benefit from solar energy in lower average assessed value

community. Those residence, where mainly consists of 1-2 or 1-3 family or 2 units, tend to cluster together and use solar energy collectively. With the introduction of shared solar, homeowners can subscribe to “community solar gardens”, and generate solar electricity without actually having solar panels on their own rooftops.

7. Intended Follow-Up

While we are able to plot the solar panel usage and measure the heat pump rate of change for both the Boston Building Permit and Boston Property Assessment datasets with great accuracy, we believe there is potential for follow-up future work. Steps could include more text processing for the permit data, using the American Census Housing Survey to get a detailed inventory of each building parcel in Boston, and also a larger time interval (defined as the past decade) for determining the usage of solar panels and the heat pump rate of change.

The target of our project investigation lies in heat pumps in heating systems. However, the heat pumps percentage in Boston Property Assessment and Permit dataset is in very small percentage (less than 0.1%), which poses a significant challenge for our model prediction and classification. Even though we attempted to adjust the class weight and regularization method, the model would still cannot provide enough information for class P. If we could have more dataset regarding heat pumps or other feature vectors(such as race, income level, housing price) we may be able to have more classification on heat pumps system.

There are also a large number of forced air and hot water systems. If we have more time, we would like to talk to people in residence using these two types of heating system and ask them the practical benefit of them compared to heat pumps. We will talk to Michael and ask more pros and cons of other heating systems.

For energy efficient systems, the zip code might not be a good measure of housing price level in a specific area, as Professor Smith told us. If we have more time, we would like to use Zillow as a measure of housing level, and divide communities based on housing price range, instead of zip code.

BU ISE will be able to determine with greater accuracy which type of housing will most benefit from a new energy system. The research group will be able to determine, given the type of house, income, and other identified characteristics which heating system each building has currently. For the Boston community, it would make the surrounding area more sustainable in terms of energy-efficiency. The results from the analysis of our model would set a precedent for a cleaner Boston in the future.

8 Link to Github Repo (Private Repo)

https://github.com/jasonlu6/CS-506-Final-Project-BU-ISE
Please contact Jason (jasonlu6@bu.edu) in order to check our Python Notebooks and graphs.

9 Appendix

Type of Property data key table

LU	Type of Property (land use)			
	A	Residential 7 or more units	E	Tax-exempt
	AH	Agricultural/Horticultural	EA	Tax-exempt (121A)
	C	Commercial	I	Industrial
	CC	Commercial condominium	R1	Residential 1-family
	CD	Residential condominium unit	R2	Residential 2-family
	CL	Commercial land	R3	Residential 3-family
	CM	Condominium main (physical structure housing all related condo units with no assessed value)	R4	Residential 4 or more family
			RC	Mixed use (res. and comm.)
	CP	Condo parking	RL	Residential land

Heat Type

R_HEAT_TYP	Structure heat type:				
	E	Electric	O	Other	W Hot Water
	F	Forced Air	P	Heat Pump	
	N	None	S	Space Heater	