CS 506 Interim Report

CS 506
Spring 2018
Computational Tools for Data Science
Professor Adam Smith
TF: Sofia Nikolakaki
3/12/2018 - 3/20/2018

Group:

1) **Jason Lu (jasonlu6@bu.edu)**
2) **Zixin "Cindy" Ding (cindydzx@bu.edu)**
3) **Fang Qu (hughqu@bu.edu)**

Collaboration with BU Spark!

Coordinators:
Ziba Cranmer (zcranmer@bu.edu)
Professor Dora Erdos (edori@bu.edu)
Professor Adam Smith (ad22smith@bu.edu)

Contact: Michael Jay Walsh, Ph. D.

Topic: Using Permit, Parcel, Census, data to populate building energy systems inventory

Section I) Investigative Procedures:

Question to Investigate:

Using the parcel and permit databases, how well correlated is the data for the varying amount of people having heating systems as opposed to having other energy systems? For example, is there any significant effect if we were to cluster with a varying amount of family members per household, while keeping the feature vector of longitude, latitude, and comments constant?

Hypothesis:

We hypothesize that the data will be fairly well correlated for the discrepancy between the heating systems and the other energy systems. There is likely a significant difference with respect to the assessor data from 2018 fiscal year and the 2017 fiscal year, which would be advantageous for a time series analysis.

Section II) Date Source Chosen / Prepared:

Data Sources:

- **buildingpermits.csv - building permit database**
- **masterdb.csv - master database (created by our own)**
- **ast2018full.csv - assessor property value database for 2018 fiscal year**

Parsing procedure:

For parsing the database, we decided to merge the building permits database and the assessor database together to form the master database. Then, we used some basic parsing techniques (appending the data from the building permits database and the assessor database into a master list) in order to extract the most important data for the analysis. For the permit database, we extracted only the parcel ID, comments, and location (separated by the longitude and latitude). For the assessor database, we extracted many more attributes, given that we are required to connect the two databases together. Those attributes that were extracted were from the parcel id (concatenated with building permit database), zip code (for demographics and location), land use, average total valuation (housing and property cost), living area, heating systems, r_ac (air conditioning and cooling systems), and comments as well. Finally, we combined the building permits and assessor database via a python concatenation function (attempted the pandas Dataframe merge and join function, but could not find a common attribute to join the two databases) to create the master database.

Creation of the Master Database (for further analysis):

For the Master Database, we first splitted longitude and latitude into 2 separate columns by using string separation method. Then we concatenated two datasets from approved building permits and property assessment by parcel_id.

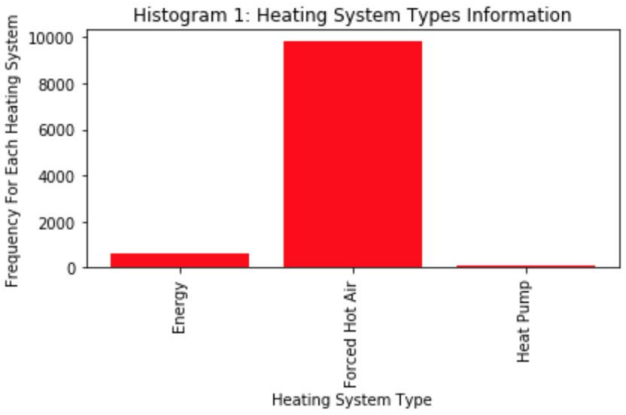
Getting the Top 10 Most Frequent Words in the Comments (TFIDF Vectorization):

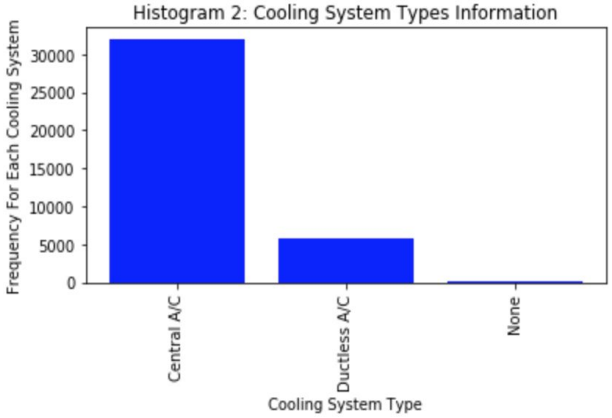
For getting the top 10 most frequent words in the comments, we use the TFIDF (term frequency). We then got a sample of 10,000 terms from the term list corpus, by displaying the shape of the v-transpose vector in the Singular Value Decomposition function. We then used the approach described in lecture (`np.argsort()`) to sort the terms in the corpus by the most frequent in the comments, and then extracted the most frequent words from the sorted corpus (using the terms that Michael specified for us for the BU Spark! Project).

Section III) Initial Exploratory Analysis:

Histograms:

Histogram for the Heating Systems:





Data information:

Histogram 1:

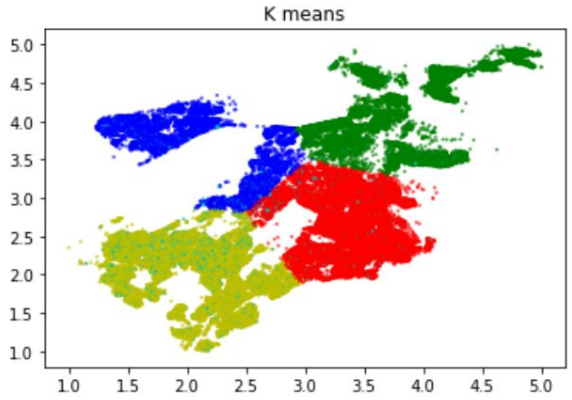
Type of Heating	Frequency
Energy	618 households
Forced Hot Air	9836 households
Heat Pump	69 households

Histogram 2:

Types of Cooling	Frequency
Central A/C	31905 households
Ductless A/C	5560 households
None	167 households

Based on the histograms above, we can know that the main cooling systems that people use in Boston area is definitely Central A/C with a small portion of home using ductless A/C. As for the heating system, most home are likely to choose Forced Hot Air with very small amount of homes using Energy and Heat pumps. More analysis will be needed to compare year by year heating and cooling systems, probably with a metric for comparison.

Clustering Plots:



Rate of Change:

After talking to Michael on March 16th, he emphasized the importance of seeing the trends of different heat types, especially "heat pumps",(decoded as "P" in column"R_HEAT_TYP" in Boston Parcel Data(from 2014-2018). Then we compared the rate of change for 5 major types and constructed a table below, displaying the rates of change in terms of percentages.

Heat Type	2014-2015	2015-2016	2016-2017	2017-2018
Space Water	-1.46%	-2.50%	-5.02%	-3.44%
Hot Water	-0.13%	-0.67%	-0.43%	-0.55%
Electric	-0.49%	-1.15%	-1.49%	-1.85%
Heat Pump	-3.08%	-0.79%	0.00%	1.60%
Forced Air	1.54%	0.89%	1.91%	1.03%
Overall Parcel ID	1.17%	0.64%	1.03%	1.16%

Section IV) Plans / Next Steps:

Solar Panel Text Extraction: Our next step in the project is to use regular expressions and the `fuzzywuzzy()` Python library import to do some string manipulation to get the solar panel capacity. We can then do a regression testing on the solar panel capacity, to determine the rate of change within the years.

Cross-Validation of the Energy Systems: We will also use the approach of confusion matrices / cross validation of our energy systems later on in the project, as a way to check if our hypothesis matches up to the evidence provided in the initial analysis. Collaboration with Adam Pollack and Michael Walsh.

For comments column, we will also assign comments into a number of groups, then make each comment a small point, and plot it in graph, instead of direct clustering.