Milestone 5

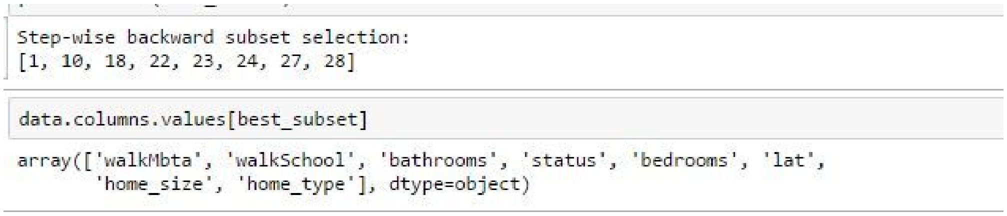
Introduction

In milestone 4 , we explored a baseline model where the housing price in the city of boston is predicted using ridge regression with cross validation methodology. We obtained data from both zillow and trulia. Zillow data contains housing price with useful house features such as hous types, the year that the building is built, the property size, longitude and latitude. Trulia data that we obtained are renting prices with similar features . We furthermore gathered urban environment data such as closeness to transportation and school facilities, the image recognition of urban environment, and crime data. These features are listed in the appendix for more detailed elaboration. The data processing and merging these data according to their geocodes is an intensive process which we won't describe. Here , we will elaborate some finding from the preliminary models .

Findings

Model for sale prices obtained from zillow:

Zillow listings does not contains a ton of houses with sale prices. In total we had 724 cases. The features number is about 30. Among the 30 features, we used backward,forward, and exhaustive selection(BIC criteria) to see which ones are more useful . And the result is rather intuitive , both backward and forward selection selected that walking distance to T-stop, walking distance to school, bathrooms number, bedroom numbers, latitudes, home size, home types are the significant features. Which we thought is interesting that room number and bathroom number and home size are all important as one would think they would be redundant variables as they code repeated information . None of the urban features appear important in this exploration stage.



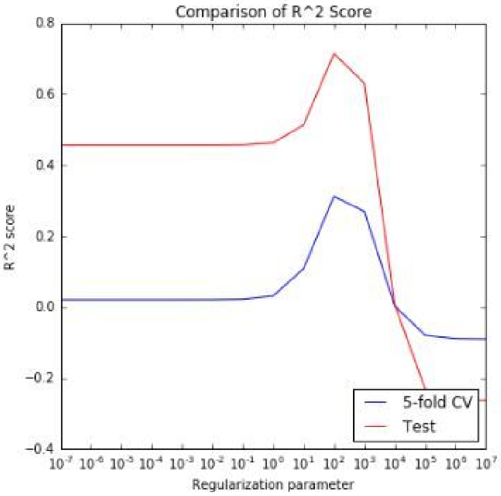
```
Step-wise backward subset selection:
[1, 10, 18, 22, 23, 24, 27, 28]
```

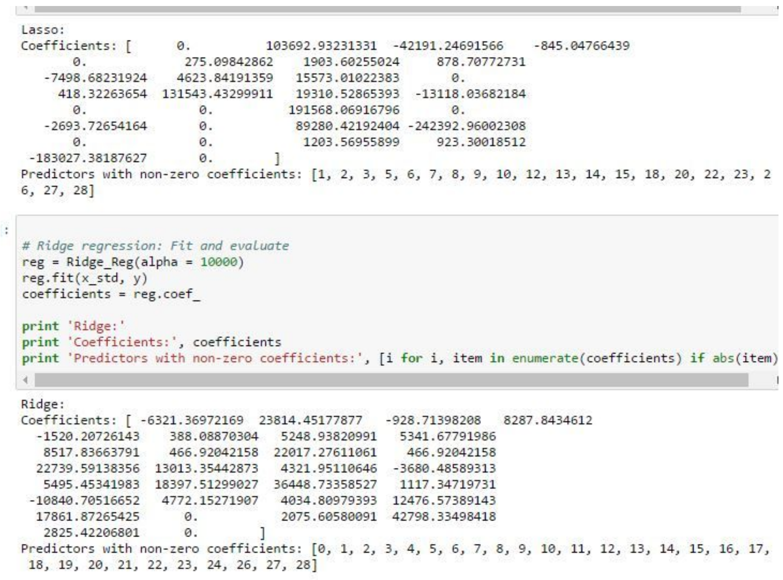
```
data.columns.values[best_subset]
```

```
array(['walkMbta', 'walkSchool', 'bathrooms', 'status', 'bedrooms', 'lat',
       'home_size', 'home_type'], dtype=object)
```

Furthermore, we implemented simple linear regression with all the features in the model, and the r square for the training data is about 0.6 the r square for test data is about 0.45, which is not so idea. We further, implemented Ridge regression and Lasso to see their coefficient choices , and implemented k fold cross validation . Which gave a better result for R square, but we still need to interpret the model as the coefficients are relatively large. But using the tuning parameter is able to give as a model with r square around 0.7. This comes with a lot more features included in the model compare to the forward and backward selection result.

```
Ridge regression: Test R^2 score for CV choice 0.714159829286
Ridge regression: Max Test R^2 score 0.714159829286
Plain regression: Test R^2 score: 0.456852277299
```





```
Lasso:
Coefficients: [      0.          103692.93231331  -42191.24691566    -845.04766439
       0.            275.09842862    1903.60255024     878.70772731
   -7498.68231924    4623.84191359   15573.01022383       0.
     418.32263654  131543.43299911   19310.52865393  -13118.03682184
       0.              0.          191568.06916796       0.
   -2693.72654164      0.           89280.42192404 -242392.96002308
       0.              0.            1203.56955899     923.30018512
 -183027.38187627      0.        ]
Predictors with non-zero coefficients: [1, 2, 3, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 18, 20, 22, 23, 2
6, 27, 28]
```

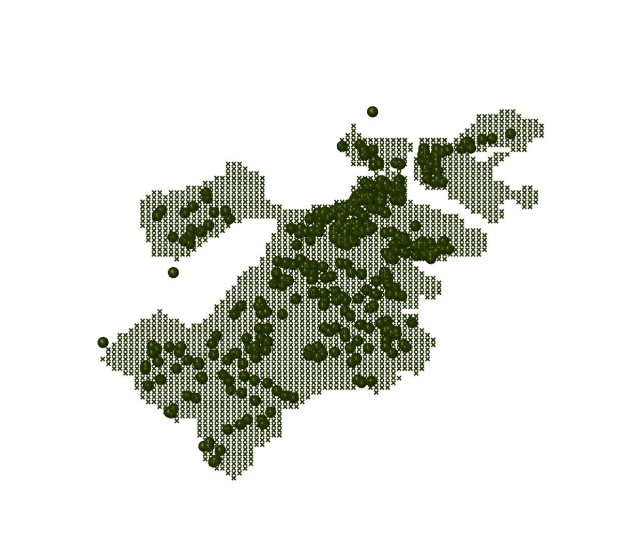
```
# Ridge regression: Fit and evaluate
reg = Ridge_Reg(alpha = 10000)
reg.fit(x_std, y)
coefficients = reg.coef_

print 'Ridge:'
print 'Coefficients:', coefficients
print 'Predictors with non-zero coefficients:', [i for i, item in enumerate(coefficients) if abs(item)
```

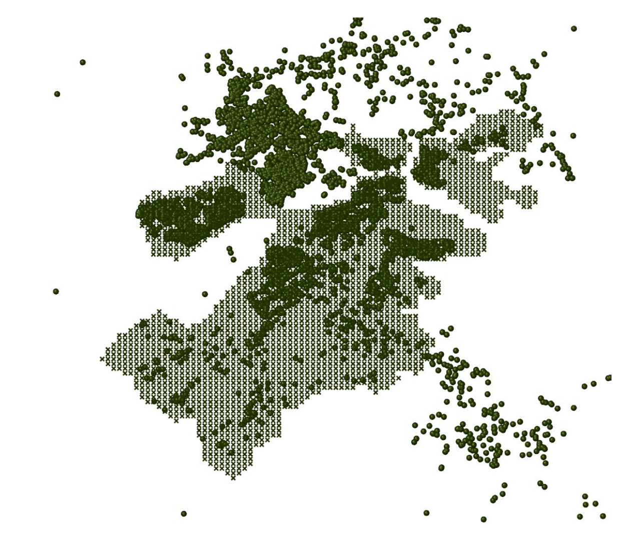
```
Ridge:
Coefficients: [ -6321.36972169  23814.45177877   -928.71398208   8287.8434612
  -1520.20726143    388.08870304   5248.93820991   5341.67791986
   8517.83663791    466.92042158  22017.27611061    466.92042158
  22739.59138356  13013.35442873   4321.95110646  -3680.48589313
   5495.45341983  18397.51299027  36448.73358527   1117.34719731
 -10840.70516652   4772.15271907   4034.80979393  12476.57389143
  17861.87265425      0.           2075.60580091  42798.33498418
   2825.42206801      0.        ]
Predictors with non-zero coefficients: [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,
 18, 19, 20, 21, 22, 23, 24, 26, 27, 28]
```

Interestingly , in ridge regression, most urban environment features have large coefficients , some of them had positive effects : pixel of sky, pixel of late, number of crime, pixel of person, pixel of car, pixel of grass, pixel of sidewalk. And some of them had negative effects: pixel of building, and pixel of house , pixel of roads, and number of craigslistings .
This shows preliminary results of how urban environment contribute to land price, although weirdly crime coefficient is positive. We will need to further investigate this.

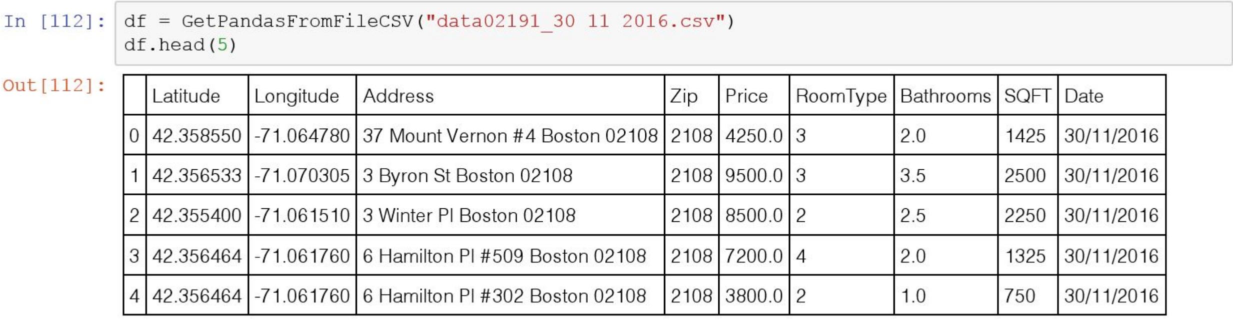
Model for rent prices obtained from trulia:



Zillow Housing Data



Trulia Rent Data

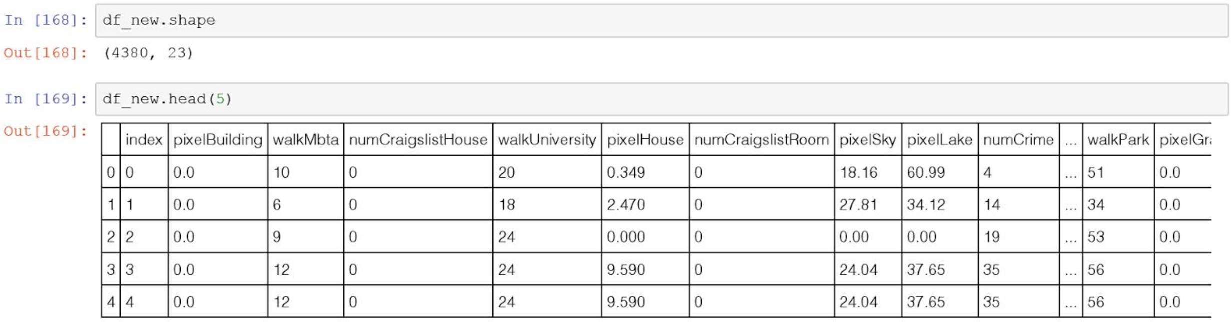


```
In [112]: df = GetPandasFromFileCSV("data02191_30 11 2016.csv")
          df.head(5)
```

Out[112]:

	Latitude	Longitude	Address	Zip	Price	RoomType	Bathrooms	SQFT	Date
0	42.358550	-71.064780	37 Mount Vernon #4 Boston 02108	2108	4250.0	3	2.0	1425	30/11/2016
1	42.356533	-71.070305	3 Byron St Boston 02108	2108	9500.0	3	3.5	2500	30/11/2016
2	42.355400	-71.061510	3 Winter Pl Boston 02108	2108	8500.0	2	2.5	2250	30/11/2016
3	42.356464	-71.061760	6 Hamilton Pl #509 Boston 02108	2108	7200.0	4	2.0	1325	30/11/2016
4	42.356464	-71.061760	6 Hamilton Pl #302 Boston 02108	2108	3800.0	2	1.0	750	30/11/2016

Total number of data we got from the Zillow housing data is relatively small, thus we looked at the rent data from Trulia. 13049 data points with 9 features can be obtained after scrapping Trulia website, which contain location, price, room type, bathrooms, and total areas. We used the location data, total areas and the number of bathrooms to try to find the correlation. However, it is difficult to get the result with unrefined datasets from web. Therefore, we mapped the data we got from google street views, which consist of multiple different parameters such as walkability to schools and MBTA.



```
In [168]: df_new.shape
Out[168]: (4380, 23)
```

```
In [169]: df_new.head(5)
Out[169]:
```

	index	pixelBuilding	walkMbta	numCraigslistHouse	walkUniversity	pixelHouse	numCraigslistRoom	pixelSky	pixelLake	numCrime	...	walkPark	pixelGr
0	0	0.0	10	0	20	0.349	0	18.16	60.99	4	...	51	0.0
1	1	0.0	6	0	18	2.470	0	27.81	34.12	14	...	34	0.0
2	2	0.0	9	0	24	0.000	0	0.00	0.00	19	...	53	0.0
3	3	0.0	12	0	24	9.590	0	24.04	37.65	35	...	56	0.0
4	4	0.0	12	0	24	9.590	0	24.04	37.65	35	...	56	0.0

What to do next:

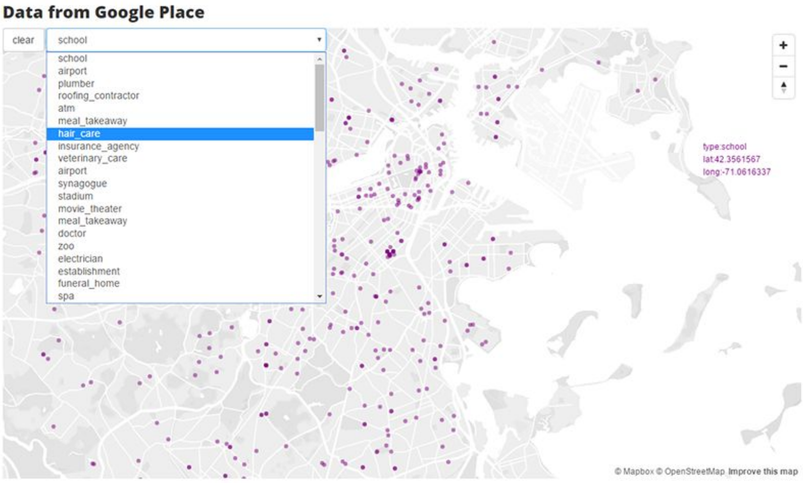
For the sale price model , the coefficients are very large which make things a little bit difficult to interpret. In the initial model, we have dropped some missing values because there are not many of them , but to improve we can use knn method to generate values such as built year using their geo location code. Visualization the results will help us understand what the model means more. We would implement 3d visualization of location, price,and the useful features to interpret the story better. We will also compare the housing price trend with the rent price trend in the city of boston.

Feature Appendix:

link for data visualization

There are two main data sets: 1) Housing Price from Zillow and 2) Rent Price from Trulia.
In term of features, each data set has its own data and pixel data that we produce from Google Place, Google Street View and craigslist. Among many features we extract, we pick the key factor

Google Place:

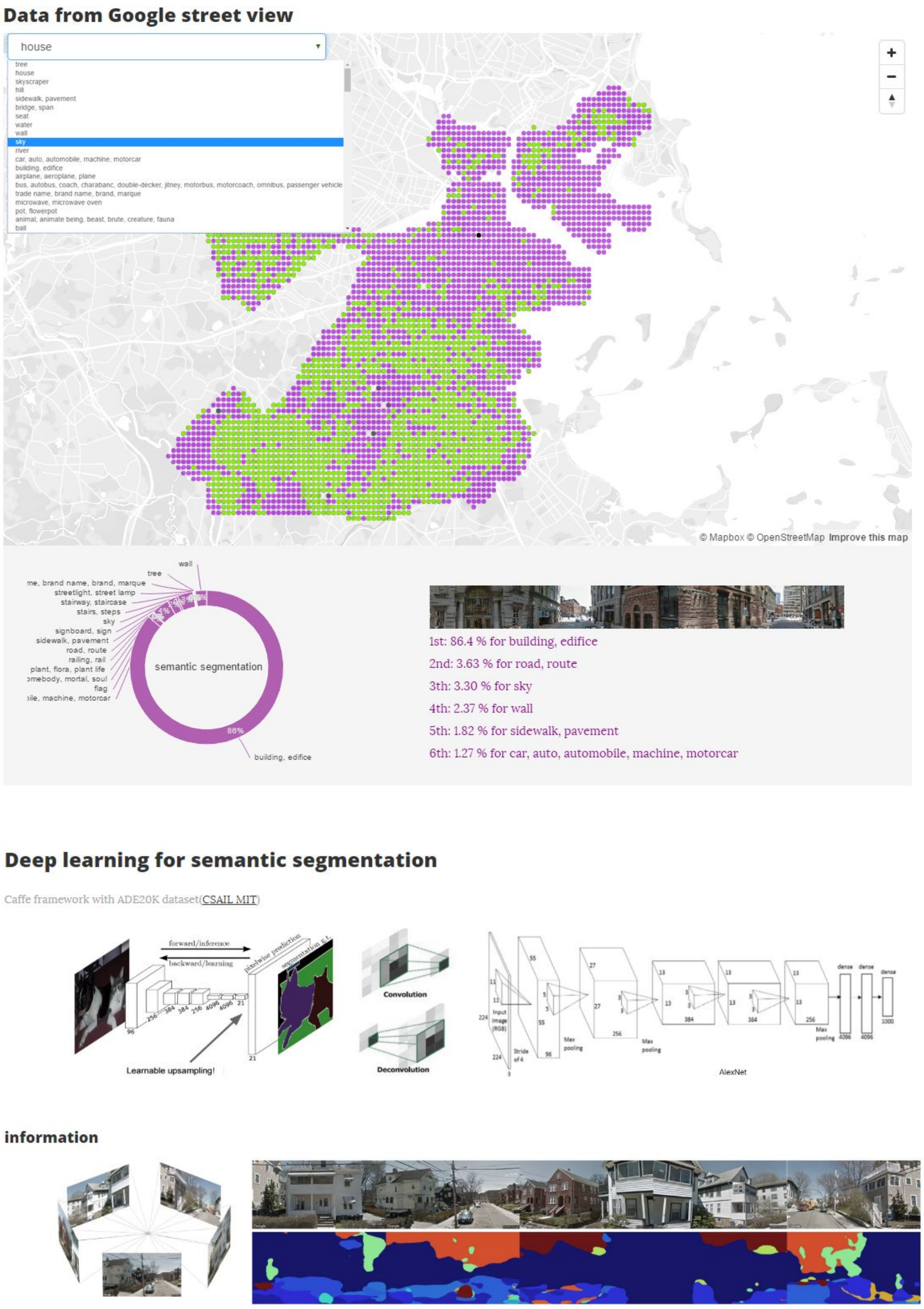


Based on walkable distance(10 min [800 meter, 0.497097 mile]) each Rent and Housing location count the number of place

The number of data used: 20,559 places

Feature:
"walkMbta" : the number of MBTA within the distance
"walkUniversity" : the number of university within the distance
"walkSchool" : the number of school within the distance
"walkPark": the number of park within the distance

Google Street View:



Data from Google street view

1st: 86.4 % for building, edifice
2nd: 3.63 % for road, route
3th: 3.30 % for sky
4th: 2.37 % for wall
5th: 1.82 % for sidewalk, pavement
6th: 1.27 % for car, auto, automobile, machine, motorcar

Deep learning for semantic segmentation

Caffe framework with ADE20K dataset(CSAIL MIT)

information

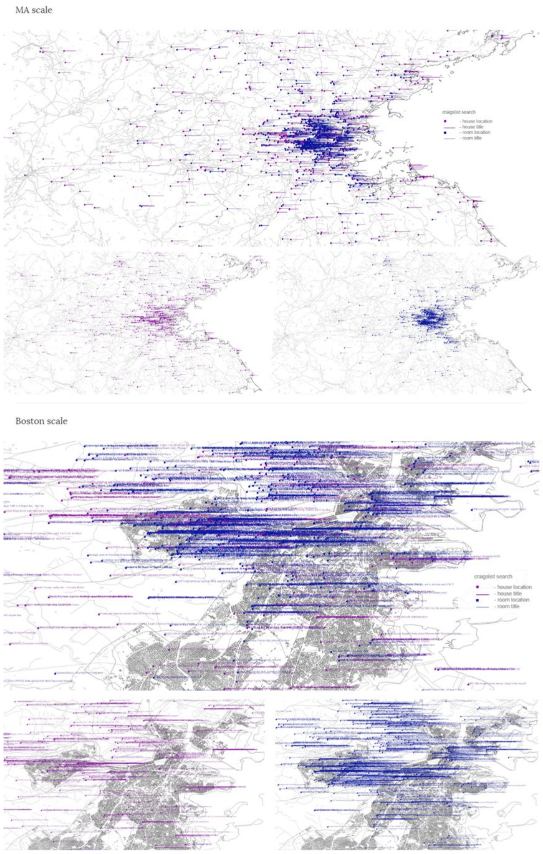
Based on semantic segmentation using Deep Learning for google 360 street view, we are able to get the pixel ratio of individual features from each image.

Image data used: 18,588 images

Feature:

“pixelSky” : number of pixel for sky in the panorama picture
“pixelLake” number of pixel for lake in the panorama picture
“pixelPerson” number of pixel for person in the panorama picture
“pixelCar” number of pixel for car in the panorama picture
“pixelBuilding” : number of pixel for building in the panorama picture
“pixelHouse” : number of pixel for house in the panorama picture
“pixelSidewalk” : number of pixel for sidewalk in the panorama picture
“pixelRoad” : number of pixel for road in the panorama picture
“pixelGrass” : number of pixel for grass in the panorama picture

Data from Craigslist Boston

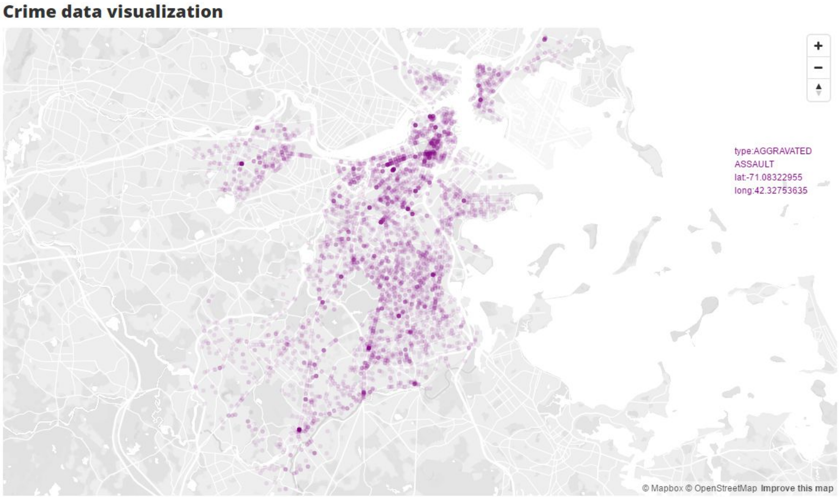


Purple : house post
Blue : room post

number of data used and keyword:
house - 2,500 posts including text data
room - 2,500 posts including text data

Feature:
“numCraigslistRoom” : number of room post in the pixel in Boston area
“numCraigslistHouse” : number of house post in the pixel in Boston area

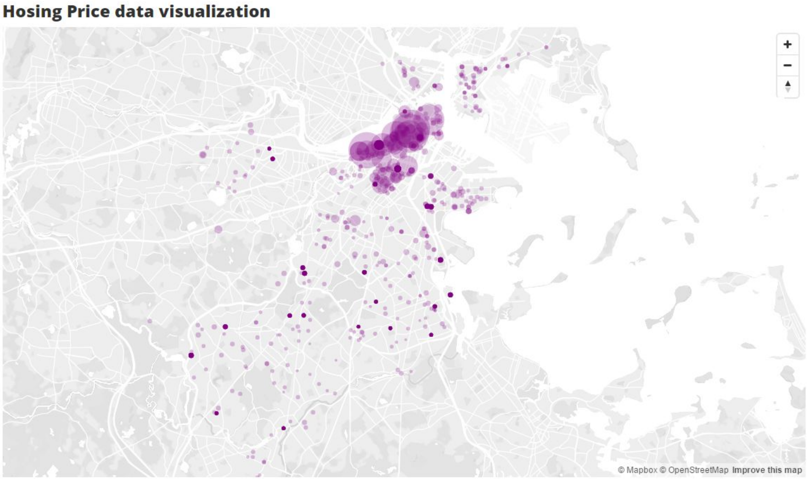
Crime data from City of Boston


Crime data visualization

The crime data visualization shows the relationship between area and violence in a sense of social problem in the urban context. Although the data covers range from trivial one to serious incident, it could be categorized and harnessed for train data.

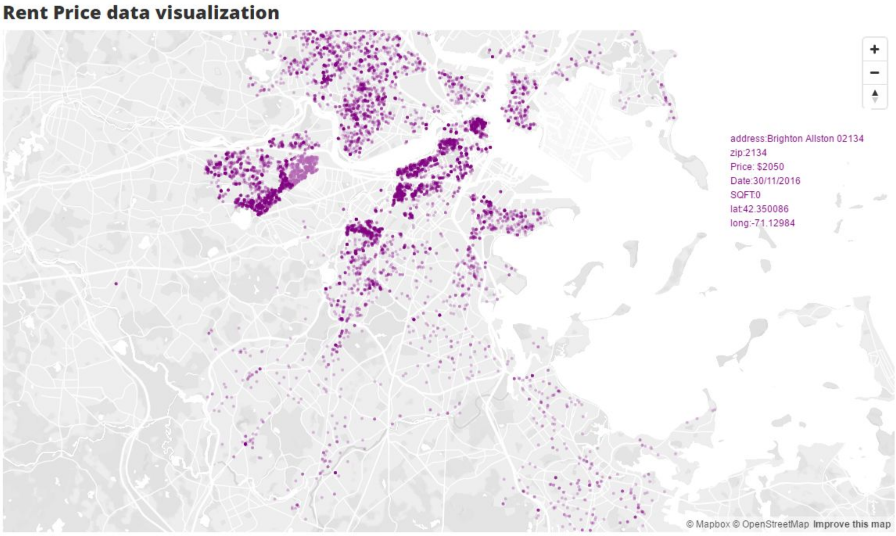
The number of data used: 5,001 crimes

Housing Price data from Zillow


Hosing Price data visualization

Most importantly, the housing price data from Zillow is a key data for training model. This is because one of our assumption is that there should be a strong relationship between the prices and urban environmental condition.
number of data used: 988 houses

Rent Price data from Trulia


Rent Price data visualization
address:Brighton Allston 02134
zip:2134
Price: $2050
Date:30/11/2016
SQFT:0
lat:42.350086
long:-71.12984
© Mapbox © OpenStreetMap Improve this map

The number of data used: 13049 rents
In addition to the housing price, the rent price is also useful data for training the model because it shows a similar pattern to housing data in a sense.