CS591 Lab Report
Team: Ailing Yu, Jin Tang, Yuxiao Wang, Yunyu Zhang

Analyzing Real Estate Investment in Boston

Introduction

Boston is a beautiful and convenient city to live in. Because half of the Boston population are students, the Boston population is fluid. People need to rent a house to live. Moreover, because more and more technology companies are coming to Boston, the labour populations will grow. Boston may become the next Palo Alto. Therefore, the housing market will bloom in Boston. The purpose of this project is to find the best spot in Boston to invest in real estate. We take education, natural environment, rent, transportation, social facilities and potential danger into consideration to get the best spot in Boston to live and invest in.

Datasets

Boston Housing Rental Price:

http://datamechanics.io/data/boston_rentalPrice.csv

Boston Colleges and Universities:

http://datamechanics.io/data/alyu_sharontj_yuxiao_yzhang11/Colleges_and_Universities.geojson

Boston Fire Incidents:

http://datamechanics.io/data/2013fireincident_anabos2.json

http://datamechanics.io/data/2014fireincident_anabos2.json

http://datamechanics.io/data/2015fireincident_anabos2.json

Boston Garden:

http://datamechanics.io/data/alyu_sharontj_yuxiao_yzhang11/garden_json.json

Boston Hospital:

http://datamechanics.io/data/alyu_sharontj_yuxiao_yzhang11/hospitalsgeo.json

Boston Hubway:

http://datamechanics.io/data/hubway_stations.csv

Boston MBTA:

http://datamechanics.io/data/alyu_sharontj_yuxiao_yzhang11/MBTA_Stops.json

Data transformation

We performed several transformations to product six new datasets.

We divide the big boston area by zip codes.

1. education_trans_avg

   We processed Colleges and Universities Dataset, Hubway Dataset, MBTA Dataset to get, in each zip code area, the number of colleges and universities and the average number of transportations around those colleges and universities (within 0.8km).

2. average_rent_zip

   We processed Rental Dataset to get the average rent rate for each zip code area.

3. education_rent

   We processed Colleges and Universities Dataset and average_rent_zip Dataset to get, in each zip code area, the number of colleges and universities and the average rent rate.

4. Fire_Hospital_vs_Rent

   We processed Fire Dataset, Hospital dataset, and average_rent_zip Dataset to get, in each zip code area, the ratio: the number of fire divided by the number of hospital and the average rent.

5. garden_vs_rent

   We processed average_rent_zip dataset and garden dataset to get the number of garden and the average rent rate for each zip code area.

6. Correlation

We took all the previous datasets and calculated the correlations between rent and all other factors.

Algorithms, Techniques and Analysis

We used constraint satisfaction to get the top zip codes in the Boston area. For each factor we considered, we calculated the mean μ and standard deviation σ . We only took valid data form $[\mu - 3\sigma, \mu + 3\sigma]$. We calculated the correlation between factors and then we used correlations to calculate weight for each factors. For each zip code, we multiplied weights to each factors and got the final evaluation score for each area. We then used greedy algorithm to get the most valuable areas to invest.

We use correlation to see the relations between different factors. For example, the relationships between rent price and garden number in the area. Here are some coefficient graphs:

Figure 1.0 Number of Schools vs. Average Rent in the same zip code

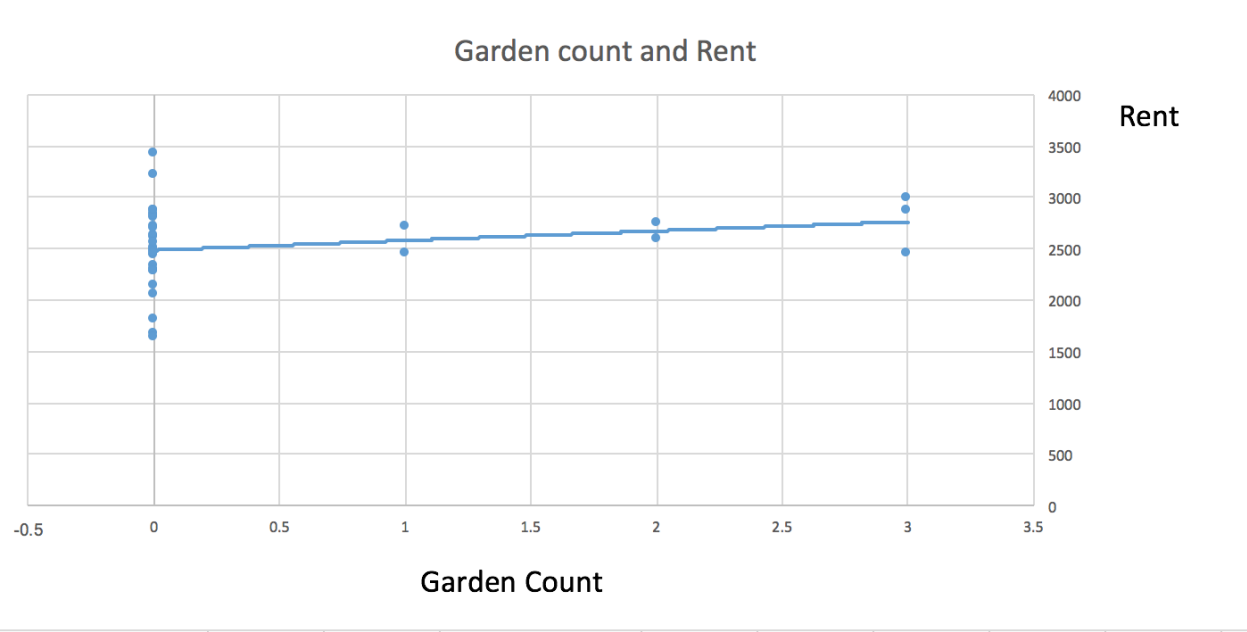

Figure 1.1 Number of Gardens vs Average Rent in Certain Zip Code

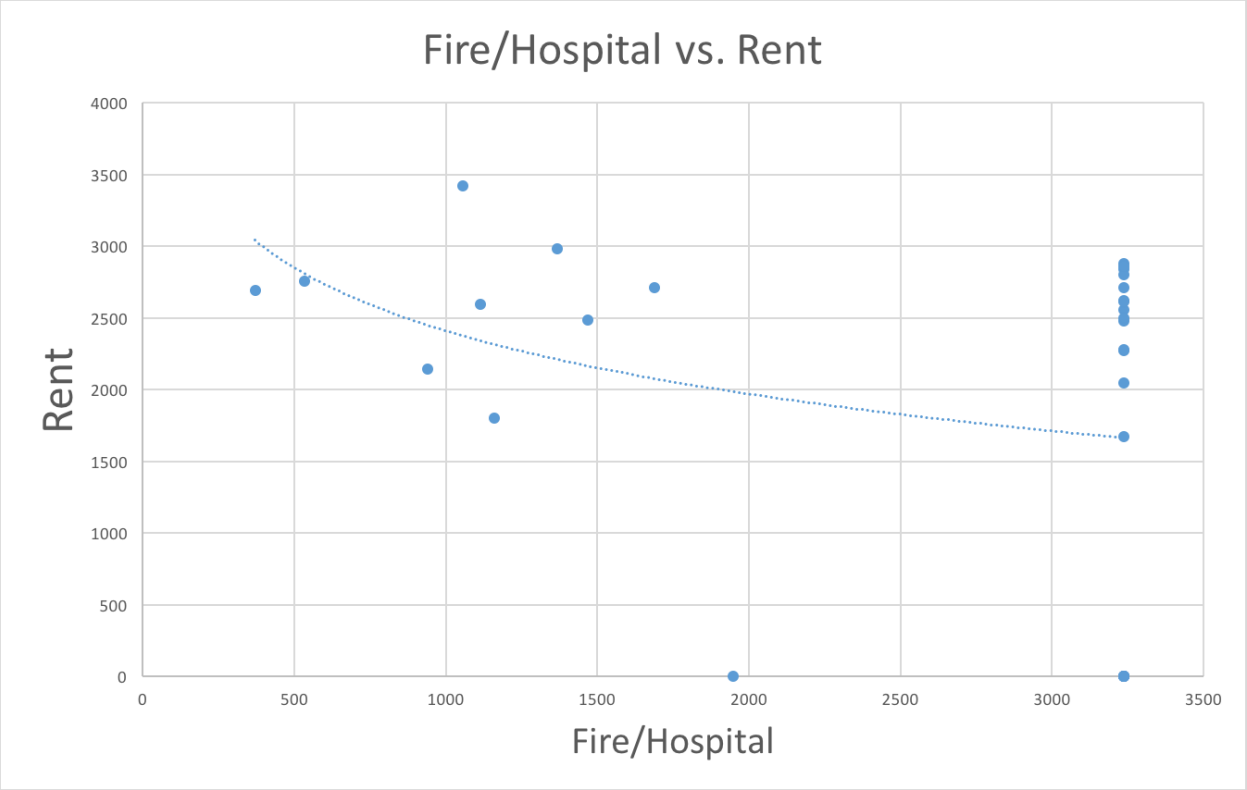

Figure 1.2 Number of Fire/ Hospital Rate vs Average Rent in certain area

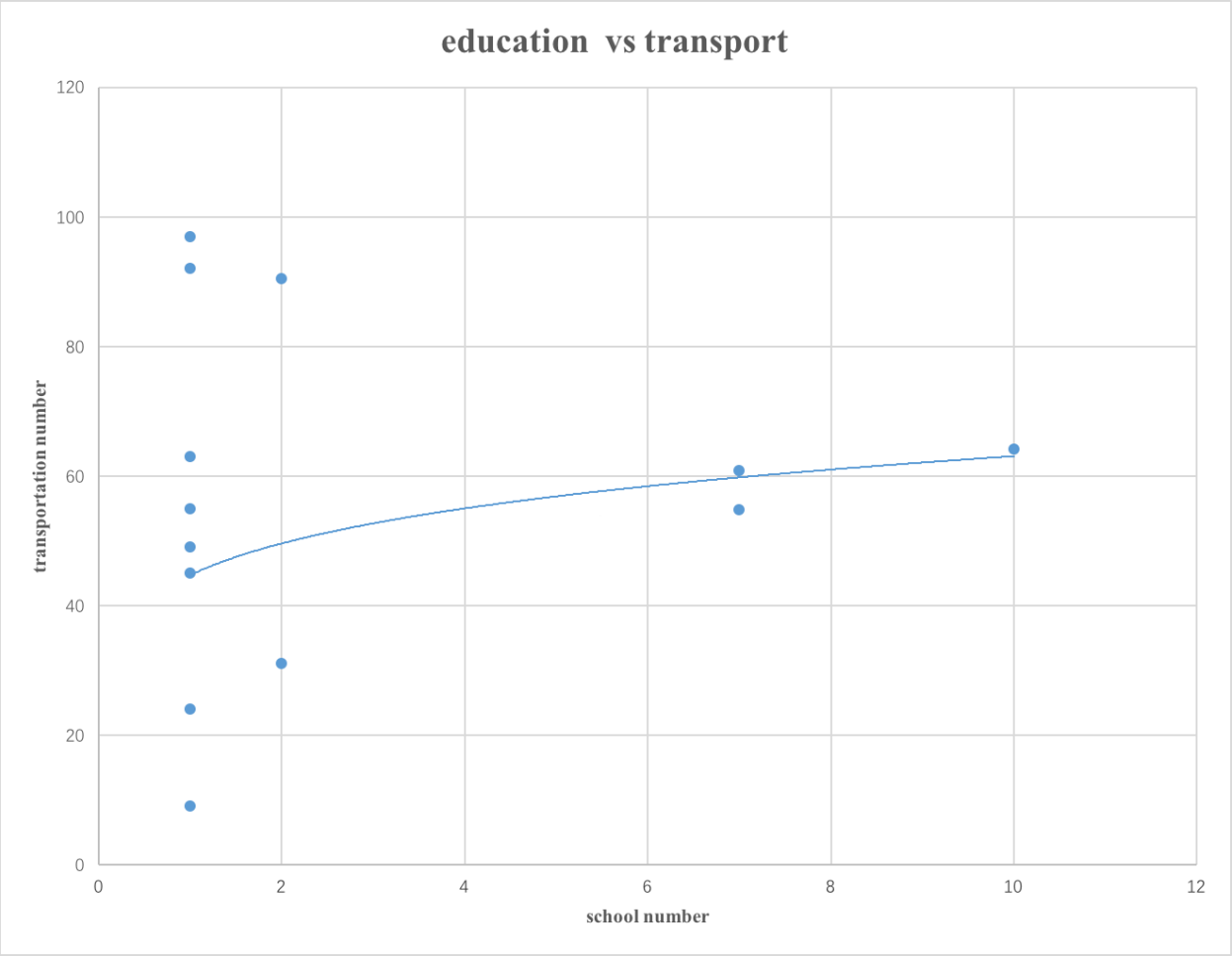

Figure 1.3 Number of Public Transportation Stops vs Number of Schools in Certain Area

Final Conclusion

We calculated the score of each zip code in order to get the top 5 zip code areas to invest. First, We define several symbols as follows:

S = the score of a certain zip code

Ng = Number of gardens in a zip code

Wg = Weight of gardens

Ns = Number of schools in the zip code

Ws = Weight of Schools

Nt = Number of MBTA stops and hubways stops in the zip code

Wt = Weight of transportation stops

R = Average rent in the zip code

Wr = Weight of Average Rent

Second, we calculated the correlation coefficients and corresponding weights. We calculated the correlation coefficient and corresponding weight for the four main factors(number of gardens, number of colleges and universities, number of transportations, and the value of the ration by the number of fire divided by number of hospitals) in the Correlation.py file. The result is as follows.

Correlation	Correlation Coefficient	Weight
Rent vs Education	0.5418132534754234	0.5418132534754234
Rent vs Garden	0.04038425343016952	0.05082871879198115
Rent vs Fire/Hospital	-0.32365265581635083	0.40735802773259544
Education vs transportation	0.04157	0

Table 1.1 **Correlation Coefficient and Weight Table**

Third, we calculated the final score for each zip doe by the formula

S= Ng*Wg + Ns*Ws + Nt*Wt + R*Wr

Finally, the final ranking and score for Boston zip codes were the following

Zip code	Score
02116	70.93619422
02115	69.54518613
02215	61.80635799
02129	58.68282882
02111	56.58042778
02113	51.35646218
02114	48.24488218
02135	47.07520201
02118	45.32702261
02132	43.04113225
02127	41.53460123
02130	40.23334884
02125	36.71758491
02134	36.34109667
02128	36.06230542

02108	33.31033685
02109	21.12737158
02110	12.18296527
02136	12.18296527
02124	12.18296527
02120	12.18296527
02210	0
02199	0
02163	0
02131	0
02126	0
02122	0
02119	0
02121	0

Table 1.2 Zip Code Score Table

The map visualization is the shown in Figure 2.0. Areas with darker color had the higher final scores.

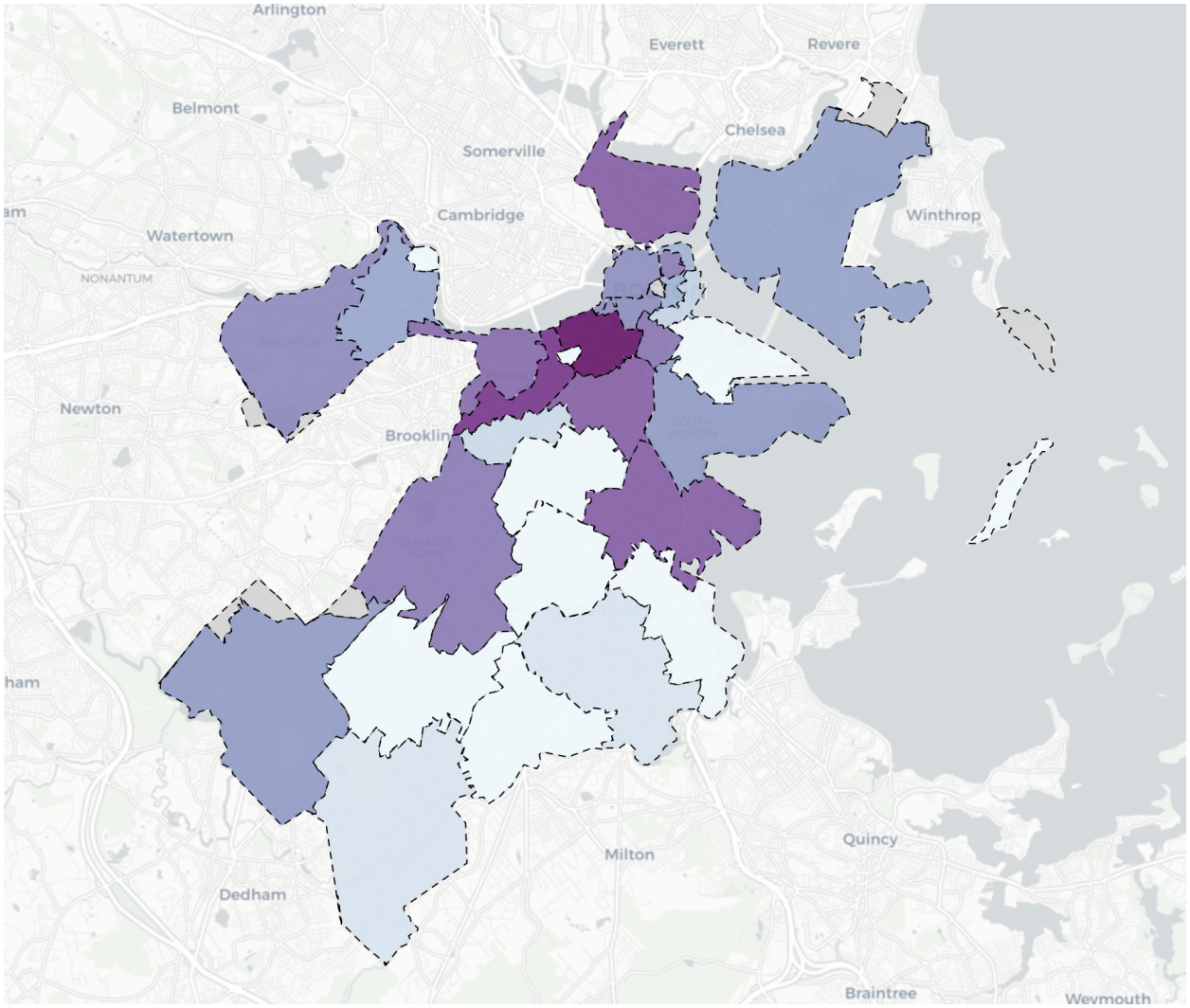

Figure 2.0 Heat map of final score

We claimed that the top 3 areas in Boston to invest were 02116, 02115 and 02215.

Limitation and Future Work

There were many limitations and things we could improve in this project. Looking for a useful dataset is the biggest limitation we had. We take 5 factors in our consideration because these datasets were accessible. Even though we found a dataset, some of the data are not in good quality. For example, in the average rent price dataset from Zillow, there are many blanks in

certain zip codes. To fill up the blank, our team calculated the average rent price in Boston and used the rent to run the tests. In the future, we should take more factors into our calculation.

How to run the project:

Take out *app.py* and *mapping.py* in the folder.
Run $ python3.6 execute.py alyu_sharontj_yuxiao_yzhang11
This previous step will generate the database.
Put back the files *app.py* and *mapping.py.*
Run app.py.