Public Facilities, Crimes and Average Income

Yehui Huang, Chang Gao, Yingqiao Xiong, Hongyu Zhou
Professor: Andrei Lapets

Introduction

It is commonly thought that the public facilities, such as schools, hospitals, transportations, are unevenly distributed due to the favor to high-income communities. Such bias forms a deprivation of low-income communities. However, according to the work of Ellaway et al. 2008, "it appears that in the early 21st century access to resources does not always disadvantage poorer neighbourhoods in the UK".

In the study of Ellaway et al., twelve resources had higher density in, and/or were closer to or more common in, more deprived neighbourhoods: public nurseries, public primary schools, police stations, pharmacies, credit unions, post offices, bus stops, bingo halls, public swimming pools, public sports centres, outdoor play areas, and vacant and derelict land/buildings. Sixteen had higher density in, and/or were closer to, or more common in, more affluent neighbourhoods: public secondary schools, private schools, banks, building societies, museums/art galleries, railway stations, subway stations, tennis courts, bowling greens, private health clubs, private swimming pools, colleges, A&E hospitals, parks, waste disposal sites, and tourist attractions. Private nurseries, Universities, fire stations, general, dental and ophthalmic practices, pawn brokers, ATMs, supermarkets, fast food chains, cafes, public libraries, golf courses, and cinemas showed no clear pattern by deprivation.

To support the claim of Ellaway et al. and to exam their conclusion in the greater Boston area, we analyze the relationships between average income of residents in certain neighborhoods and number of public facilities, crimes in the same neighborhoods throughout the greater Boston area. In order to analyze the relationship, we use the calculate the numbers of the hospitals, public gardens, schools, stores and crimes in each neighborhood, and then perform MapReduce algorithms on the data to derive the intended format and content. We also operate machine learning algorithms, such as k-means and linear regression to process data such that the correlation between each factor and income can be derived, and the numerical coefficient can be estimated. We conclude that there is no biased distribution of public facilities and no clear evidence of deprivation.

Data Sets

The seven datasets we used:

1. Hospital_locations:
   (Locations of Hospitals within the city)
   https://data.cityofboston.gov/resource/u6fv-m8v4.json
2. Employee_earnings_report_2015:
   (Earnings for City of Boston Employees for year 2015.)
   https://data.cityofboston.gov/resource/bejm-5s9g.json
3. Health_corner_stores:
   (Corner stores and bodegas participating in the Healthy Corner Stores Initiative sponsored by the Boston Public Health Commission, and other community health center healthy corner store project locations.)
   https://data.cityofboston.gov/resource/ybm6-m5qd.json
4. Community_gardens:
   (Community garden sites throughout Boston. Access may be restricted to community members. Inaccurate information? Contact us with corrections.)
   https://data.cityofboston.gov/resource/rdqf-ter7.json
5. Public_schools:
   (Location of Boston Public Schools)
   https://data.cityofboston.gov/resource/492y-i77g.json
6. Crime_incident_reports:
   (Crime Incident Reports provided by Boston Police Department including longitude and latitude)
   https://data.cityofboston.gov/Public-Safety/Crime-Incident-Reports-July-2012-August-2015-Sourc/7cdf-6fgx/data
7. Approved_building_permit:
   (Data of approved building/construction permits with zip Code)
   https://data.cityofboston.gov/Permitting/Approved-Building-Permits/msk6-43c6/data

Methodologies

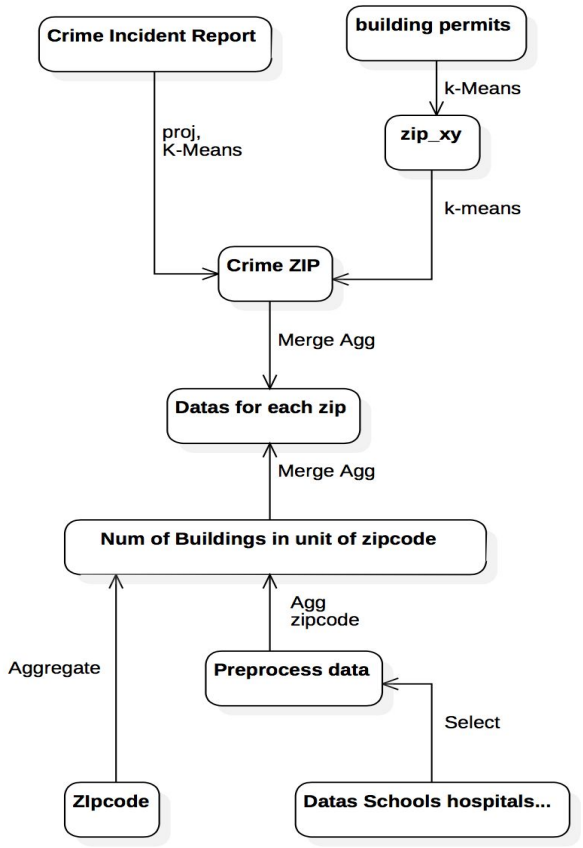
Firstly we can obtain the number of average earnings in each zip code area by using MapReduce to project and reduce to get the total earnings and counts in each zip code, then we can obtain the average earning by (total earning divides counts) for each zipcode area.

Secondly we can obtain number of hospital, number of garden, number of store and number of school group by the zip code by applying the simple MapReduce Function.

We extract two new datasets CrimeIncidentReport and ApprovedBuildingPermit. We use the k-means s an optimization to calculate the approximated mean point in each zip code area and use the argmin to place the crime coordinates into his closest zip code mean point.

Combining all datasets generated by the second step, we obtained a number of these building grouped by the zipcodes.

Lastly, we merge the previous dataset with the average earnings dataset to obtain a dataset that contains the number of these buildings and average earning using zipcode as unit.

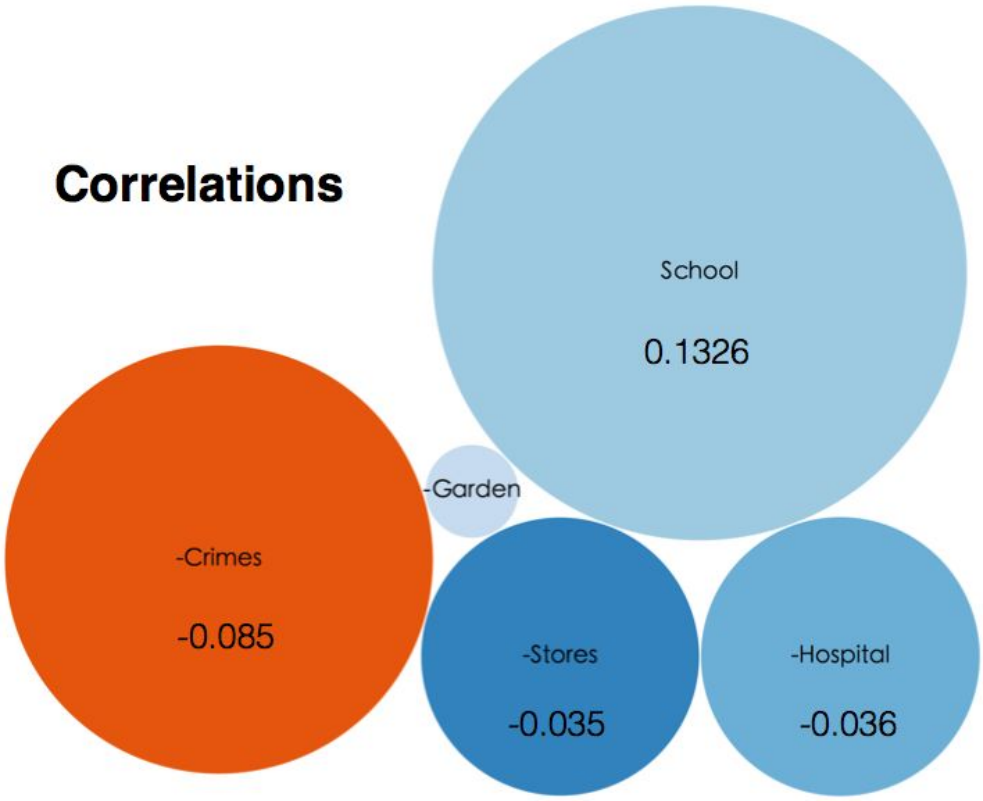


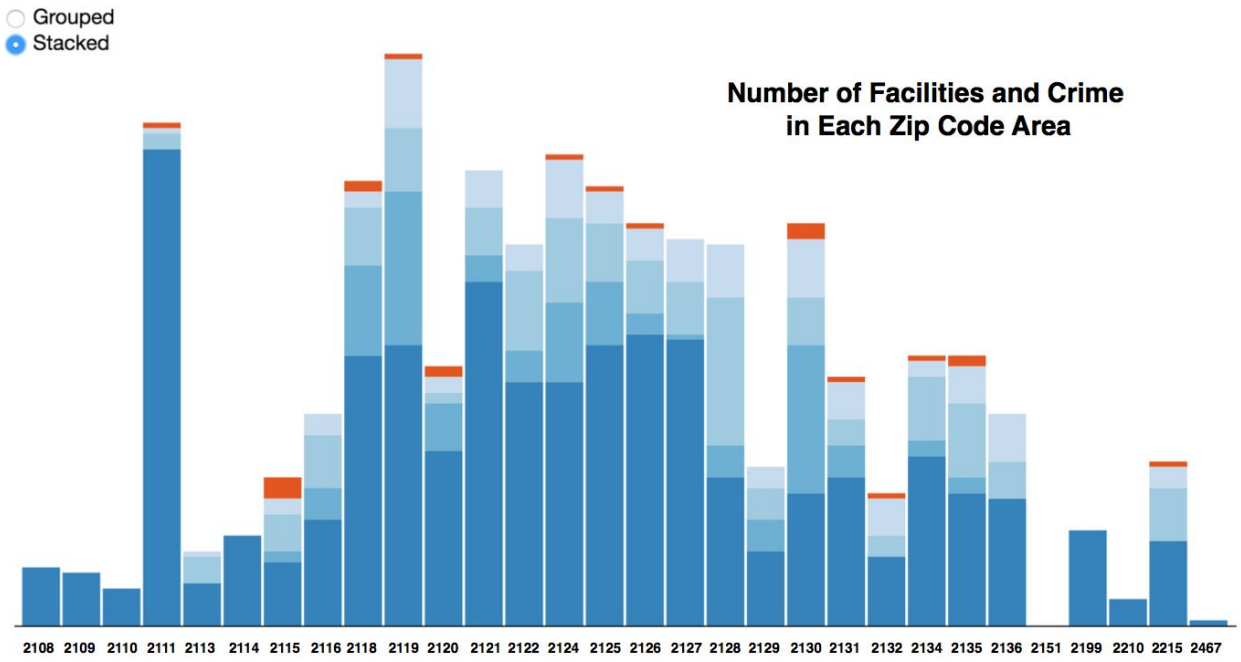
Statistic Operations:

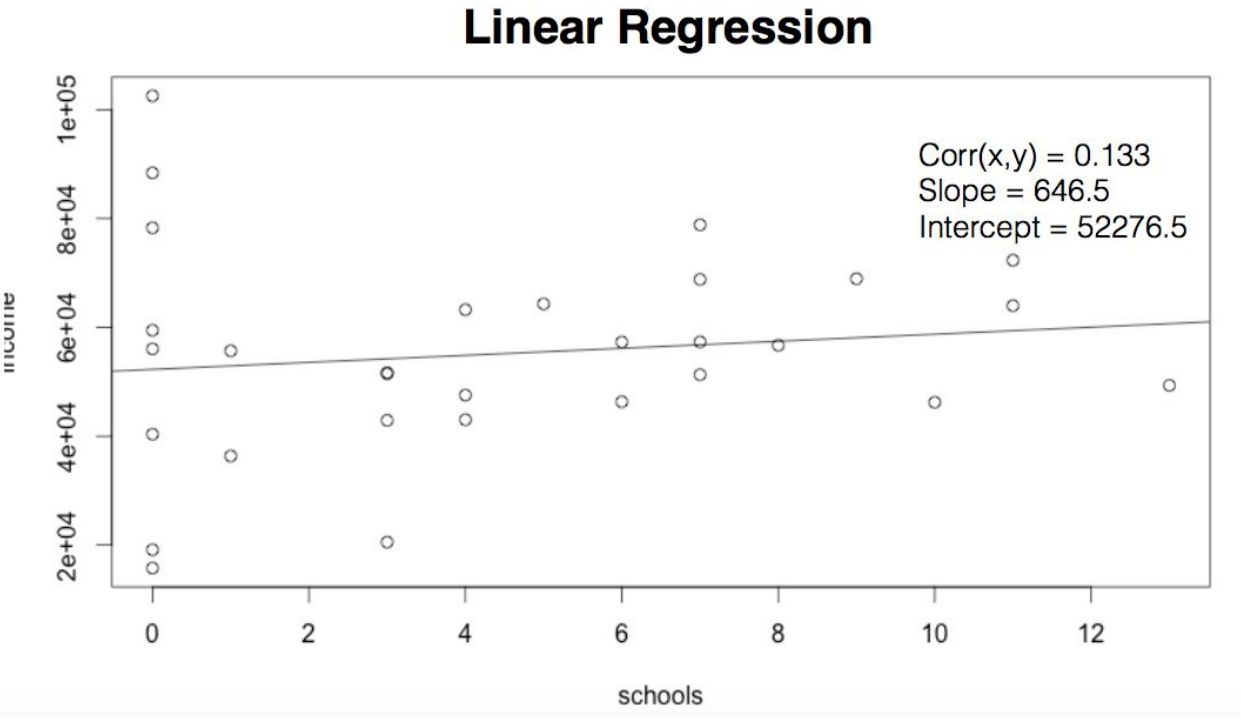
- Correlation Coefficient: taking the numbers of corner stores, hospitals, public schools, community gardens and crimes as independent variables, and average income as dependent variable, we calculate the correlation coefficients to measure the dependency between the dependent variable and each of the independent variable, to estimate their relationships. The formula applied are identical to those shown in the class notes.
- Linear Regression: taking the variables described above, we fit the data into the linear least square regression model, and calculate the estimated slope and interception of the regression line. Using the coefficients returned, we can quantify the effect of each independent variable. This method shall be improved using nonlinear regression, as many of the relationships are not linear.
- Coefficient of Determination (R-squared): taking the estimations derived above, we calculate coefficients of determination for each of the regression, to measure how well the linear model fits into the data.

Results

The correlations between average income and each of the possible factors are vague, as most of the correlation coefficients fall in the range of [-0.1, 0.1]. The linear regression model does not fit well for the data sets, as the coefficients of determination are noticeably close to 0. Thus we conclude that there is no biased distribution of public facilities and no clear evidence of deprivation. There is also no clear pattern of higher criminal rates in low-income areas.




Grouped
Stacked
Number of Facilities and Crime
in Each Zip Code Area
2108 2109 2110 2111 2113 2114 2115 2116 2118 2119 2120 2121 2122 2124 2125 2126 2127 2128 2129 2130 2131 2132 2134 2135 2136 2151 2199 2210 2215 2467


Linear Regression
Corr(x,y) = 0.133
Slope = 646.5
Intercept = 52276.5
2e+04
4e+04
6e+04
8e+04
1e+05
0
2
4
6
8
10
12
schools

Possible Extensions

Because the income data we use includes only those of employees of City of Boston, we should collect much more data spanning different enterprises and occupations. We shall also consider other factors which could have affected the average income in certain areas, and also consider the correlations between many of the independent variables.

In addition, we should investigate how satisfied of residents towards the public facilities in the neighborhood they live in, so that we should be able to provide suggestions to City of Boston regarding the issues of urban property management, such as the planning of constructions of schools and gardens.

Reference

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2570170/