**CS591 Data Mechanics – Report**

**Introduction**

We are interested in the relation between personal income and amount of infrastructures in different regions of Bostion city. Do people with higher income share more infrastructure resource than others, or no significant relation of income and infrastructure resource?

To anwser the question, in this project, data of employee earnings, hospital and public/private school in Boston city are investigated to generate four result datasets, which may help figure out the question.

**Input datasets**

The datasets used from the City of Boston Data Portal were:

* Employee Earnings Report 2013: <https://data.cityofboston.gov/resource/54s2-yxpg>
* Employee Earnings Report 2014: <https://data.cityofboston.gov/resource/4swk-wcg8>
* Hospital Locations: <https://data.cityofboston.gov/resource/46f7-2snz>

The datasets used from the Boston Wicked Open Data was:

* Public Schools: <https://boston.opendatasoft.com/explore/dataset/public-schools>

The datasets used from the BostonMaps Open Data was:

* Nonpublic school:

<http://bostonopendata.boston.opendata.arcgis.com/datasets/0046426a3e4340a6b025ad52b41be70a_1>

**Transformation**

1. Fetch 5 datasets from web services and store them in mongodb.
2. Process raw data of employee earnings(2013, 2014) and calculate average personal income for each region of Boston city
3. Process raw data of public/private schools and count amount of schools(public, private and all, respectively) in each region of Boston city
4. Process raw data of hospital locations and count amount of hospitals in each region of Boston city
5. Join data of income, school, hospital by region to see relation between those factors

**Output datasets**

4 main datasets were generated based on transformations above.

1. regionincome: average personal income for each region of Boston city
2. regionschool: amount of schools(public, private and all, respectively) in each region of Boston city
3. regionhospital: amount of hospitals in each region of Boston city
4. income\_infrastructure: relation between income and infrastructure resource in each region of Boston city

**Statistic Analysis**

Two methods of statistic analysis (1. pearson correlation coefficient & 2. T-test) are applied on dataset income\_infrastructure. Based on data of personal income and amount of infrastructures in each region of Boston city, the relation between income and amount of infra is investigated by calculating pearson correlation coefficient. On the other hand, infra amount data of regions is splited into two parts (regions with higher income and lower income). T-test is applied to find the significance of the difference between the means of the two parts.

**Webservice & Visualization**

We build a webservice providing RESTful APIs of fetching datasets based on Flask, a lightweight Python web framework. Figure 1 shows the infrastructure of the webservice. MongoDB is used to store datasets derived by the transformation above.

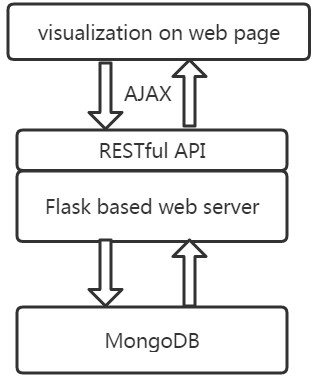


Figure 1: the infrastructure of the webservice

Analysis result and statistic information of datasets are visualized on a web page, which is generated using RESTful API provided by the webservice. For nice visualization, we use a javascript library named ECharts to draw related graphs. Sample visualization result of datasets are showed in Figure 2, 3, 4.

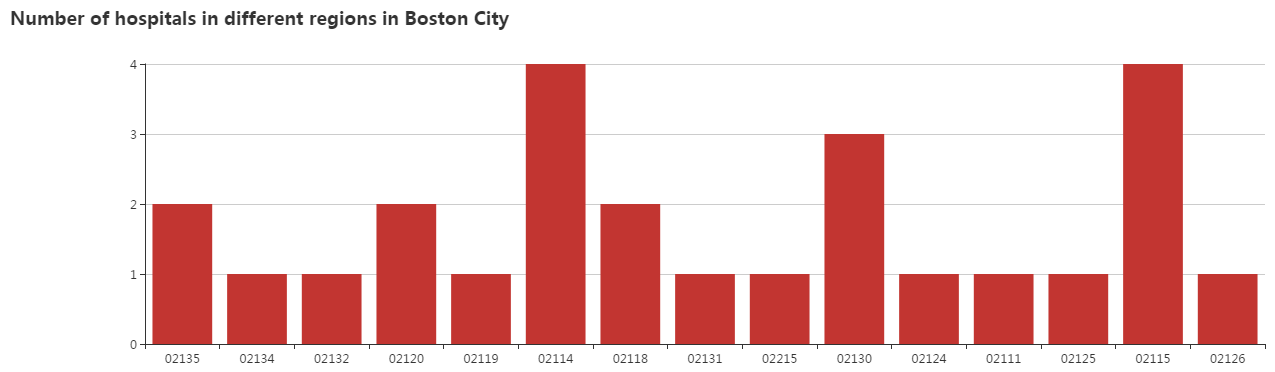


Figure 2: Number of hospitals in different regions of Boston City

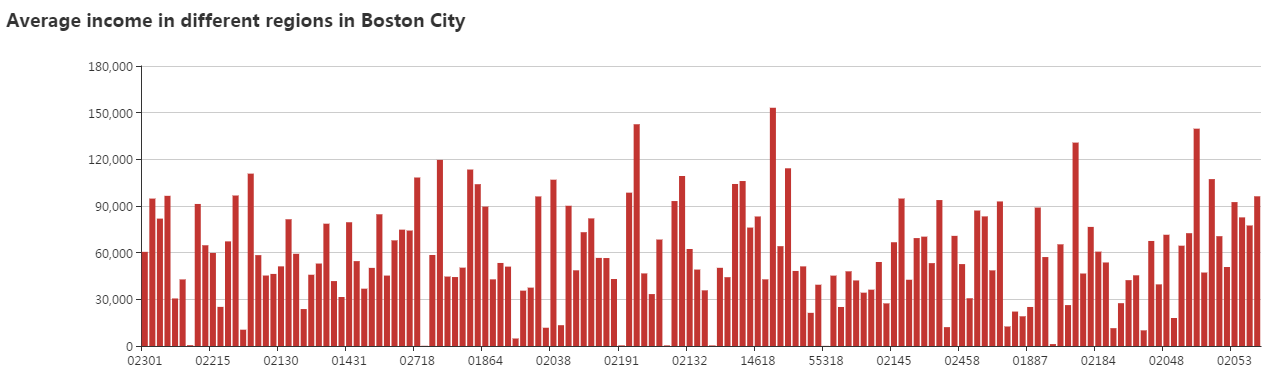


Figure 3: Number of average income in different regions of Boston City

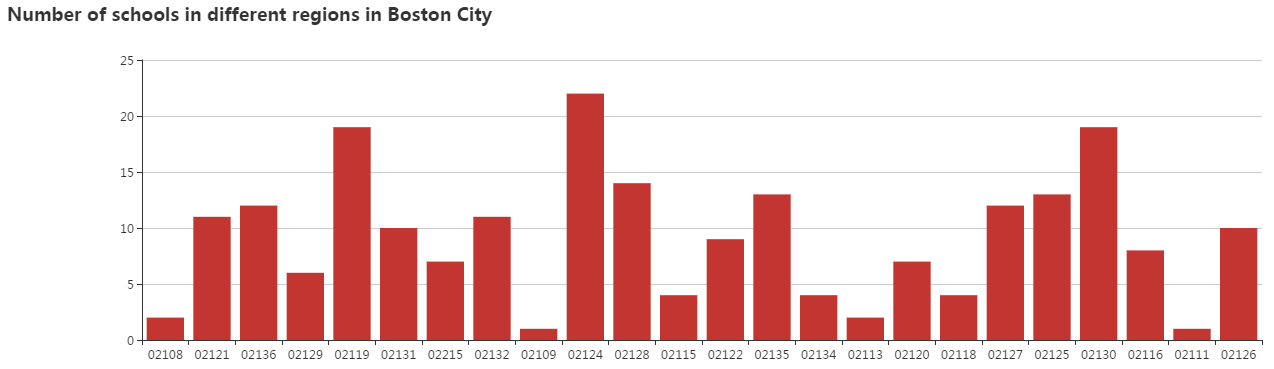


Figure 3: Number of schools in different regions of Boston City

**Analysis Result**

Pearson correlation coefficient on region income and infrastructure is showed in Figure 4. Average income was negatively related to number of school and hospitals based on this result.

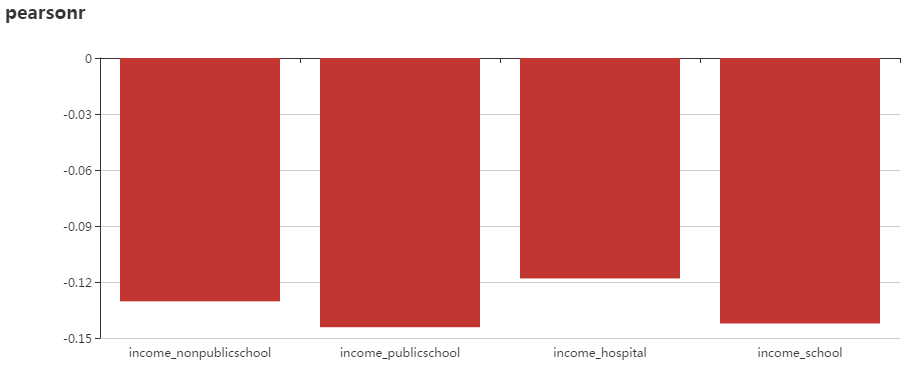


Figure 4: correlation coefficient on region income and infrastructure in Boston city

In order to check whether the result found from pearson correlation coefficient is significant, we compared regions with higher income and that with lower income and applied ttest on them. The results is showed in Figure 5, 6 and Table 1.

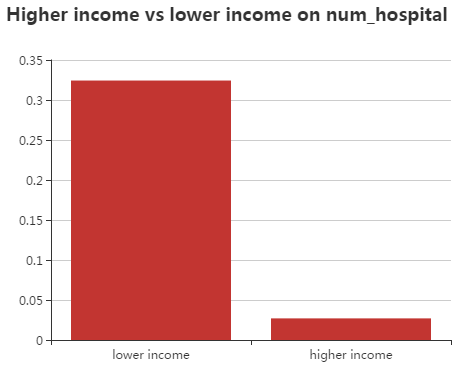


Figure 5 Average number of hospitals

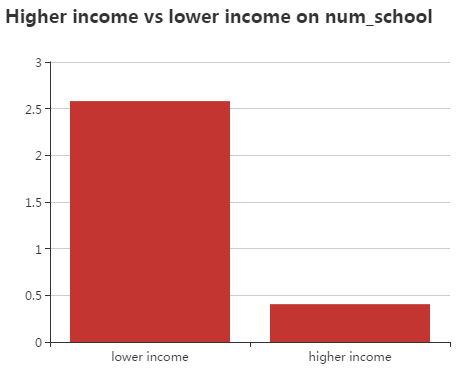


Figure 6. Average number of schools

|  |  |  |  |
| --- | --- | --- | --- |
|  | Avg num in lower income regions | Avg num of higher income regions | p-value |
| Num of hospitals | 0.324 | 0.027 | 0.0035 |
| Num of schools | 2.581 | 0.405 | 0.0011 |

Table 1

**Conclusions**

The result made in this project may suggest that regions with higher income do not share more infrastructure in Boston City. In contrast, there may be more schools and hospitals in regions with lower income than that with higher income.