CS591 Final Report

Tianyou Luo, Linshan Jiang

Professor: Andrei Lapets

Date: May, 4th, 2016

Objective:

Two ways of characterizing the living standard of different areas is average income and frequency of crime happened in that area - we would describe a place with high living standard "people earn much, and security level high on average". However, is there a relation between average income and number of crimes?

Many of us may expect more crime in low-income areas, due to lower living standards. In the project, we would examine the relationship between the average income and the number of crime incidents among Massachusetts using several mechanisms such as diagrams, covariance, and p-value in grouping units of zip code. This project may help people obtain overall information about humanistic environment about neighborhoods, especially for those who are searching for a new place to settle down.

Resources Used:

1. 'crime_incident_reports':'https://data.cityofboston.gov/resource/7cdf-6fgx.json?year=2014'
   This dataset contains information about crime reports with specific locations (longitude and latitude).
2. 'employee_earnings_report_2014':'https://data.cityofboston.gov/resource/4swk-wcg8.json'
   From this data file we know the link between zipcode and earning.
3. 'approved_building_permits':'https://data.cityofboston.gov/resource/msk6-43c6.json
   From the third dataset we can get the relationship between location and zip codes.

These three Jason files are all downloaded from *data.cityofboston.gov*.

Data Flow:

We use a diagram below to demonstrate how we manipulated the datasets to get new relationships for later use.

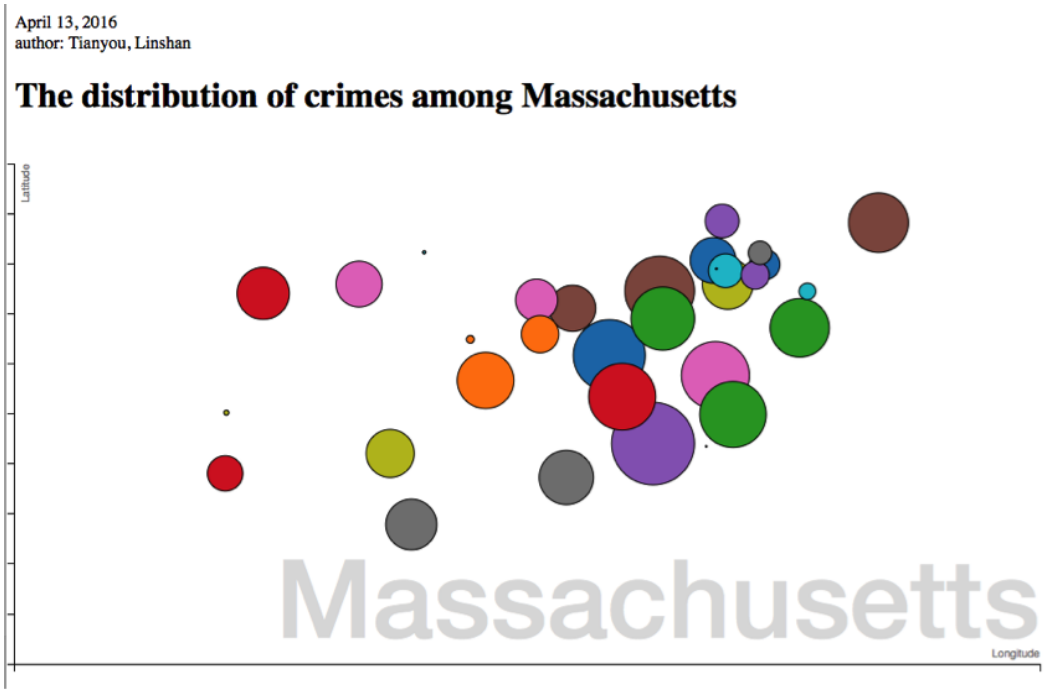
The crime incident reports and the employee earnings report are the two main datasets, which contain the key information we care. Because they do not share the same attribute, the approved building permits acts as a glueing dataset. More specifically, the crime dataset has location in the

forms of (longitude, latitude) and other crime related information, but the earnings dataset only contains zip code. Then a map between zips and locations is needed to join these two, in which the permit dataset contains. Combining(union) the first and third dataset by location, aggregating by (zipcode,sum, K mean of locations) we can get the number of crimes in each zip code area, and also the location of each zip code. Combine(union) the second and third dataset by location, aggregate by (zip code, avg) we can get average income in each zip code area. Then combine the result datasets by zip code, we can get the relationship between average earning and number of crimes of each zip code.

Analytic Diagrams:

The visualization tool we used is D3, which can be reference here: *https://d3js.org*

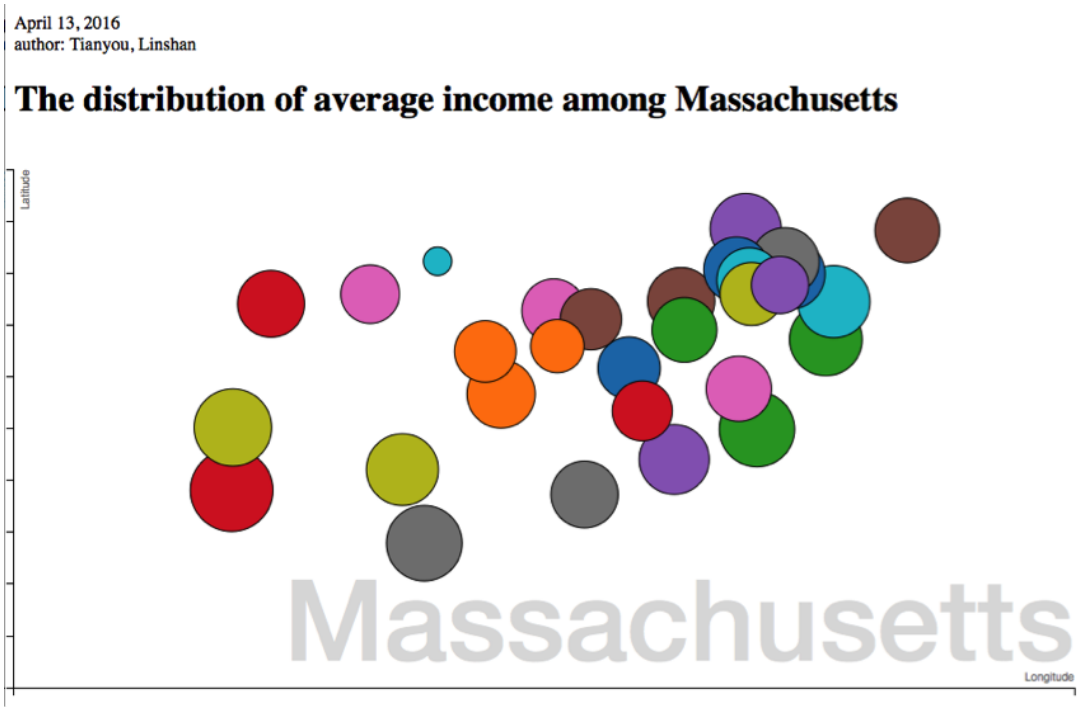
1. Distribution of Crimes:



The above diagram is the distribution of crimes in Massachusetts. We plot the circles on coordination composed of longitude and latitude, thus one can view this diagram as a "map". Each center of circle resides on the corresponding longitude and latitude of that zip code location it represents (The location of each zip code is determined by running k-means algorithm on all

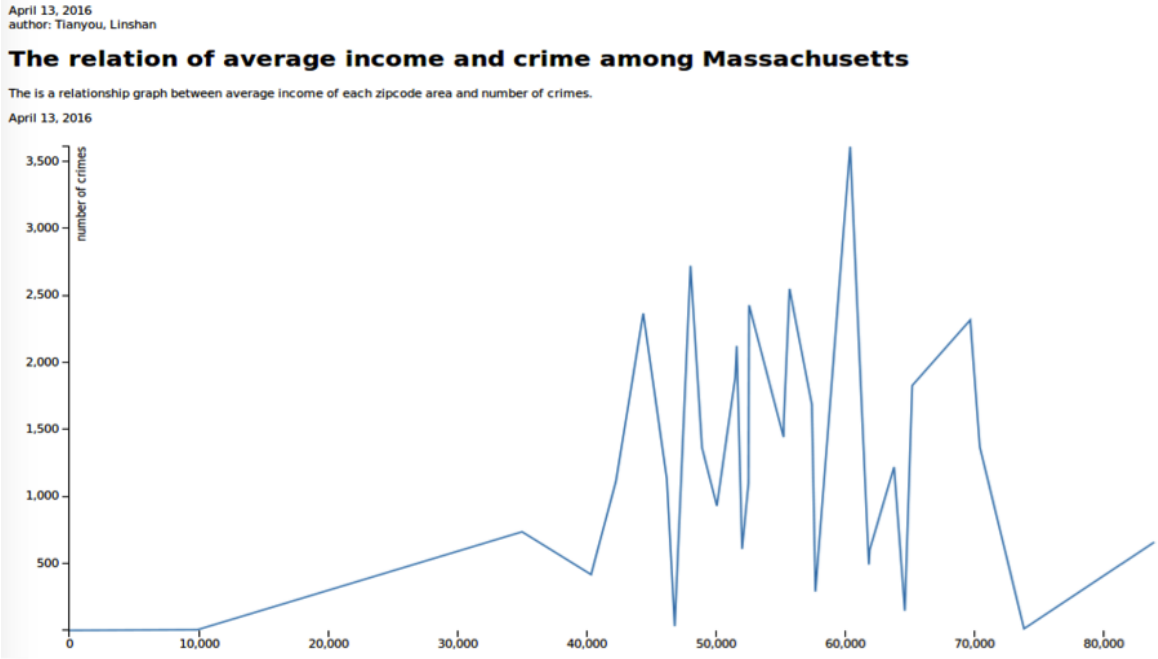
data coordinates we have for that zip code). Furthermore, the larger the radius of the circle suggests the more crimes happened in that area. The different colors do not have actual representations and are just for differentiating circles (make it easier to read).

2. Distribution of Income:



The above diagram shows the distribution of income in Massachusetts. Each center of circle resides on the corresponding longitude and latitude of that zip code location it represents. The larger the diagram, the higher average income in that area.

3. Combining Income and Crime Together:



The above diagram shows crime number in terms of income.

Conclusion:

Taking a first look at the relation diagram, we can see that there is no linear relationship between income and number of crime.

Furthermore, we computed correlation coefficient between income and crime number, which yields the result of -0.072 with p-value 0.720. This further indicates that there is barely any linear relation between them.

This Suggests from the data we obtained, we shouldn't assume low income means high crime number!

To Be Improved:

As the result may seem somehow counter-intuitive, we take a further step and analyze some limitations of our project that may affect the result we got.

1. Data can be normalized.

The income (x-axis) we used is average income in each zip code area. However, we did not manage to obtain an average number of crimes for each zip code area. This is because we need further information about population in each zip code areas. If the data are normalized, the variations between two variables may be smaller.

2. The location of a crime ≠ people living in that place encountered that crime.

To illustrate this point, an example could be that one individual living at Brookline went to eat at Cambridge and encountered criminal. This would be reported as a crime whose location is at Cambridge, but not at where that individual lives. This results in the potential limitation that the statistics we got for number of crimes for one location do not match the statistic for number of crimes people got living at that location.