Capstone Project - The Battle of Neighborhoods
Housing Sales Prices and Data Analysis of Boston

Sophie Xu

A. Introduction

Boston is one of the oldest cities in the United States where over **690,000** people live in the city and it has a population density of **14,345** people per square mile. It's the 10th largest city in the United States. Boston is sometimes called a "city of neighborhoods" because of the profusion of diverse subsections. The city is officially divided into 23 neighborhoods in total. I decided to use Boston in my project. However, the fact that districts are squeezed into an area of less than 90 square miles causes the city to have a very intertwined and mixed structure [1].

As you can see from the figures, Boston is a city with high population and density. Being such a crowed city leads the owners of shops and social gathering places in the city where the population is dense. When we think of it by the investor, we expect them to prefer both lower real estate cost and less intense business area. If we think of the city residents, they may want to live where lower real estate values are. At the same time, they may want to choose the district according to the social places density. However, it is difficult to obtain information that will guide either investors or residents in this direction.

B. Data

To consider the problems, we can list the data as below:

- The Boston house-price data of Harrison, D. and Rubinfeld, D.L.[2].

```
Data Set Characteristics:

    :Number of Instances: 506

    :Number of Attributes: 13 numeric/categorical predictive

    :Median Value (attribute 14) is usually the target

    :Attribute Information (in order):
        - CRIM     per capita crime rate by town
        - ZN       proportion of residential land zoned for lots over 25,0
00 sq.ft.
        - INDUS    proportion of non-retail business acres per town
        - CHAS     Charles River dummy variable (= 1 if tract bounds river
; 0 otherwise)
        - NOX      nitric oxides concentration (parts per 10 million)
        - RM       average number of rooms per dwelling
        - AGE      proportion of owner-occupied units built prior to 1940
        - DIS      weighted distances to five Boston employment centres
```

```
        - RAD      index of accessibility to radial highways
        - TAX      full-value property-tax rate per $10,000
        - PTRATIO  pupil-teacher ratio by town
        - B        1000(Bk - 0.63)^2 where Bk is the proportion of blacks
by town
        - LSTAT    % lower status of the population
        - MEDV     Median value of owner-occupied homes in $1000's
```

- I found the corrected Boston housing data source [3]. `That file contains the Harrison and Rubinfeld (1978) data corrected for a few minor errors and augmented with the latitude and longitude of the observations.`

I used Foursquare API to search for venues of Chinese restaurants within certain distances of Boston.

C. Methodology

I used GitHub repository in my study. My initial data which has the main components crime rate, proportion of residential land & of non-retail business acres, average # of rooms per dwelling, age, distances, etc. It has 13 numeric/predictive attributes and Median Value as last attribute (attribute 14)/target. It has 506 samples.

I used **histogram** below to visualize pieces of the data.

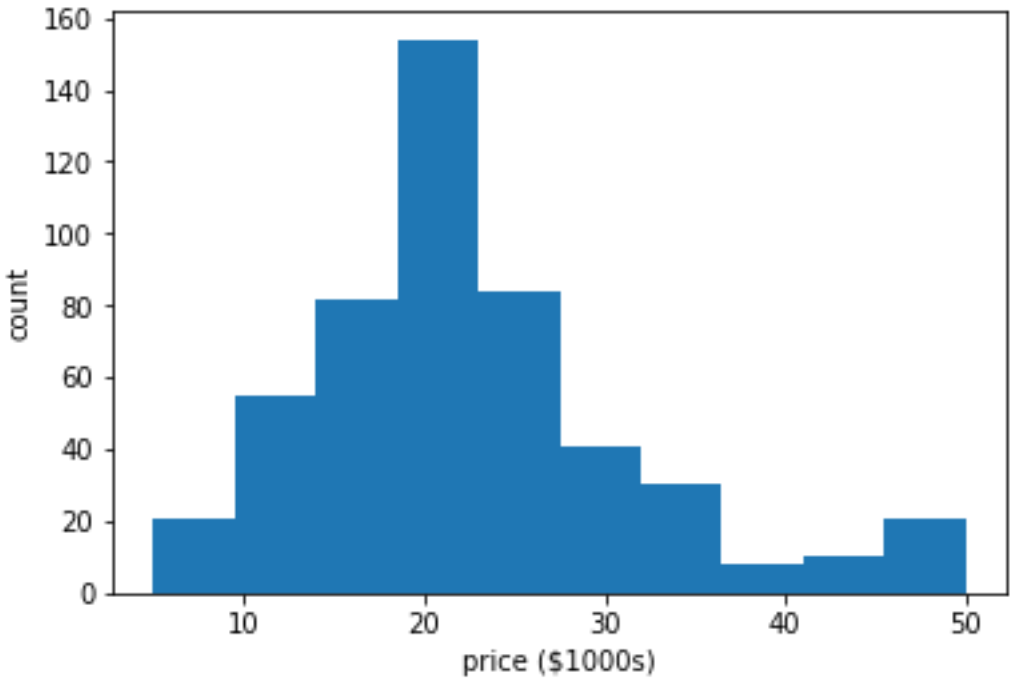

Figure 1: histogram of the target values: the median price in each neighborhood

In addition to the initial data, the corrected Boston housing data source augmented with the borough, latitude and longitude, of the observations.

I used python folium library to visualize geographic details of Boston and its boroughs and I created a map of Boston with boroughs superimposed on top.

I utilized the Foursquare API to explore the boroughs and segment them.

D. Results

In addition to the initial data, the corrected Boston housing data source augmented with the borough, latitude and longitude, of the observations. It has 92 boroughs and 506 neighborhoods.

E. Discussion

As I mentioned before, Boston is a big city with a high population density in a narrow area. The total number of measurements and population densities of the 92 districts in total can vary. As there is such a complexity, very different approaches can be tried in clustering and classification studies. Moreover, it is obvious that not every classification method can yield the same high quality results.

I also performed data analysis through this information by adding the coordinates of districts as static data on GitHub. In future studies, these data can also be accessed dynamically from specific platforms or packages.

I ended the study by visualizing the data on the Boston map. In future studies, web or telephone application can be carried out to direct investors.

F. Conclusion

As a result, people are turning to big cities to start a business or work. For this reason, people can achieve better outcomes through their access to the platforms where such information is provided.

Not only for investors but also city managers can manage the city more regularly by using similar data analysis types or platforms.

G. References:

[1] Boston – Wikipedia

https://en.wikipedia.org/wiki/Boston

[2] The Boston house-price data of Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.

[3] http://lib.stat.cmu.edu/datasets/boston_corrected.txt

[4] Foursquare API