An Analysis and Recommendation of Better Places to Live in Boston

May 30, 2019

# Introduction

## 1.1 Background

What makes a great place to live and how do we decide that? Is health care most important, or affordable housing? Which matters more, commute times, climate or crime rate?[1] Different people have different preferences and different perspective for their societies. Here, I have tried to narrow down my friend’s preferences and tried to analyze better places to reside in Boston.

## 1.2 Problem

A friend of mine is looking to move and settle down in Boston. Boston, being one of the oldest cities in the United States, it is most populous city of the Commonwealth of Massachusetts. Its rich history attracts many tourists and its many colleges and universities make it an international center of higher education [2]. A friend of mine is looking for a proper neighborhood to reside in Boston. But he has some priorities for the location. His preferences are as follows:

* His younger boy needs special-needs education. He is looking for a neighborhood which has more number of special schools, so that he can choose the nearest or the best among these schools for his kid.
* Neighborhood with low criminality rate, which is safe for them and their kids.
* City which has lots of restaurants in town, as he used to work as a chef in a Chinese restaurant and he wants to continue the same in Boston.

But his first priority is the area with more number of special schools.

## 1.3 Target audience

The target audience for this project are:

* Potential movers, who are willing to settle down in Boston.
* Parents focusing on the schools for their children around Boston.
* Real estate builders and planners who can decide what kind of neighborhoods are more attractive to open their apartments.
* Potential researcher on the diverse information about venues in the neighborhoods of Boston.
* All the students, curious to learn python’s pandas dataframe, python’s visualization tools and data visualization using python libraries.

# Data Acquisition and Cleaning

## 2.1 Data Description

Boston has been chosen as the city of target because, it is one of the oldest city in the United States and the dataset required for our analysis are readily available in the web. The datasets used for the analysis of the problem are as follows:

1. [Crimes in Boston](https://www.kaggle.com/ankkur13/boston-crime-data)
2. [Boston public Schools data](https://www.kaggle.com/crawford/boston-public-schools)
3. [Boston Police Districts geo location data](https://data.boston.gov/dataset/police-districts)
4. [Boston Neighborhoods geo location data](https://data.boston.gov/dataset/boston-neighborhoods)
5. [Boston Police Districts and Neighborhoods served](https://en.wikipedia.org/wiki/Boston_Police_Department)

## 2.2 Data Collection

The datasets collected are mostly from Airbnb’s [Kaggle portal](https://www.kaggle.com/datasets?search=boston) and geo location data are from official [sources](https://data.boston.gov/dataset/boston-neighborhoods) of Boston.

The crime dataset includes the crimes reported and recorded at the Boston Police Districts in the year 2018 and the Boston public schools dataset includes the name, address, school type of the public schools in Boston for the year 2018-2019. Those datasets from Kaggle portal are in CSV format which can be readily read into Panda’s dataframe. But for my ease, I just downloaded the CSV datasets into my local folder and read into Panda’s dataframe, as shown in Figure:1 and Figure:2.

The geo location data available are in the JSON format as shown in Figure:3, is used to render Boston city’s map with boundaries according to the Police Districts and according to Neighborhoods in Boston.

List of Boston Police Districts (BPDs), the Neighborhoods served under each police district, the latitude and longitude of each BPD and each Neighborhoods were not easily found in a single dataset. So, these data are entered manually in an excel sheet. The latitude, longitude for each BPD and each Neighborhood are calculated using the online geographical lat-long converter tool: <https://www.latlong.net/convert-address-to-lat-long.html>. This excel data is later read into a Panda’s dataframe.

In order to explore the venues in every Neighborhood and to segment and cluster the Neighborhoods around Boston, I have used the Foursquare API.

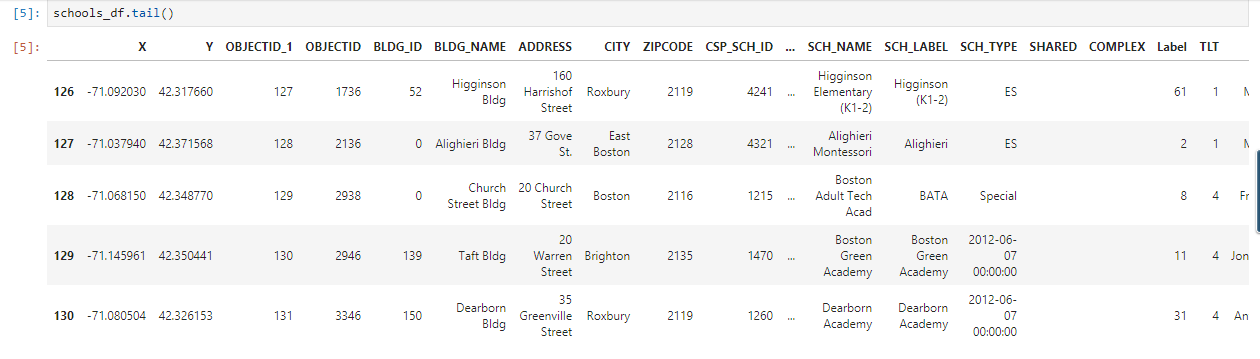


Figure 1: Dataframe showing Boston public schools data

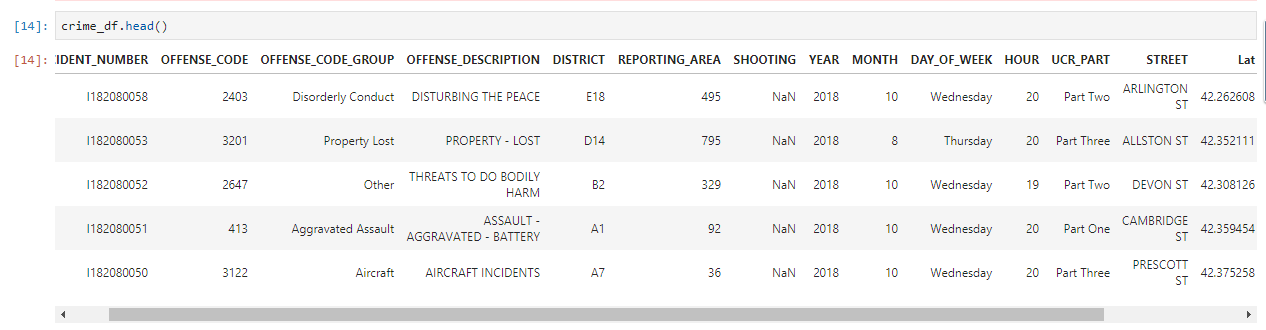


Figure 2: Dataframe showing Boston's crime data



Figure 3: JSON data showing the geo location of BPDs.

## Data Preparation

Data collected are from different sources. In the Crime data CSV, the crime reported in the Boston Police Districts are recorded, while in the Public Schools CSV, the schools are listed according to the Neighborhoods in Boston. To link those two dataset, an excel sheet with list of Boston Police Districts (BPDs), the Neighborhoods served under each police district, the latitude and longitude of each BPD and each Neighborhoods are entered manually in an excel sheet.

The Crime data CSV consisted of a column with garbage value (values as ######). These columns have been removed from the dataset. Also in the Public Schools dataset, some of the schools’ school type are missing and instead some date value are inserted. Those entries have also been removed from the analysis.

## References

1. <https://livability.com/best-places/top-100-best-places-to-live/2018>
2. <https://en.wikipedia.org/wiki/Boston>