Live in Boston

'The battle of Neighborhoods'

IBM Apllied Data Science CAPSTONE

Claudia Cerasale

05/11/2020

INDEX

- Introduction
  - Business Understanding

- Data
  - Data Cleaning
  - Data Manipulation
  - Data Anlysis & Data Visualization

- Methodology
  - Foursquare
  - K-Means

- Results and Discussions

- Conclusions

1. Introduction:

How safe is Boston and which is the most liveable neighborhood? First let's start with some considerations.

From the polls, the 27% of surveyed as of end of 2018 said they considered Boston to fairly or very unsafe. According to recent surveys the city's low murder rate per capita (9 per 100.000) could likely be a contributing factor to this perception. The city also has the second-lowest incidence of vehicle theft at a rate of 258 per 100.000 and fares well on the other crimes rates.

Is it true to say that Boston was more violent than New York and Seattle, but less violent than Chicago and Las Vegas, according to numbers from the FBI, based on crimes committed back in 2015. In 12/21/18 nationally, Boston ranked 14 out of 50 according to US News. After digging up and analysing the data what we found out that in recent numbers from the Boston Police Department, or BPD, show that violent crime, as well as property crime, has continued to drop, and has been steadily dropping for past years.

1.1 Business Understanding:

Public safety is vital to public health and happiness and the safety of a state can be a crucial factor in deciding where to move with the family, or more simply for those who need to move to another city for study purposes or work, but more generally it can also be useful for a tourist to know which is the safest neighborhood where to stay. In fact, going around the various forums, it can easily be seen that the questions are often addressed to the safety of the neighborhood as well as which interesting places to visit. So the project aims to find the safest and most comfortable neighborhoods in Boston, both through the analysis of the crime rate in the various neighborhoods and through the evaluation of the best served areas for traveling, for example, or for buying food, college presence etc. .

Therefore we will select the safest district of Boston based on the total of crimes, we will explore the various neighborhoods to find the 10 most common venues in

each neighborhood and finally we will group the neighborhoods using k-mean clustering, identifying the most suitable neighborhood for the needs of the individual.

Crime statistics will provide an overview of this problem.

Having outlined our problem, the factors, therefore, that will influence our decision are:

1. The total number of crimes committed in each district in a year (2017)
2.The most common venues in each selected safest neighborhood.

To extract / generate the requested information the following data sources will be required:

- Preprocessing of a Boston Police Department (BPD) dataset showing Boston crimes from 2016 to 2018.
- Conversion of District Codes into the names of their neighborhoods, with web scraping (Beatiful Soup Libraries)
- Distinction between crimes and accidents, and consequent filtering of the types of crimes
- Creation of a new dataset of Boston's safest neighborhoods and generation of their coordinates: the neighborhood coordinates will be developed using the geocoding of the Google Maps API
- Foursquare to identify the most common venues
- Clustering (K-Means)

2. Data

The dataset used for this study is 'Crime Incident Report', extracted from kaggle.

The dataset ‘Crime Incident Report’ is an open data initiative program led by the Boston government to document the initial details surrounding an incident to which Boston Police Department (BDP) officers respond.

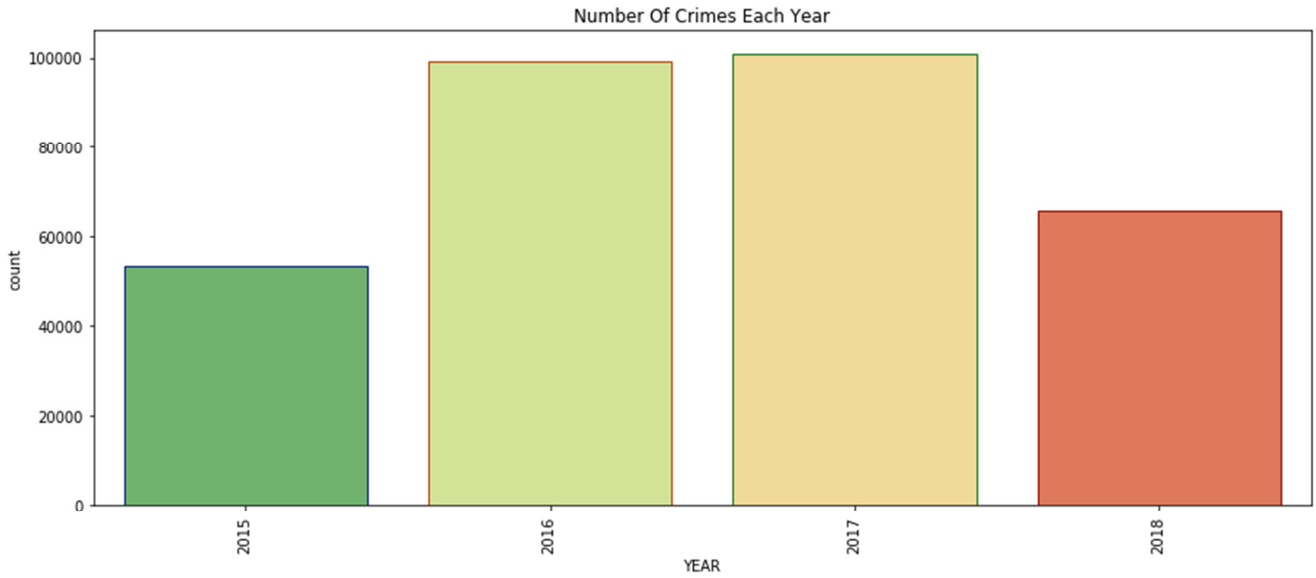
The dataset contains records from the Boston government’s new crime incident report system, which includes a reduced set of fields focused on capturing the type of incident as well as when and where it occurred. The Boston government took an initiative to improve the city of Boston by releasing its data sources to the public. Over the past few decades, the way we look at the field of climate, genetics, sports, have been altered dramatically due to big data technology advancements; similarly, the way crime data was traditionally held by law enforcement agencies has als changed, crime prediction is a niche trend in this era.

The dataset begins from August 2015 to December 21st 2018, there are 349073 incidents and 17 variables; ranging from types of offense, reported area and reporting area, date occurred, street, and the latitude and longitude of the incident.

	INCIDENT_NUMBER	OFFENSE_CODE	OFFENSE_CODE_GROUP	OFFENSE_DESCRIPTION	DISTRICT	REPORTING_AREA	SHOOTING	OCCURRED_ON_DATE	YEAR	MC
0	I182070945	619	Larceny	LARCENY ALL OTHERS	D14	808	NaN	2018-09-02 13:00:00	2018	
1	I182070943	1402	Vandalism	VANDALISM	C11	347	NaN	2018-08-21 00:00:00	2018	
2	I182070941	3410	Towed	TOWED MOTOR VEHICLE	D4	151	NaN	2018-09-03 19:27:00	2018	
3	I182070940	3114	Investigate Property	INVESTIGATE PROPERTY	D4	272	NaN	2018-09-03 21:16:00	2018	
4	I182070938	3114	Investigate Property	INVESTIGATE PROPERTY	B3	421	NaN	2018-09-03 21:05:00	2018	

2.1 Data Cleaning

After a first look at the data, to evaluate its size, basic statistics and basic information regarding the fields of the dataset, I made a first graph to get an idea of the change in the rate of crimes over the years.



For the purpose of our project, I decided to eliminate some columns not strictly necessary. Furthermore, the "SHOOTING" column has also been eliminated, due to the high presence of missing values. Only the data relating to the year 2017 was filtered, the most recent year but also fully monitored.

Here is what the dataset looks like after the first phase of the cleaning process:

	OFFENSE_CODE_GROUP	DISTRICT	YEAR	MONTH	Lat	Long
0	Fraud	D4	2017	11	42.339268	-71.072088
1	Investigate Property	D14	2017	10	42.346501	-71.136128
2	Property Lost	B2	2017	8	42.331925	-71.084112
3	Fraud	D14	2017	6	42.338482	-71.152890
4	Property Lost	B3	2017	12	42.279591	-71.081323

The columns that I intended to be useful for my purpose are: 'Offense Code Group', which gives us an indication of the types of crimes, 'District', the district where they occurred, ‘Latitude and Longitude’, coordinates that will be useful for Foursquare and the crime mapping, and finally ‘Year’ and ‘Month’.

2.2 Data Manipulation

At this stage, with the use of Beatiful Soup, the scraping library, I convert the codes of the police districts into their respective neighborhood names.

The information was obtained from the site: "https://bpdnews.com/districts". It should be noted that although the Boston City Neighborhood Services Office has designated 23 neighborhoods in the city, there are 12 police districts in the dataset, probably because some districts may refer to multiple neighborhoods or simply do not part of the crime register program.

In any case, following the data and after various manipulations, I obtain the following data frame, with the districts and neighborhoods of reference, representative of the area:

	District	Neighborhoods
0	A7	EastBoston
1	B2	Roxbury
2	B3	Mattapan
3	C6	SouthBoston
4	C11	Dorchester
5	D4	SouthEnd
6	D14	Brighton
7	E5	WestRoxbury
8	E13	JamaicaPlain
9	E18	HydePark
10	A1	DowntownCharlestown
11	A15	DowntownCharlestown

This data will be combined with the original data set, through a join on 'District', obtaining the following dataframe, functional for the analysis we intend to face:

	Group	District	Year	Month	Lat	Long	Neighborhoods
0	Fraud	D4	2017	11	42.339268	-71.072088	SouthEnd
1	Fraud	D4	2017	11	42.353940	-71.078030	SouthEnd
2	Larceny	D4	2017	12	42.352569	-71.079788	SouthEnd
3	Recovered Stolen Property	D4	2017	12	42.352569	-71.079788	SouthEnd
4	Counterfeiting	D4	2017	10	42.342528	-71.076789	SouthEnd
5	Auto Theft	D4	2017	12	42.335968	-71.081299	SouthEnd
6	Larceny	D4	2017	11	42.343144	-71.095893	SouthEnd
7	Harassment	D4	2017	6	42.337002	-71.082276	SouthEnd
8	Harassment	D4	2017	3	42.337002	-71.082276	SouthEnd
9	Motor Vehicle Accident Response	D4	2017	11	42.341386	-71.080826	SouthEnd

Below is the distribution of crimes by neighborhoods:

```
Roxbury                14391
Dorchester             12732
DowntownCharlestown    11997
SouthEnd               11930
Mattapan               10622
SouthBoston             6557
Brighton                6134
HydePark                5407
JamaicaPlain            5148
WestRoxbury             4143
EastBoston              4075
Name: Neighborhoods, dtype: int64
```

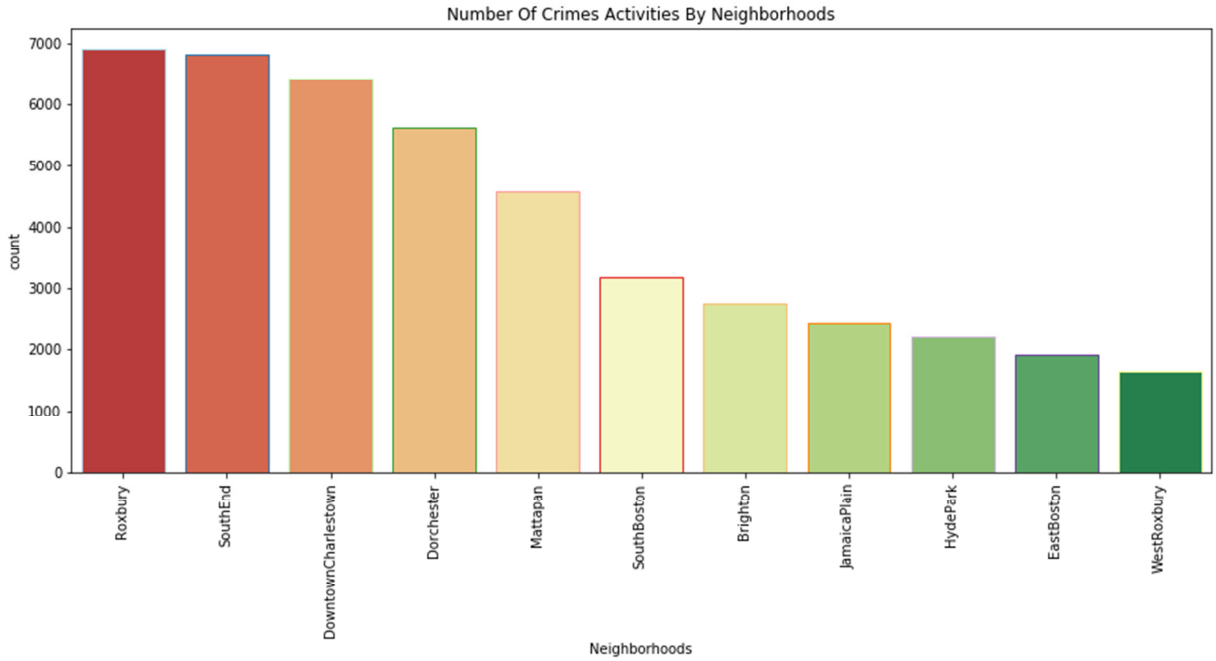
Looking at the types of crimes reported in the dataset, we realize that not all of them are crimes, such as: 'Property Lost' or 'Property Found', etc...

As I intend to evaluate the safest areas of Boston where I can move, I have considered selecting only the crimes that can be a problem for personal security, obtaining a dataset with dimensions (44420, 9), and a reassessment of the total number of crimes by neighborhood :

```
Roxbury                6892
SouthEnd               6803
DowntownCharlestown    6424
Dorchester             5612
Mattapan               4581
SouthBoston            3173
Brighton               2744
JamaicaPlain           2442
HydePark               2195
EastBoston             1913
WestRoxbury            1641
Name: Neighborhoods, dtype: int64
```

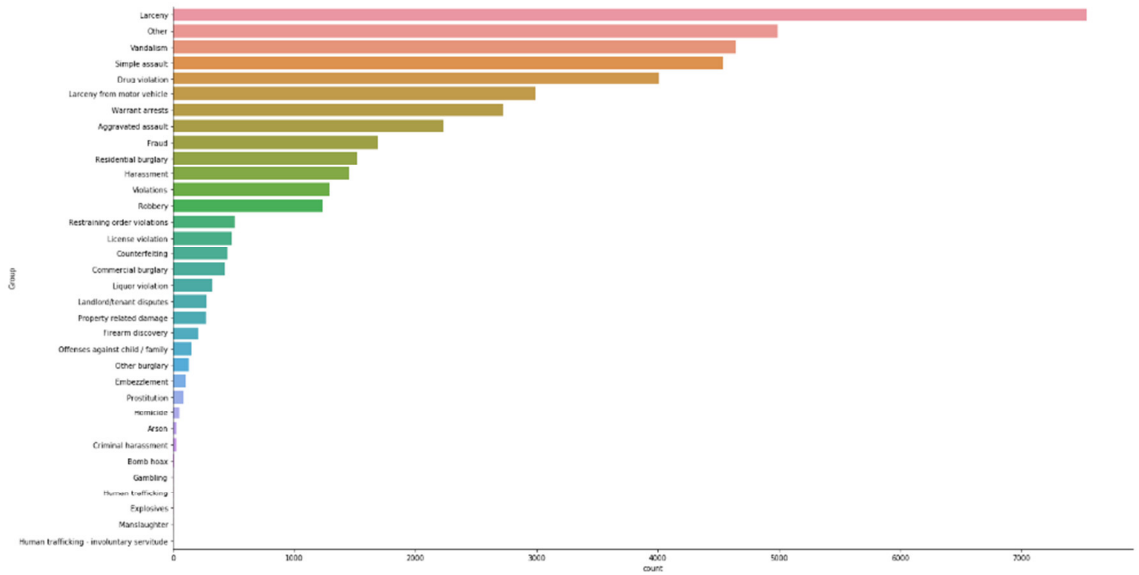
2.3 Data Visualization & Data Analysis

Number Of Crimes Activities By District:



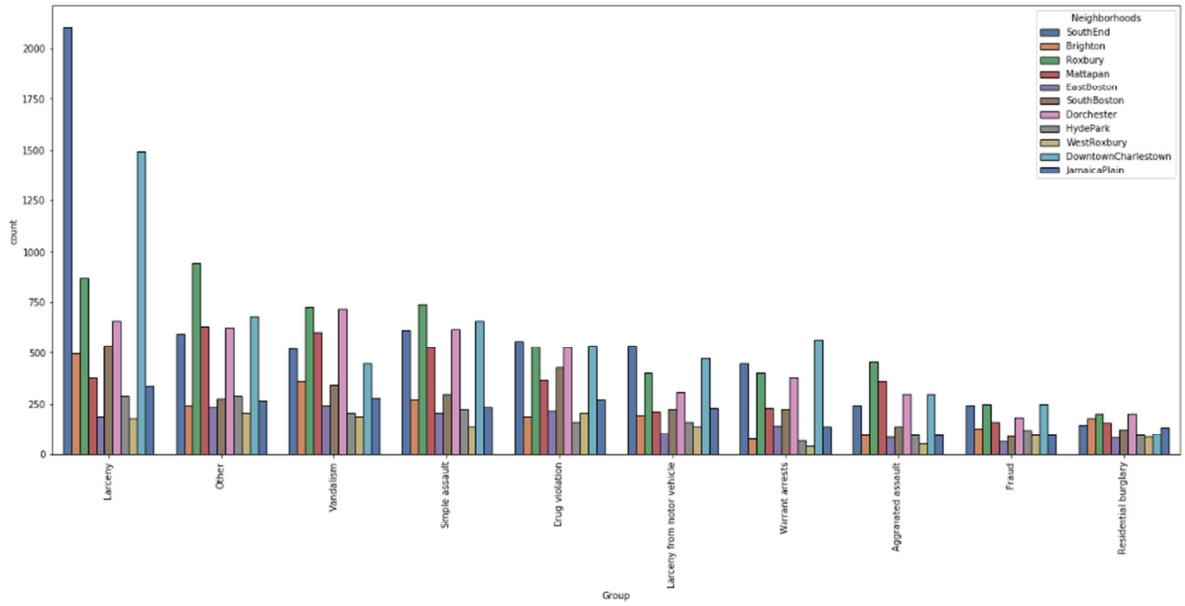
In this graph, we organize data by neighborhood based on observation of the occurrence of crimes in each of them. With the date ordered, we can see and conclude that the neighborhood with the greatest crime episodes is Roxbury, followed respectively by SouthEnd, DownTown/Charlestown, Dorchester and Mattapan.

Distribution of the typology of crimes:



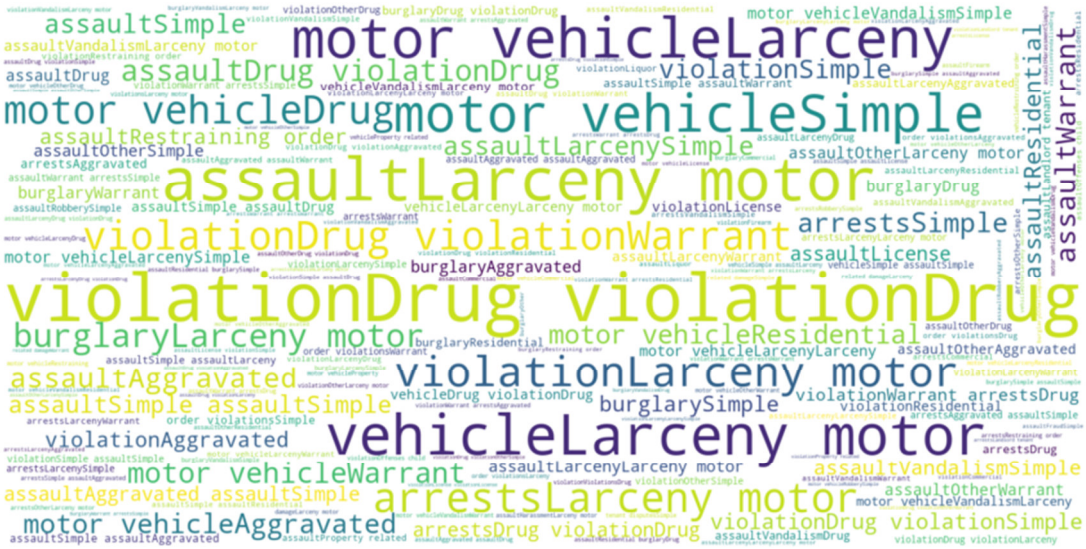
The 5 most frequent types of crimes, having previously filtered the data for crimes and excluding accidents, appear to be: Larceny, Other, Vandalism, Single Assault and Drug Violation.

Relationship between type of crimes and Neighborhoods:

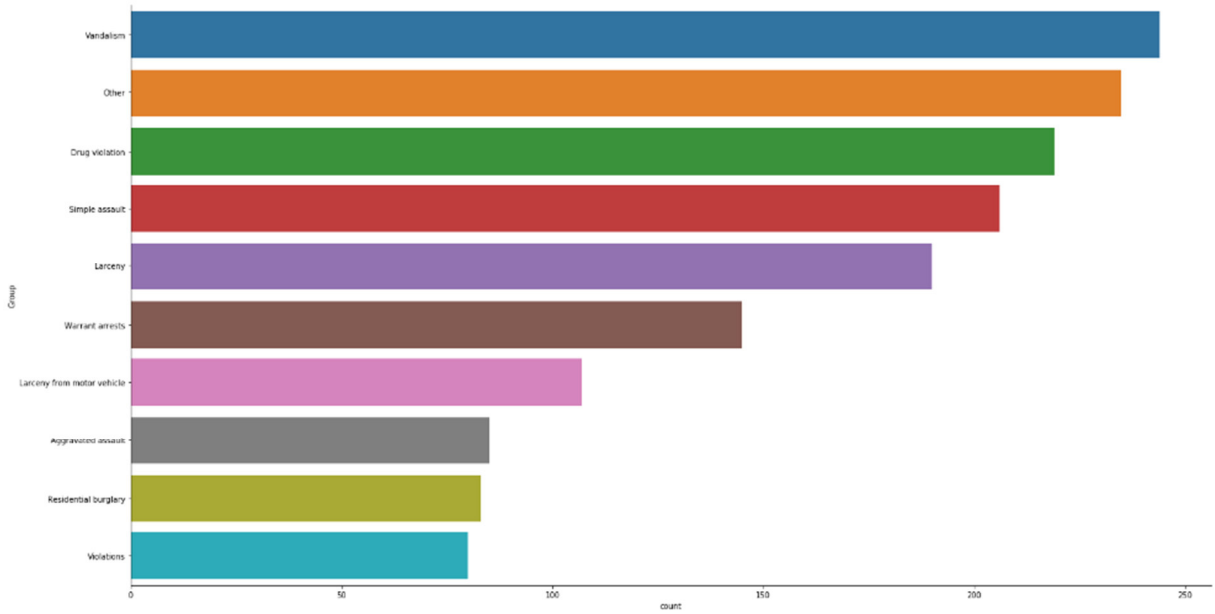


In this graph we observe the distribution of the various types of crimes by neighborhood. We observe how the crime type 'larceny' has a high concentration in SouthEnd and, immediately followed, by DowntownCharleston. Furthermore, West Roxbury is confirmed as the safest neighborhood in Boston on this chart.

Wordcloud representation of the most common types of crime in Boston. Words that appear larger give us an immediate indication of what they are:



Let's take a look in particular at the "West Roxbury" neighborhood, which turns out to be the neighborhood with the lowest crime rate compared to the others. Let's see what types of crimes take place in this neighborhood, and I mean the purpose of creating a new dataframe with only data concerning a West Roxbury.



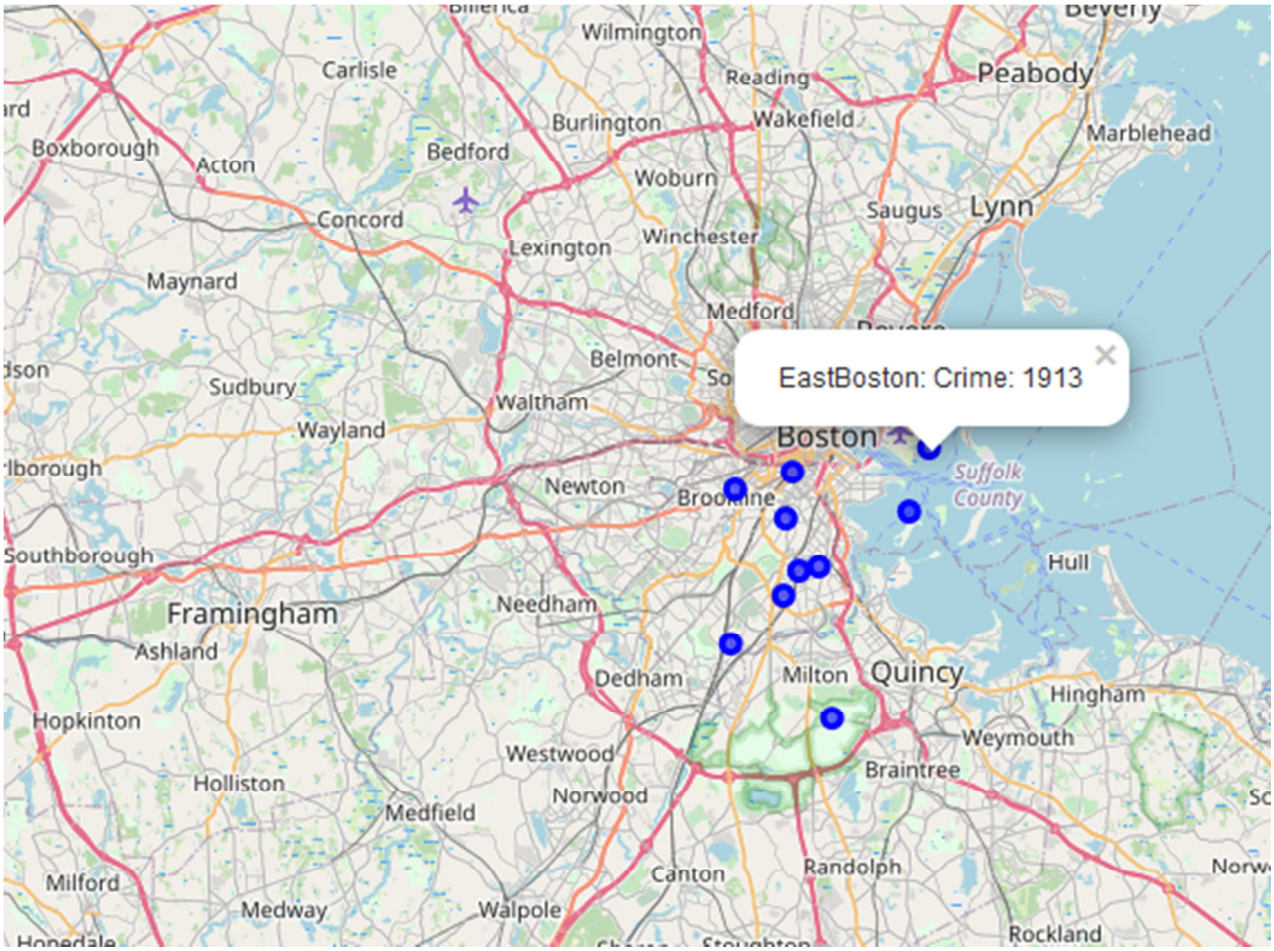
The top 5 most common types of crimes in West Roxbury are: Vandalism, Other,

Simple Assault and Larceny. Note that Larceny is a type of crime very common throughout Boston.

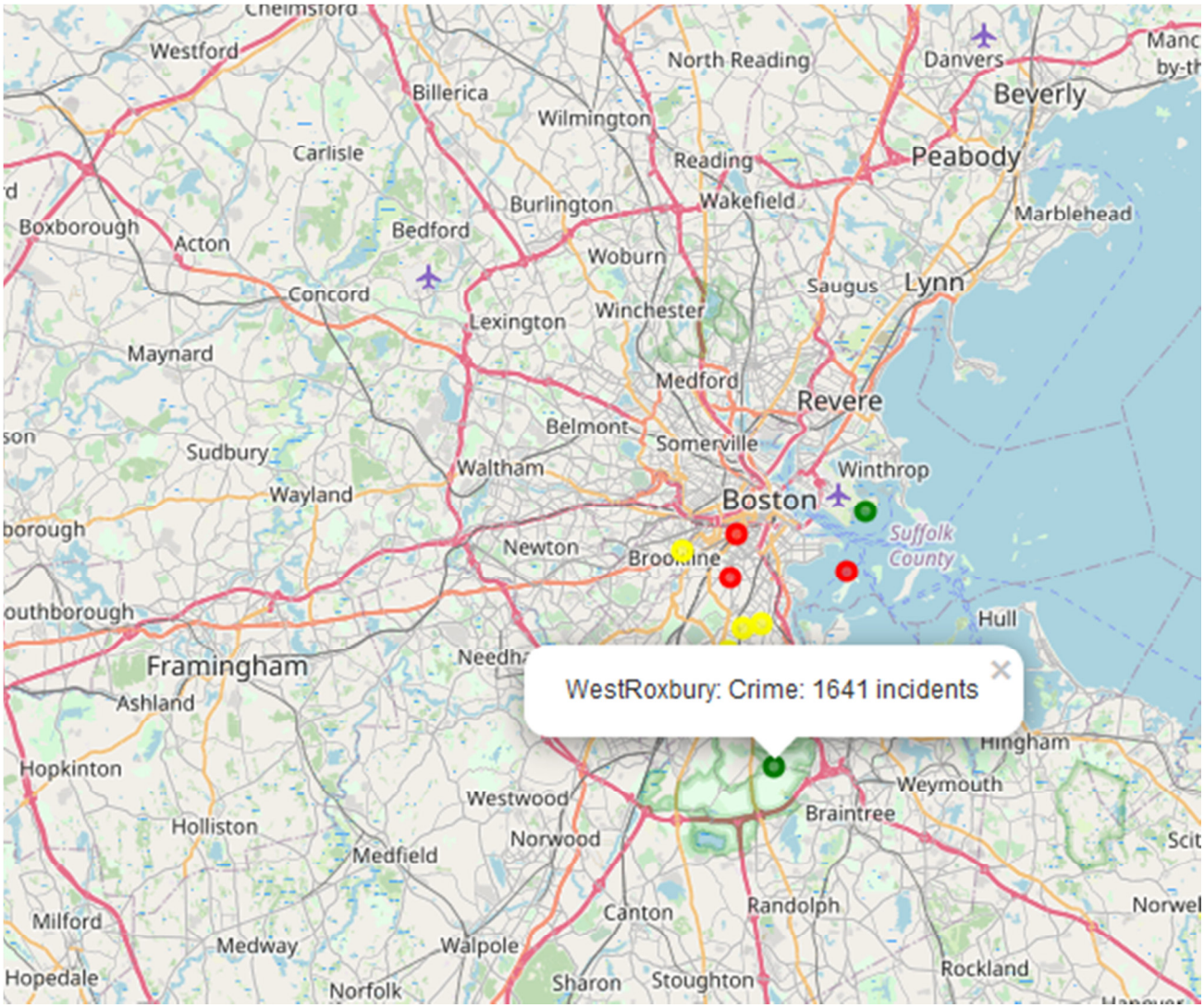
Below is the total number of the top 5 types of the West Roxbury neighborhood:

Vandalism	244
Other	235
Drug violation	219
Simple assault	206
Larceny	190

Thanks to the Geopy library, we display the map of Boston, initially marking only the neighborhoods:

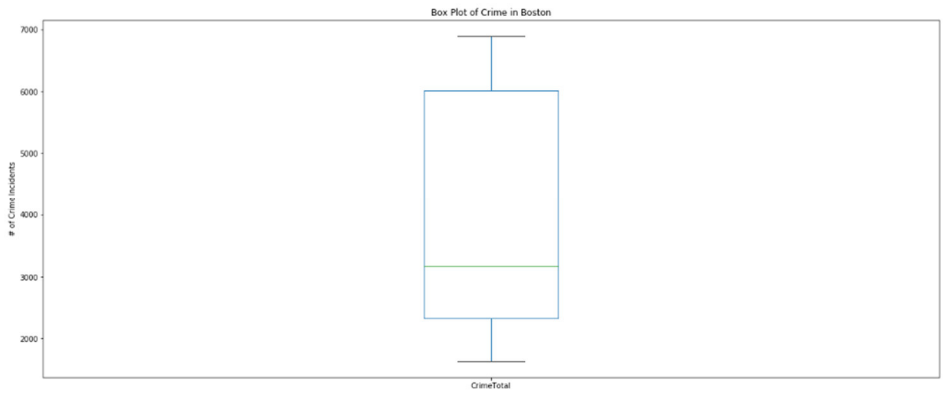


Below is a map distinguishing for safer or less safe neighborhoods based on the total of crimes:



The red markers are the areas with a higher crime rate, while the green ones with low intensity, and the yellow markers we consider them intermediate values.

This first distinction by neighborhood was made considering the interquartile range.



3. Methodology: Foursqaure & Clustering

Foursquare:

Although the previous processing steps may also fall under the "Methodology", I preferred to distinguish the data manipulation and exploration operations in the "Data" section.

Having obtained the final data set on which we will work in this phase, we now intend to find out which of these neighborhoods has services and places of greatest interest for those who intend to move or simply have a peaceful holiday and with the possibility of moving easily from where they are staying.

Through Foursquare we will try to identify the most common venues of the various neighborhoods first by concentrating, first of all on West Roxbury, as it has proved to be the safest neighborhood compared to the others.

Later, by crossing these new data with the number of crimes, we will evaluate whether our first intuition on the safest and livable neighborhood will be confirmed.

For West Roxbury, Foursquare, in this case returned, returned 4 venues:

	name	categories	lat	lng
0	Blue Hills Overlook	Scenic Lookout	42.226305	-71.060998
1	Granite link golf club	Golf Course	42.229141	-71.051724
2	Blue Hills Reservation Pond	Fishing Spot	42.228282	-71.050894
3	Chickatawbut Hill	Trail	42.225689	-71.061120

Let's see what returns for all the other neighborhoods; so we get a ranking of the top 10 venues for each neighborhood:

	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Brighton	Italian Restaurant	Gym	Pizza Place	Chinese Restaurant	Coffee Shop	Dessert Shop	Sandwich Place	Library	Pub	Playground
1	Dorchester	Vietnamese Restaurant	Pizza Place	Sandwich Place	Seafood Restaurant	Chinese Restaurant	Café	Supermarket	Donut Shop	Pub	Fast Food Restaurant
2	EastBoston	Airport Terminal	Diner	Dog Run	Donut Shop	Dry Cleaner	Electronics Store	Fast Food Restaurant	Fish Market	Fishing Spot	Food
3	HydePark	Home Service	Supermarket	Donut Shop	Cosmetics Shop	Pizza Place	Mobile Phone Shop	Yoga Studio	Dry Cleaner	Electronics Store	Fast Food Restaurant
4	JamaicaPlain	Pizza Place	Market	Fried Chicken Joint	Platform	Donut Shop	Dry Cleaner	Shoe Store	Southern / Soul Food Restaurant	Caribbean Restaurant	Liquor Store

Considering, however, specifically, the people who have to move, and therefore live in Boston for work or study needs, I carried out a more detailed search by choosing some categories of funeral venues for a longer period of stay, such as supermarket, gyms, bus stop to be able to go without the aid of the car to the places of interest.

The categories I have chosen, which in the future can be expanded, are the following:

Park	Library	Pool	Playground	Cinemas	Gym	College & University	School	Supermarket	Metro Station	Bus Stop

And here is the final dataset:

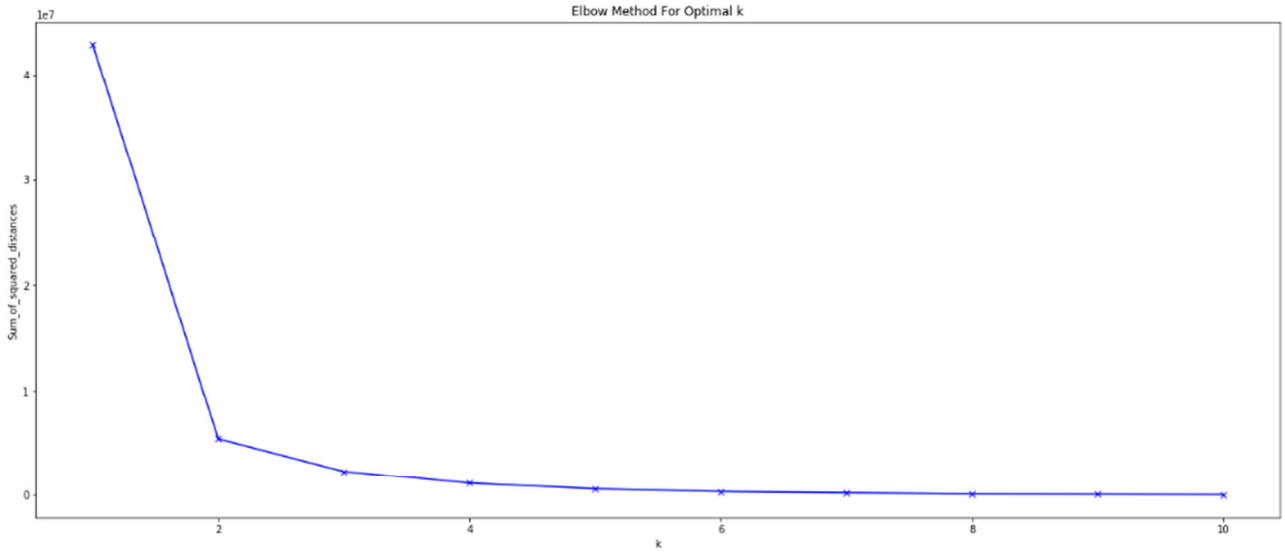
	Neighborhoods	Lat	Long	CrimeTotal	Park	Library	Pool	Playground	Cinemas	Gym	College & University	School	Supermarket	Metro Station	Bus Stop
7	Roxbury	42.322237	-71.084382	6892	6	1	0	1	0	1	12	17	1	2	2
9	SouthEnd	42.343495	-71.080143	6803	0	0	0	0	0	0	0	0	0	0	0
2	DowntownCharlestown	42.324842	-71.006856	6424	1	0	0	0	0	0	0	0	0	0	0
1	Dorchester	42.299940	-71.063290	5612	41	27	5	18	1	46	49	31	3	12	20
6	Mattapan	42.286373	-71.085469	4581	21	11	10	18	0	17	50	32	3	16	13
8	SouthBoston	42.156097	-70.766414	3173	3	1	0	4	0	3	1	12	3	0	0
0	Brighton	42.334868	-71.116300	2744	9	1	0	4	0	2	16	26	0	1	0
5	JamaicaPlain	42.297375	-71.076617	2442	0	0	0	0	0	0	0	0	0	0	0
4	HydePark	42.264184	-71.119224	2195	2	1	0	2	0	1	2	10	1	0	0
3	EastBoston	42.353989	-70.995139	1913	9	1	1	7	0	3	11	21	1	3	1
10	WestRoxbury	42.229512	-71.055121	1641	0	0	2	0	0	0	0	0	0	0	0

Clustering (K-Means):

To better understand the results and draw the right conclusions, we use clustering, in particular k-means. K-Means is a technique for partitioning / classifying data without internal structures or labels, dividing them into K non-overlapping subsets, precisely called "clusters".

In the previous maps we have identified three groups, based on the total number of crimes by neighborhood. So we choose K = 3 (therefore 3 groups).

But, verifying the optimal number of clusters with the elbow method, we obtain that the optimal nr of K is 2. In fact, as we see from the graph, the putno where the rate of decrease shifts abruptly is 2.



4. Results & Discussions

The first result obtained by analyzing the crime rate is that the safest neighborhood is **"West Roxbury":**

	clusterLabel	label	Neighborhoods	Lat	Long	CrimeTotal
0	0	0	WestRoxbury	42.229512	-71.055121	1641
1	0	0	EastBoston	42.353989	-70.995139	1913
2	0	0	HydePark	42.264184	-71.119224	2195
3	0	0	JamaicaPlain	42.297375	-71.076617	2442
4	0	0	Brighton	42.334868	-71.116300	2744
5	0	0	SouthBoston	42.156097	-70.766414	3173

The second result: the most suitable for a transfer to Boston turns out to be, considering both the total number of crimes and the total number of venues, functional to a more comfortable everyday life, it turns out to be "**East Boston**": second in our ranking for the total number of crimes and second for the total number of venues, but certainly an excellent compromise:

Out[301]:

Neighborhoods	Lat	Long	CrimeTotal	Park	Library	Pool	Playground	Cinemas	Gym	College & University	School	Supermarket	Metro Station	Bus Stop	Total_Venues
Brighton	42.334868	-71.116300	2744	9	1	0	4	0	2	16	26	0	1	0	59
EastBoston	42.353989	-70.995139	1913	9	1	1	7	0	3	11	21	1	3	1	58
SouthBoston	42.156097	-70.766414	3173	3	1	0	4	0	3	1	12	3	0	0	27
HydePark	42.264184	-71.119224	2195	2	1	0	2	0	1	2	10	1	0	0	19
WestRoxbury	42.229512	-71.055121	1641	0	0	2	0	0	0	0	0	0	0	0	2
JamaicaPlain	42.297375	-71.076617	2442	0	0	0	0	0	0	0	0	0	0	0	0

5. Conclusions

Boston is definitely a city that has improved a lot over the years, (as can be seen in the graph of the distribution of crimes over the years, in which there is clearly a sharp decrease in crimes between the years 2017 and 2018), to the point of being considered from their citizens quite safe.

From our analysis and from the data in our possession, we identify two categories of neighborhoods:

- 'Good Neighborhoods': SouthBoston Brighton JamaicaPlain HydePark EastBoston WestRoxbury

- 'Bad Neighborhoods': Roxbury SouthEnd DowntownCharlestown Dorchester Mattapan

I remember that the types of crimes were selected and therefore a part of them was worked on. This choice was made to adapt the data to our purpose: that is, to find a place to move safely from a personal security point of view, but also because looking at the data as a whole, there is a clear prevalence of crime-type events, identified as 'Motor Vehicle Accident Response' and this would have led to false results focused on this result and, therefore, non-adherent and significant to achieve our purpose.

Therefore to all those tourists who ask for a safe area of Boston where to stay I would recommend **'West Roxbury'**. But for those who intend or need to move for a longer time I think there must be a good compromise between safety and comfort

(understood as ease of movement, nearby supermarkets, but also as the presence of places for leisure), therefore I would recommend **'East Boston**'. Finally, the project could later be expanded by also including data on the cost and rents of houses.