1. 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.

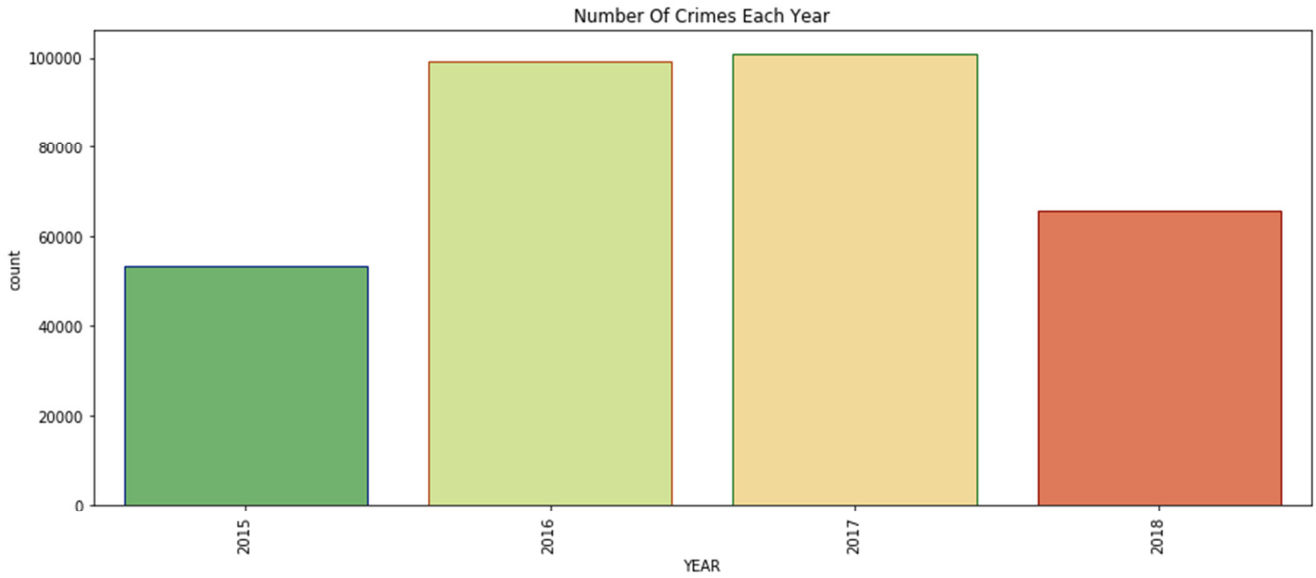
The dataset contains records from the Boston government’s new crime incident report system, which includes a reduced set of fields focused on capturing the type of incident as well as when and where it occurred. The Boston government took an initiative to improve the city of Boston by releasing its data 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
```

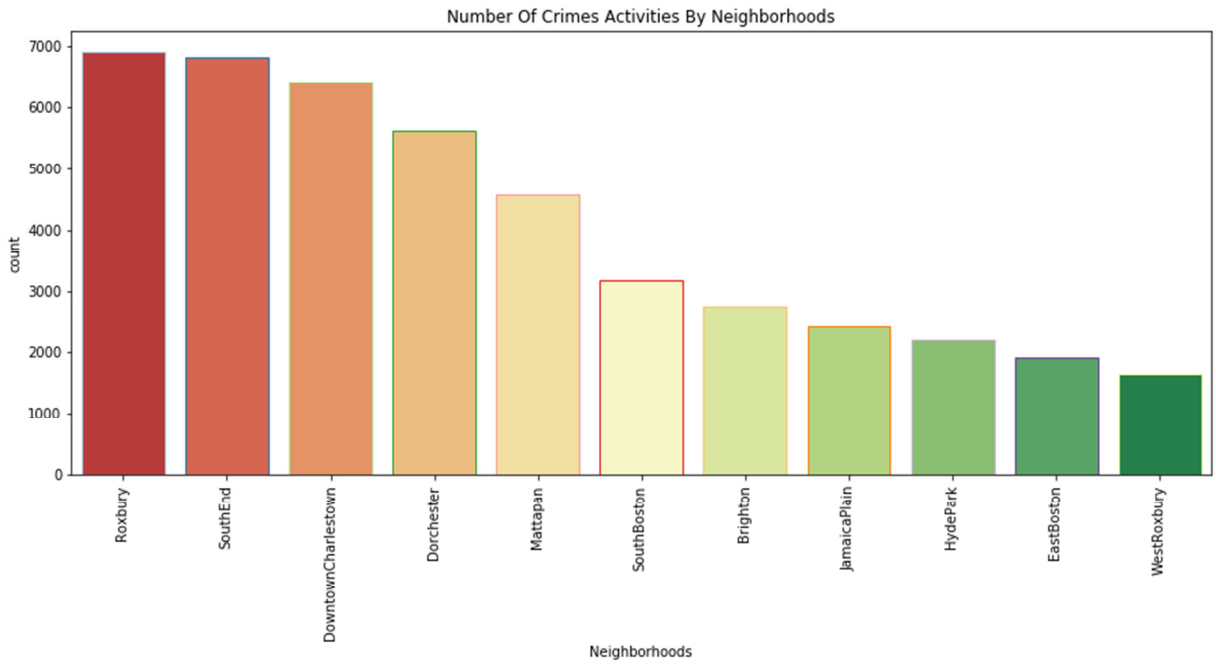
Looking at the types of crimes reported in the dataset, we realize that not all of them are crimes, such as: 'Property Lost' or 'Property Found', etc...

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
```

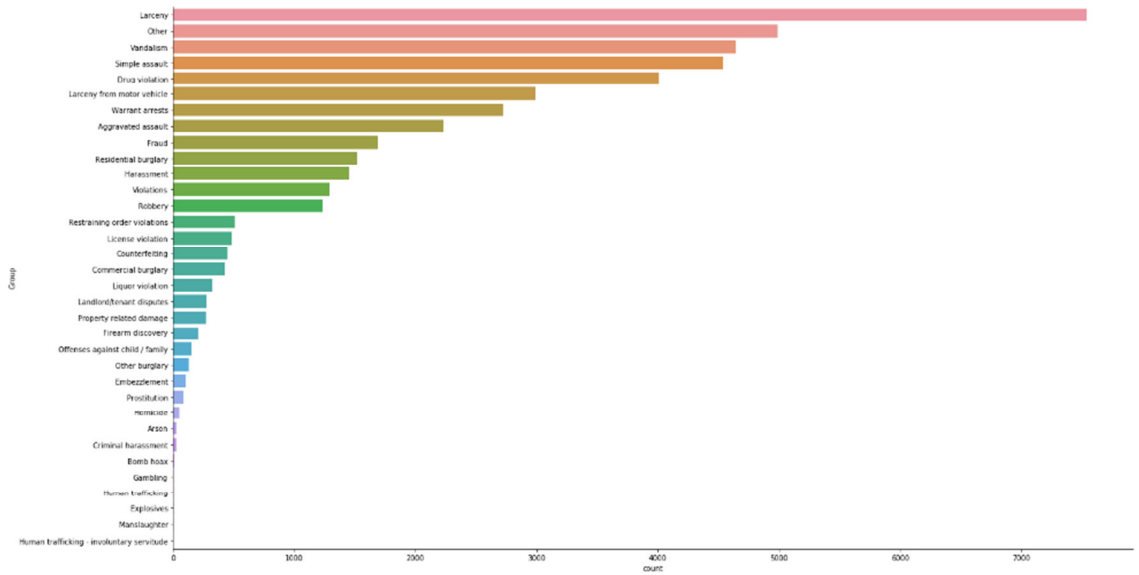
2.3 Data Visualization & Data Analysis

Number Of Crimes Activities By District:



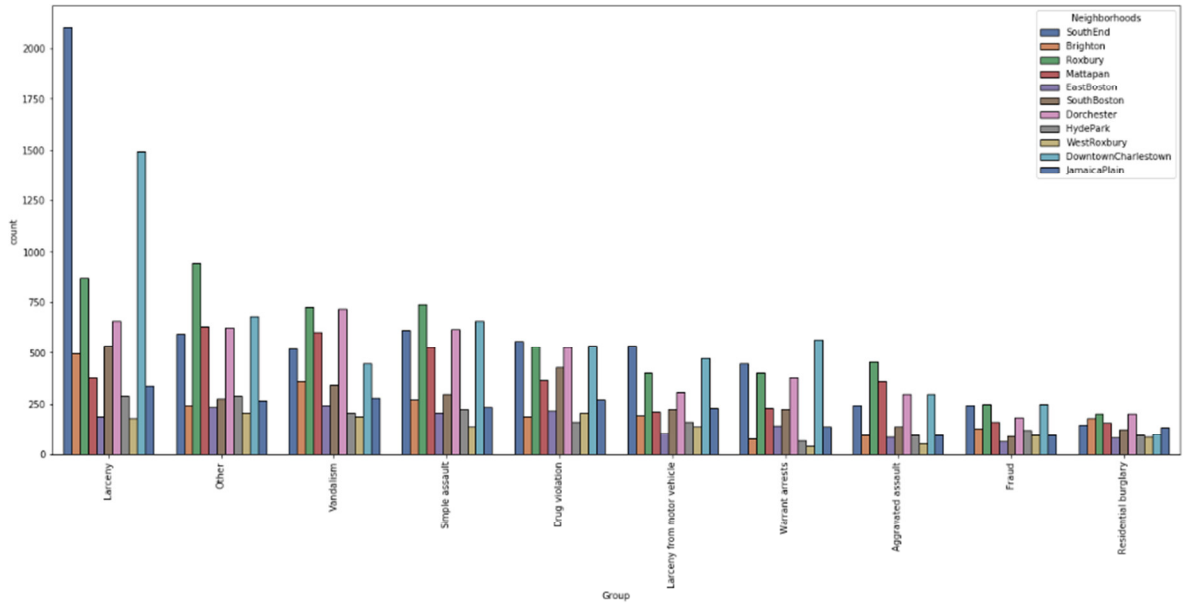
In this graph, we organize data by neighborhood based on observation of the occurrence of crimes in each of them. With the date ordered, we can see and conclude that the neighborhood with the greatest crime episodes is Roxbury, followed respectively by SouthEnd, DownTown/Charlestown, Dorchester and Mattapan.

Distribution of the typology of crimes:



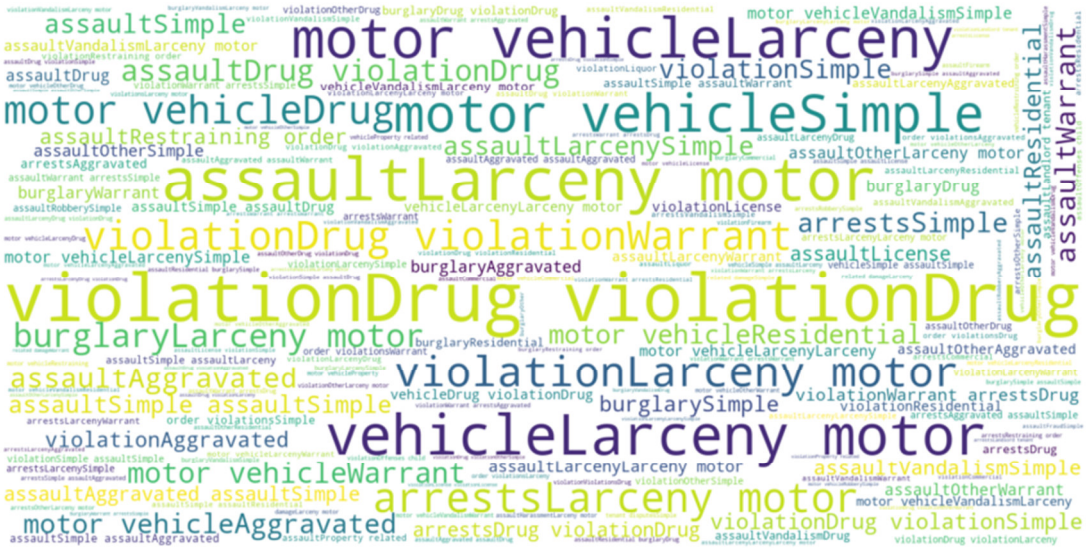
The 5 most frequent types of crimes, having previously filtered the data for crimes and excluding accidents, appear to be: Larceny, Other, Vandalism, Single Assault and Drug Violation.

Relationship between type of crimes and Neighborhoods:

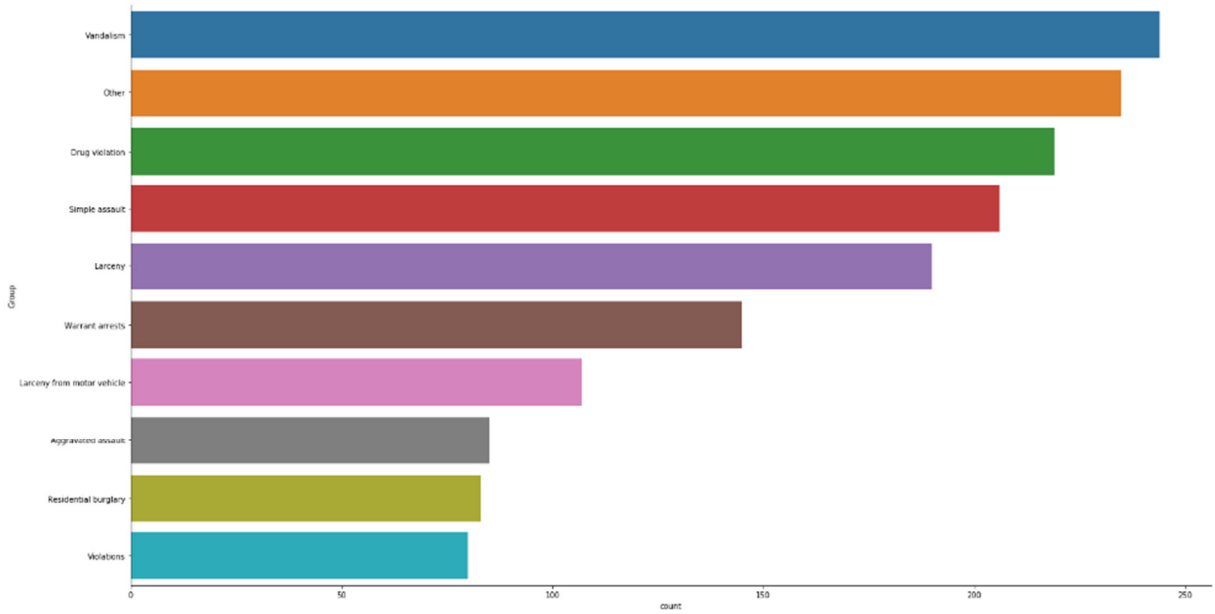


In this graph we observe the distribution of the various types of crimes by neighborhood. We observe how the crime type 'larceny' has a high concentration in SouthEnd and, immediately followed, by DowntownCharleston. Furthermore, West Roxbury is confirmed as the safest neighborhood in Boston on this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



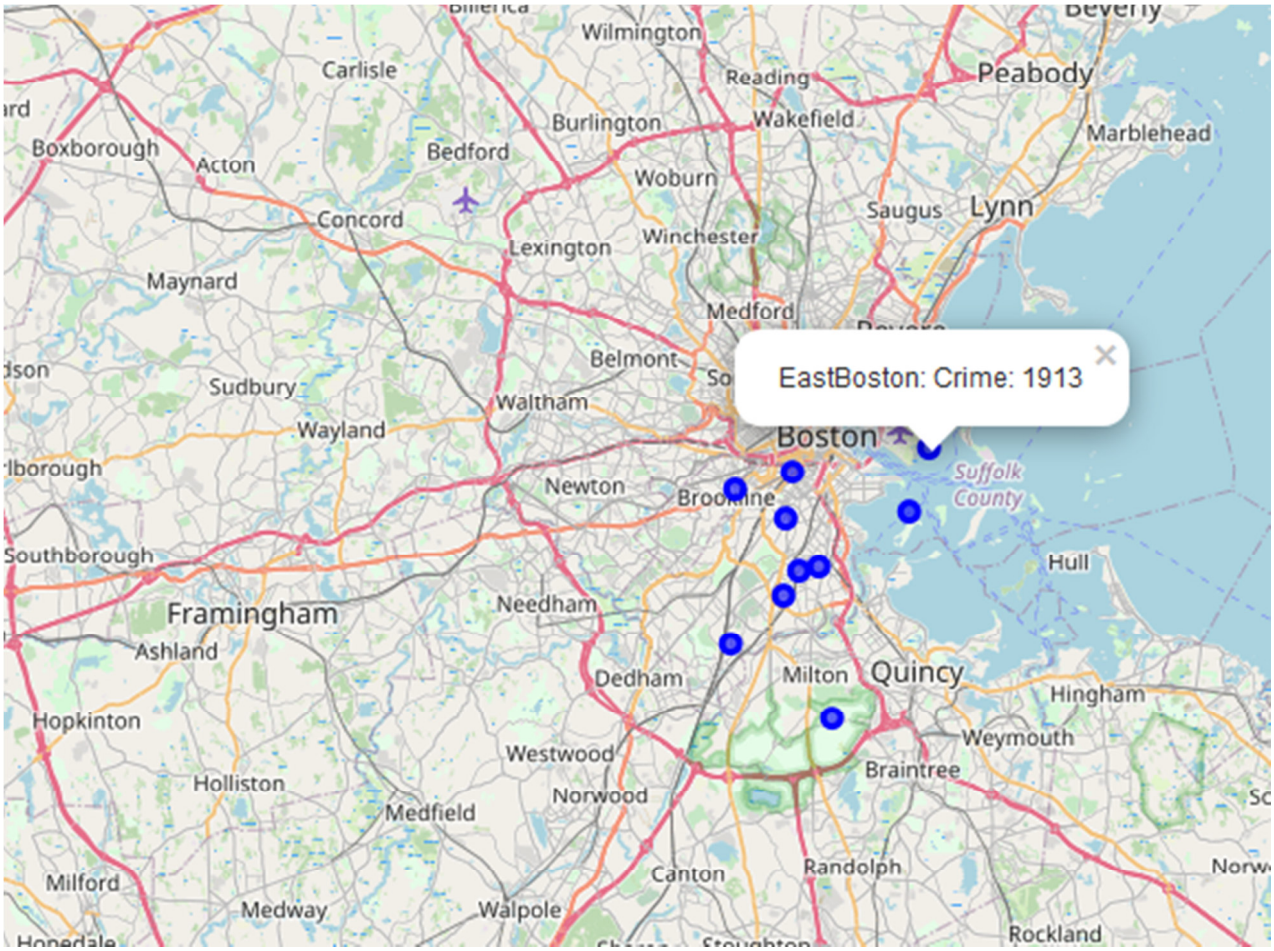
The top 5 most common types of crimes in West Roxbury are: Vandalism, Other,

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

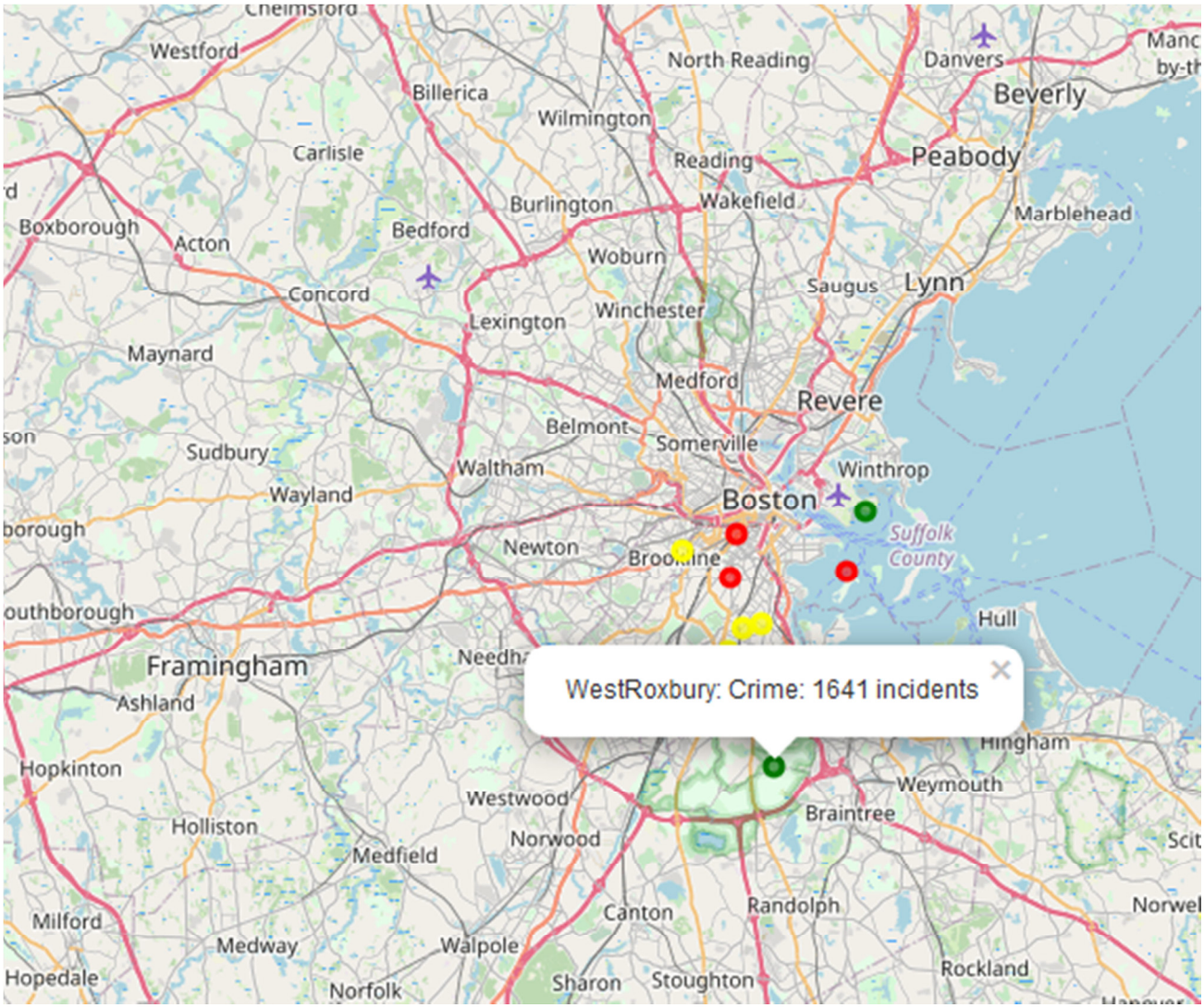
Below is the total number of the top 5 types of the West Roxbury neighborhood:

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.