Crime Coded Venues

Look before you head to your favorite venue

Published on February 27, 2019 by **Abhishek Bhansali**

Table of contents

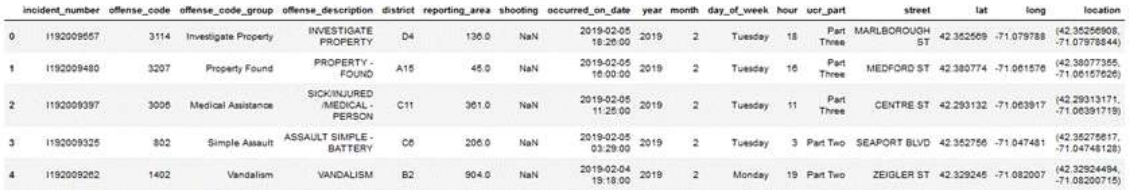
1. Introduction: Problem

The question that we are trying to answer here is: How safe are we when we visit one of the popular venues in Boston, MA? The answer to this question can not only help individuals secure their personal safety but also help them in being better prepared to face an unforeseen unfortunate event. But beyond this obvious benefit, the answers to this question can also help in identifying the concentration of types of crimes around venues and this in turn can be utilized by law enforcement agencies in evolving an effective strategy to curb crimes. In addition the business owners of such venues can utilize the information to devise better business strategies by exactly knowing what types of crimes occur around them more frequently.

2. Data

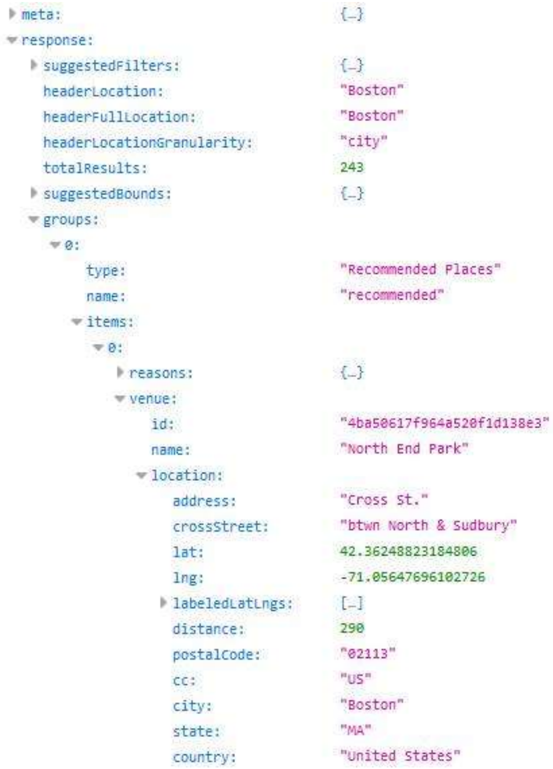
2.1 Data sources

To start answering our question we need data for venues and crime incidents in Boston. Crime incident data is provided by Boston Police Department (BPD). This dataset is updated regularly and contains records which includes set of fields focused on capturing the type of incident as well as when and where it occurred. The data set has information of crime incidents from year 2015 to date. As of February, 2019 the dataset contains 364,577 entries



	incident_number	offense_code	offense_code_group	offense_description	district	reporting_area	shooting	occurred_on_date	year	month	day_of_week	hour	ucr_part	street	lat	long	location
0	I192009567	3114	Investigate Property	INVESTIGATE PROPERTY	D4	136.0	NaN	2019-02-05 18:26:00	2019	2	Tuesday	18	Part Three	MARLBOROUGH ST	42.352569	-71.079788	(42.35256908, -71.07978844)
1	I192009480	3207	Property Found	PROPERTY - FOUND	A15	45.0	NaN	2019-02-05 16:00:00	2019	2	Tuesday	16	Part Three	MEDFORD ST	42.380774	-71.061576	(42.38077355, -71.06157626)
2	I192009397	3006	Medical Assistance	SICK/INJURED /MEDICAL - PERSON	C11	361.0	NaN	2019-02-05 11:25:00	2019	2	Tuesday	11	Part Three	CENTRE ST	42.293132	-71.063917	(42.29313171, -71.06391719)
3	I192009325	802	Simple Assault	ASSAULT SIMPLE - BATTERY	C6	206.0	NaN	2019-02-05 03:29:00	2019	2	Tuesday	3	Part Two	SEAPORT BLVD	42.352756	-71.047481	(42.35275617, -71.04748128)
4	I192009262	1402	Vandalism	VANDALISM	B2	904.0	NaN	2019-02-04 19:18:00	2019	2	Monday	19	Part Two	ZEIGLER ST	42.329245	-71.082007	(42.32924494, -71.08200715)

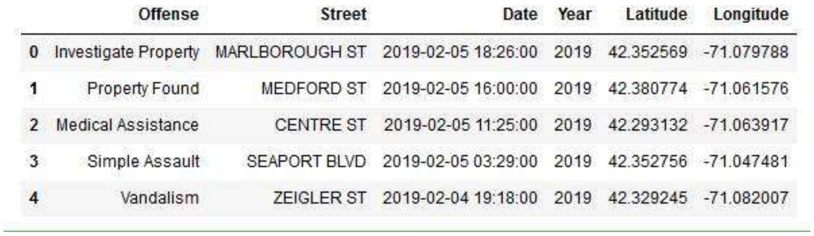
The venue data is aggregated by leveraging the Foursquare[1] location data. The explore end point of Foursquare API, *https://api.foursquare.com/v2/venues/explore* returns a list of recommended venues near the current location expressed as latitude and longitude for that location. The return results are limited by the parameters in GET request such as *radius* which governs the radius to search within (in meters) and *limit* which governs the number of results returned. In addition Foursquare Developer Account Type and Type of API call (Regular or Premium) determine the daily call quota. Venue data is aggregated as *groups* field of json response.



```
▸ meta:                              {…}
▾ response:
  ▸ suggestedFilters:                {…}
    headerLocation:                  "Boston"
    headerFullLocation:              "Boston"
    headerLocationGranularity:       "city"
    totalResults:                    243
  ▸ suggestedBounds:                 {…}
  ▾ groups:
    ▾ 0:
        type:                        "Recommended Places"
        name:                        "recommended"
      ▾ items:
        ▾ 0:
          ▸ reasons:                 {…}
          ▾ venue:
              id:                    "4ba50617f964a520f1d138e3"
              name:                  "North End Park"
            ▾ location:
                address:             "Cross St."
                crossStreet:         "btwn North & Sudbury"
                lat:                 42.36248823184806
                lng:                 -71.05647696102726
              ▸ labeledLatLngs:      […]
                distance:            290
                postalCode:          "02113"
                cc:                  "US"
                city:                "Boston"
                state:               "MA"
                country:             "United States"
```

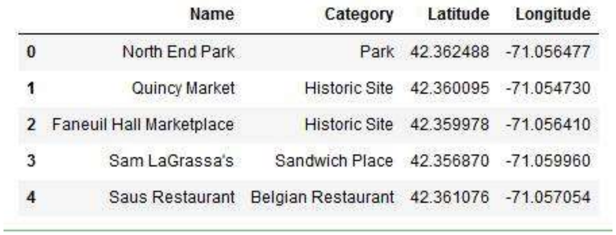
2.2 Feature Selection

The only relevant features of crime dataset are the ones that describe the type of crime, the place of occurrence, date and year when they occurred so we select *'offense_code_group', 'street', 'occurred_on_date', 'year', 'lat', 'long'* and drop rest of the features from the crime incident data provided by BPD. Subsequently the selected features are renamed as *'Offense', 'Street', 'Date', 'Year', 'Latitude', 'Longitude'* and a new Geo Spatial dataset of Boston's crime incidents ***bostonCIGeo_df*** is created.



	Offense	Street	Date	Year	Latitude	Longitude
0	Investigate Property	MARLBOROUGH ST	2019-02-05 18:26:00	2019	42.352569	-71.079788
1	Property Found	MEDFORD ST	2019-02-05 16:00:00	2019	42.380774	-71.061576
2	Medical Assistance	CENTRE ST	2019-02-05 11:25:00	2019	42.293132	-71.063917
3	Simple Assault	SEAPORT BLVD	2019-02-05 03:29:00	2019	42.352756	-71.047481
4	Vandalism	ZEIGLER ST	2019-02-04 19:18:00	2019	42.329245	-71.082007

Similarly, the relevant fields in the json response obtained from the explore endpoint of Foursquare API are venue's name, category and its latitude, longitude which are extracted to create Boston Venues dataset ***bv_df***.



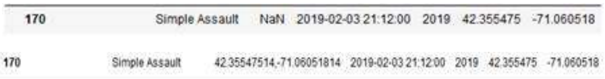
	Name	Category	Latitude	Longitude
0	North End Park	Park	42.362488	-71.056477
1	Quincy Market	Historic Site	42.360095	-71.054730
2	Faneuil Hall Marketplace	Historic Site	42.359978	-71.056410
3	Sam LaGrassa's	Sandwich Place	42.356870	-71.059960
4	Saus Restaurant	Belgian Restaurant	42.361076	-71.057054

These two datasets form the core data for further analysis. But before we can go any further we need to clean this data and remove inconsistencies.

2.3 Data Cleaning

It is observed that geo-spatial dataset of Boston's crime incidents has null values for Latitude and Longitude in many recorded incidents, but as part of our analysis, that I describe shortly, such values are required in ascertaining the vicinity of a reported incident to a given venue. Thus we have to discard all such incidents were Latitude or Longitude information is missing in geo-spatial dataset of Boston's crime incidents any attempt to substitute these values using statistical models would wrongly classify the occurrence of such incidents. There is at least a 5-10% information loss though in the process.

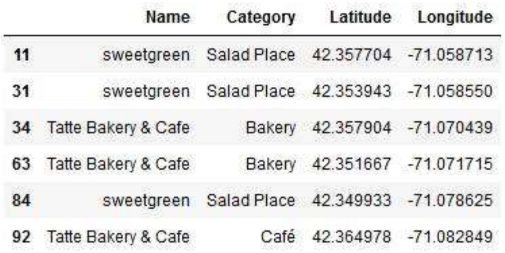
The street feature in geo-spatial dataset of Boston's crime incidents although not significant for analysis we intend to perform is shown on marker pop up of a crime incident plotted on the map which has markers for both venues and crime incidents. It is always useful when looking at a venue on the map so see what incident has occurred in vicinity of such venue and at what place however many of the reported incidents has no street information. It was decided to substitute the string Latitude, Longitude in place of missing values.



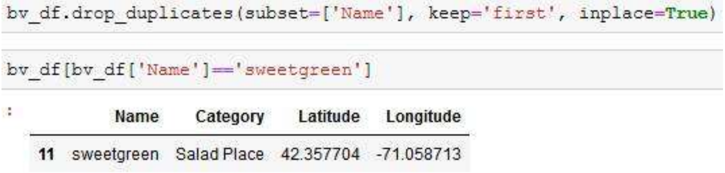
170	Simple Assault	NaN	2019-02-03 21:12:00	2019	42.355475	-71.060518
170	Simple Assault	42.35547514,-71.06051814	2019-02-03 21:12:00	2019	42.355475	-71.060518

The geo-spatial dataset of Boston's crime incidents was limited to incidents that have occurred in the year 2019 to make the analysis relevant to the current times. The intention of the analysis is to show the current scenario devoid of bias that any past data might create. A meaningful slice of the data should be selected in accordance with the objectives of the analysis.

The data obtained from Foursquare API contained duplicate values for venue names, this creates a negative bias towards venues in our analysis in the sense that they gain more weightage as being places in whose vicinity crimes are committed and thus the first occurrence of the repeated venue name was accepted and others dropped.



	Name	Category	Latitude	Longitude
11	sweetgreen	Salad Place	42.357704	-71.058713
31	sweetgreen	Salad Place	42.353943	-71.058550
34	Tatte Bakery & Cafe	Bakery	42.357904	-71.070439
63	Tatte Bakery & Cafe	Bakery	42.351667	-71.071715
84	sweetgreen	Salad Place	42.349933	-71.078625
92	Tatte Bakery & Cafe	Café	42.364978	-71.082849



```
bv_df.drop_duplicates(subset=['Name'], keep='first', inplace=True)
```

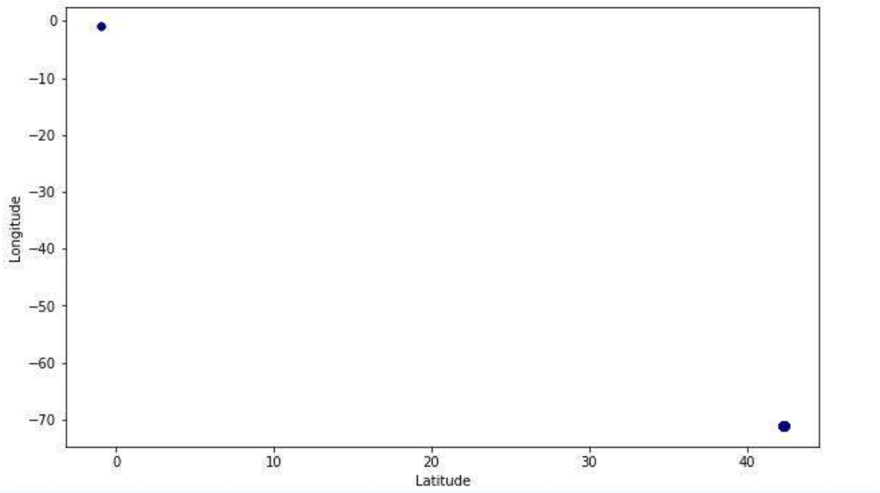
```
bv_df[bv_df['Name']=='sweetgreen']
```

:

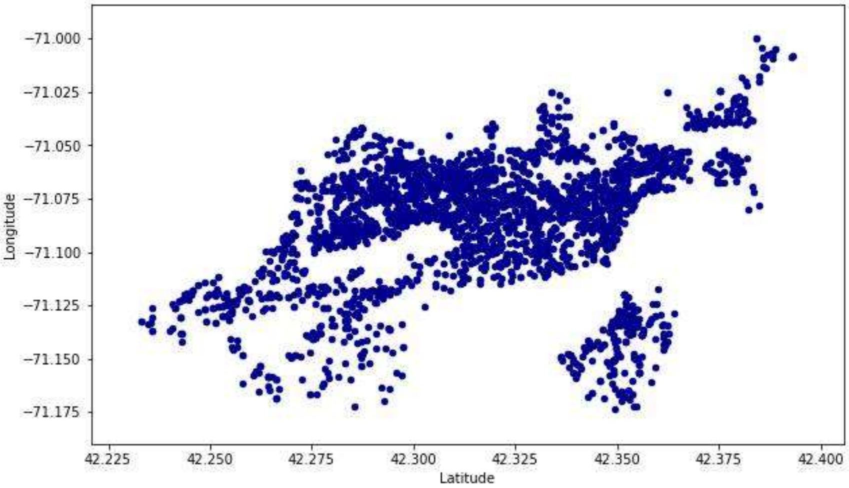
	Name	Category	Latitude	Longitude
11	sweetgreen	Salad Place	42.357704	-71.058713

3. Methodology

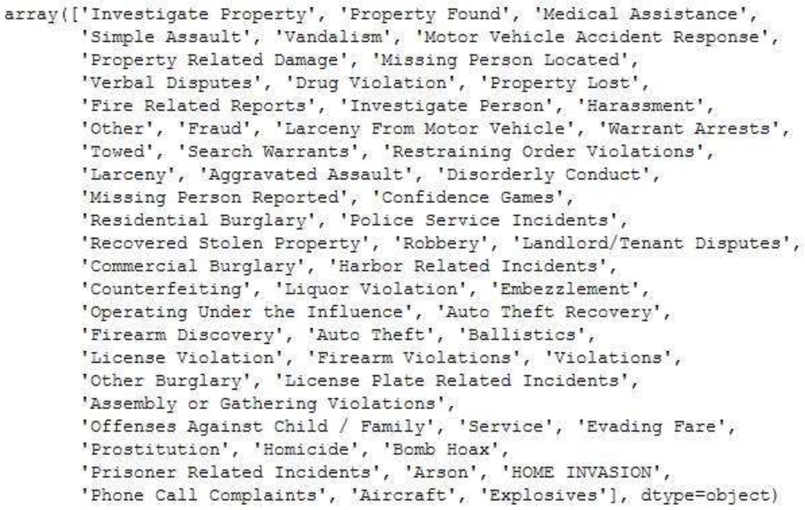
The moment to finally peek into the data has arrived, let's look at the distribution of Latitude and Longitudes in geo-spatial dataset of Boston's crime incidents using Scatter Plot and here you go – Surprise !!



The plot suggest outliers all of certain specific value and somewhere near (0,0). On further examination it is revealed as value of -1 for Latitude and Longitude in certain incident records. Let's drop these records as before and re-examine the Scatter Plot which now looks perfect



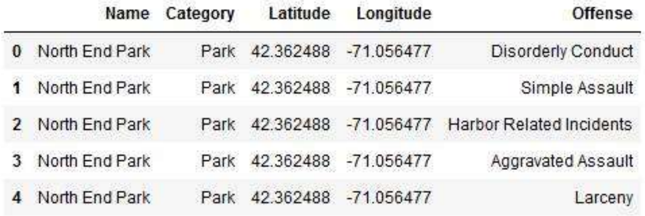
Next let us look at unique values of the feature ***Offense*** in



```
array(['Investigate Property', 'Property Found', 'Medical Assistance',
       'Simple Assault', 'Vandalism', 'Motor Vehicle Accident Response',
       'Property Related Damage', 'Missing Person Located',
       'Verbal Disputes', 'Drug Violation', 'Property Lost',
       'Fire Related Reports', 'Investigate Person', 'Harassment',
       'Other', 'Fraud', 'Larceny From Motor Vehicle', 'Warrant Arrests',
       'Towed', 'Search Warrants', 'Restraining Order Violations',
       'Larceny', 'Aggravated Assault', 'Disorderly Conduct',
       'Missing Person Reported', 'Confidence Games',
       'Residential Burglary', 'Police Service Incidents',
       'Recovered Stolen Property', 'Robbery', 'Landlord/Tenant Disputes',
       'Commercial Burglary', 'Harbor Related Incidents',
       'Counterfeiting', 'Liquor Violation', 'Embezzlement',
       'Operating Under the Influence', 'Auto Theft Recovery',
       'Firearm Discovery', 'Auto Theft', 'Ballistics',
       'License Violation', 'Firearm Violations', 'Violations',
       'Other Burglary', 'License Plate Related Incidents',
       'Assembly or Gathering Violations',
       'Offenses Against Child / Family', 'Service', 'Evading Fare',
       'Prostitution', 'Homicide', 'Bomb Hoax',
       'Prisoner Related Incidents', 'Arson', 'HOME INVASION',
       'Phone Call Complaints', 'Aircraft', 'Explosives'], dtype=object)
```

Some of these values such as *'Investigate Property', 'Property Found', 'Medical Assistance'* so on so forth are not actually criminal actions and their inclusion creates a negative bias for venues in whose vicinity they tend to get reported the records with such values for offense need to be dropped and once that is done we are actually set to start our analysis.

The analysis is based on the examination of all reported incidents in certain specific preset radius around a given venue. The specific preset chosen for present analysis is 500 mtrs. To find out if a reported incident falls within this specific preset radius around the given venue we use Python's GeoPy[2] *geopy.distance.distance* function to find the geodesic distance between them given their latitude and longitudes and in each case where the reported incident lies within this specific preset radius a new record is added to our new dataset ***bvc_df*** with following information Venue Name, Venue Category, Venue Latitude, Venue Longitude and reported Offense. The resulting new dataset now has frequency as well as crime type of all reported incidents in vicinity of a given venue for a given specific radius around it.

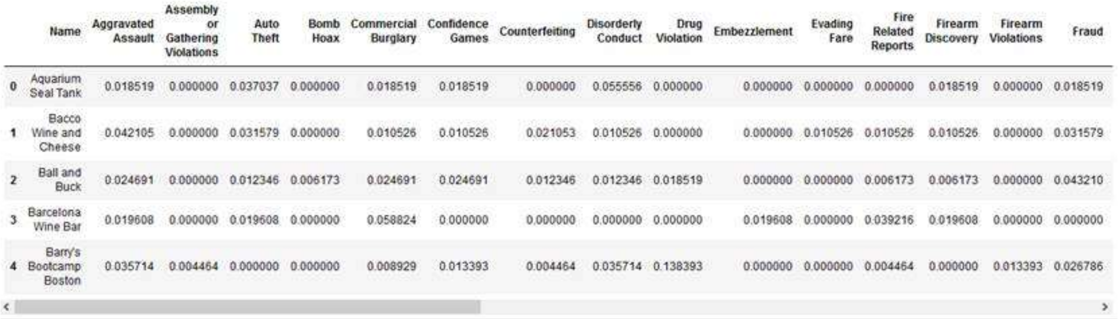


	Name	Category	Latitude	Longitude	Offense
0	North End Park	Park	42.362488	-71.056477	Disorderly Conduct
1	North End Park	Park	42.362488	-71.056477	Simple Assault
2	North End Park	Park	42.362488	-71.056477	Harbor Related Incidents
3	North End Park	Park	42.362488	-71.056477	Aggravated Assault
4	North End Park	Park	42.362488	-71.056477	Larceny

This data set could then be used to cluster venues based on the type of offenses that are reported in its vicinity using machine learning algorithms. The K-Means Clustering approach is best suited to cluster venues based on reported incidents in their vicinity.

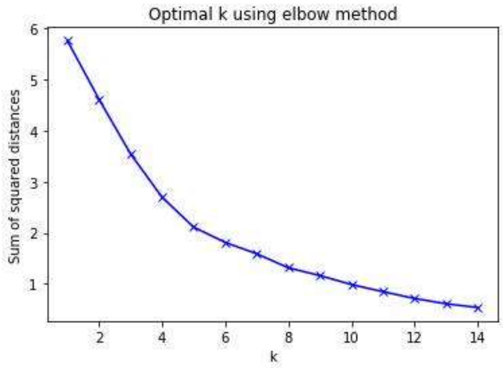
4. Analysis

The first step before we could actually utilize our new dataset ***bvc_df*** for K-Means Clustering is to convert the categorical variable *Offense* into a form that is required by K-Means Clustering package using one hot encoding. Next the dataset is grouped by venue names and a mean is obtained resulting in dataset ***bvconehotgrp_df***

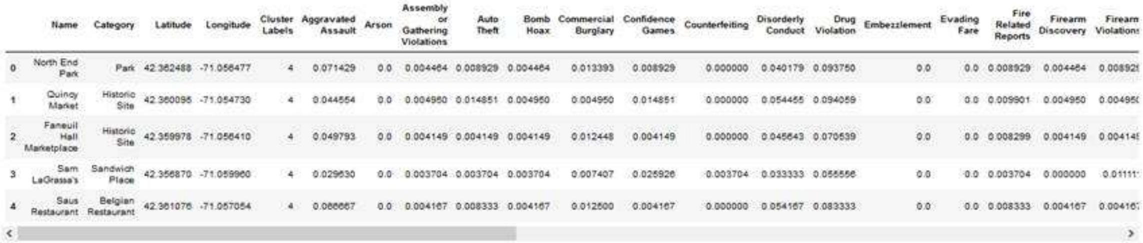


	Name	Aggravated Assault	Assembly or Gathering Violations	Auto Theft	Bomb Hoax	Commercial Burglary	Confidence Games	Counterfeiting	Disorderly Conduct	Drug Violation	Embezzlement	Evading Fare	Fire Related Reports	Firearm Discovery	Firearm Violations	Fraud
0	Aquarium Seal Tank	0.018519	0.000000	0.037037	0.000000	0.018519	0.018519	0.000000	0.055556	0.000000	0.000000	0.000000	0.000000	0.018519	0.000000	0.018519
1	Bacco Wine and Cheese	0.042105	0.000000	0.031579	0.000000	0.010526	0.010526	0.021053	0.010526	0.000000	0.000000	0.010526	0.010526	0.010526	0.000000	0.031579
2	Ball and Buck	0.024691	0.000000	0.012346	0.006173	0.024691	0.024691	0.012346	0.012346	0.018519	0.000000	0.000000	0.006173	0.006173	0.000000	0.043210
3	Barcelona Wine Bar	0.019608	0.000000	0.019608	0.000000	0.058824	0.000000	0.000000	0.000000	0.000000	0.019608	0.000000	0.039216	0.019608	0.000000	0.000000
4	Barry's Bootcamp Boston	0.035714	0.004464	0.000000	0.000000	0.008929	0.013393	0.004464	0.035714	0.138393	0.000000	0.000000	0.004464	0.000000	0.013393	0.026786

The optimum value for k is determined by observing the plot of Sum of squared distance against k and looking for the value of while a perfect elbow does not occur the value of k=8 appaears to the optimum



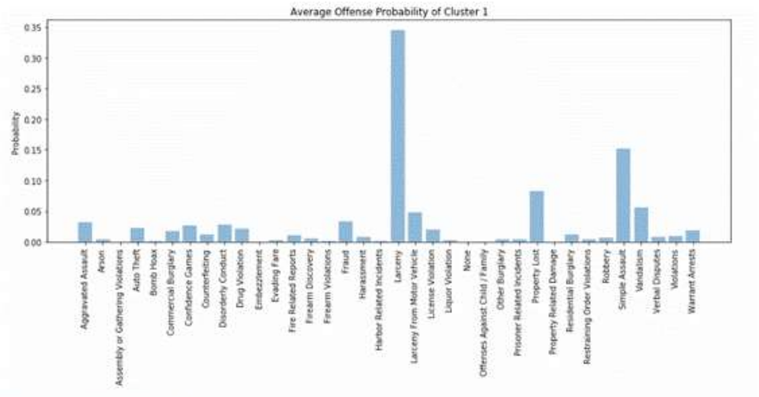
The labels so obtained are appended to the dataset containing the probabilities of occurrence of each offense in vicinity of a given venues and then finally merged with Boston venues dataset to get a final data set



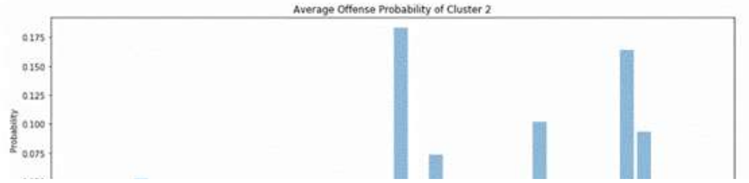
	Name	Category	Latitude	Longitude	Cluster Labels	Aggravated Assault	Arson	Assembly or Gathering Violations	Auto Theft	Bomb Hoax	Commercial Burglary	Confidence Games	Counterfeiting	Disorderly Conduct	Drug Violation	Embezzlement	Evading Fare	Fire Related Reports	Firearm Discovery	Firearm Violations
0	North End Park	Park	42.362488	-71.056477	4	0.071429	0.0	0.004464	0.008929	0.004464	0.013393	0.008929	0.000000	0.040179	0.093750	0.0	0.0	0.008929	0.004464	0.00892
1	Quincy Market	Historic Site	42.360095	-71.054730	4	0.044554	0.0	0.004950	0.014851	0.004950	0.004950	0.014851	0.000000	0.054455	0.094059	0.0	0.0	0.009901	0.004950	0.00495
2	Faneuil Hall Marketplace	Historic Site	42.359978	-71.056410	4	0.049793	0.0	0.004149	0.004149	0.004149	0.012448	0.004149	0.000000	0.045643	0.070539	0.0	0.0	0.008299	0.004149	0.00414
3	Sam LaGrassa's	Sandwich Place	42.356870	-71.059960	4	0.029630	0.0	0.003704	0.003704	0.003704	0.007407	0.025926	0.003704	0.033333	0.055556	0.0	0.0	0.003704	0.000000	0.01111
4	Saus Restaurant	Belgian Restaurant	42.361076	-71.057054	4	0.066667	0.0	0.004167	0.008333	0.004167	0.012500	0.004167	0.000000	0.054167	0.083333	0.0	0.0	0.008333	0.004167	0.00416

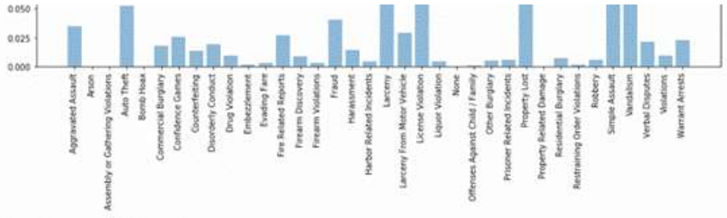
5. Results and discussion

To understand the results let's look at the plots of average probabilities of occurrence of a given offense in each cluster

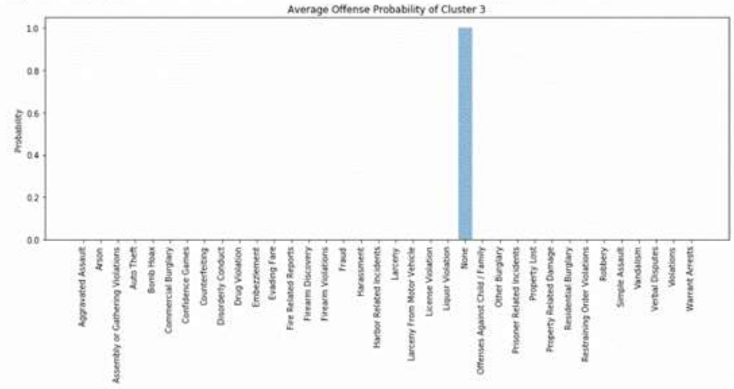


```
Number of venues in cluster: 21
Most prominently expected offenses in this cluster of venues: Larceny:34.48%,Simple Assault:15.17%
```

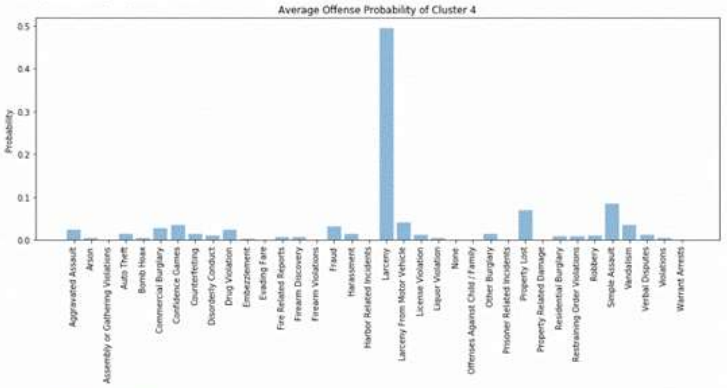




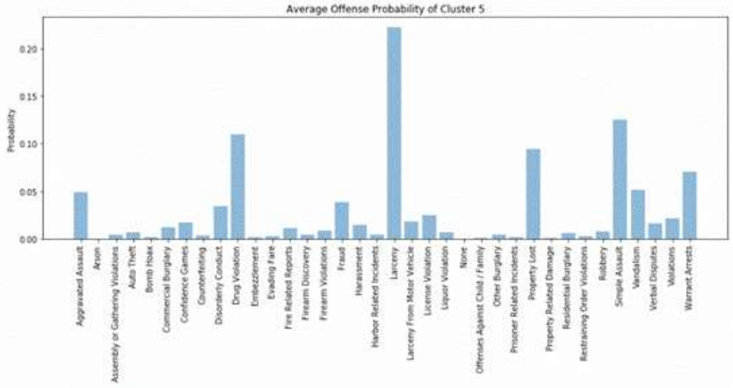
```
Number of venues in cluster: 14
Most prominently expected offenses in this cluster of venues: Larceny:18.28%,Simple Assault:16.37%
```



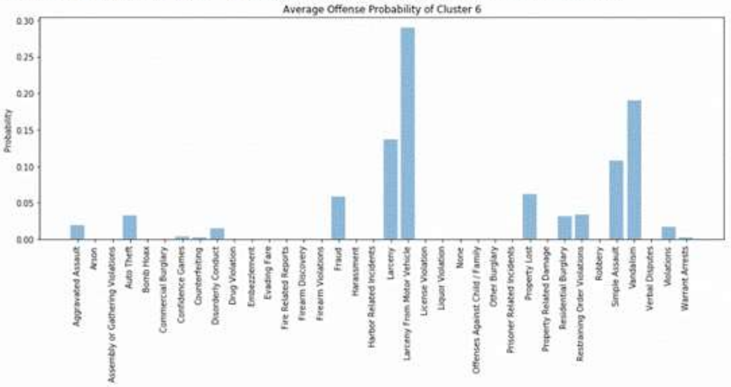
```
Number of venues in cluster: 1
Most prominently expected offenses in this cluster of venues: None
```



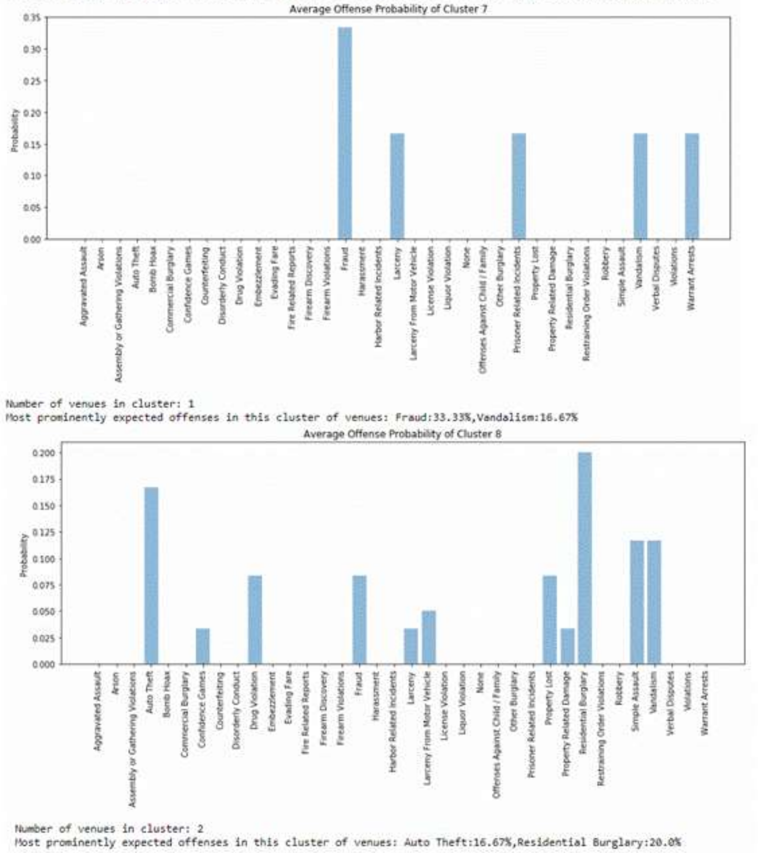
```
Number of venues in cluster: 15
Most prominently expected offenses in this cluster of venues: Larceny:49.35%,Simple Assault:8.37%
```



```
Number of venues in cluster: 36
Most prominently expected offenses in this cluster of venues: Larceny:22.18%,Simple Assault:12.51%
```



```
Number of venues in cluster: 6
Most prominently expected offenses in this cluster of venues: Larceny From Motor Vehicle:28.95%,Vandalism:19.03%
```



```
Number of venues in cluster: 1
Most prominently expected offenses in this cluster of venues: Fraud:33.33%,Vandalism:16.67%
```

```
Number of venues in cluster: 2
Most prominently expected offenses in this cluster of venues: Auto Theft:16.67%,Residential Burglary:20.0%
```

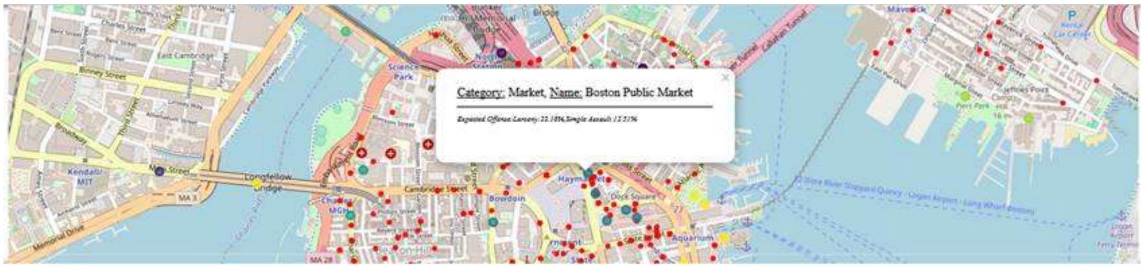
On observing the plots above we are able to identify the distribution of average probabilities of offenses in a given a cluster of venues. The prominent offense types are distinctly visible in most cases. It is important to pay attention to value of mean probabilities of offenses types in the cluster. In some cases the distribution is spread out but the mean values are small whereas in others the distribution is more concentrated and the probabilities are high. It would be beneficial to think of venues in clusters with low probabilities of offenses as safer compared to the ones that have high probabilities. Even a couple of high probabilities of offenses in a cluster would suggest that they are more unsafe as compared to venues in other clusters where the distribution is spread out but the average probabilities of offenses are low, though it is important to qualitatively understand the offense type and its weightage. Say for example, a small probability of an aggravated assault offense type is potentially more than a considerable probability value of Liquor Violation. Identifying high concertation of offense probabilities also is useful for law enforcement agencies to concrete their efforts in tackling these offenses. Apart from this the cluster of venues observed here are also found to be geographically clustered when observed on a map, indicating a relation of offense types to the geographical locations.

It would be necessary to understand that we have considered the dataset which is has more recent records. Now, sometimes it also useful to consider the data that is more post dated as offenses at a given location also tend to also follow a pattern in time, but the choice of time in selecting the records in crime incident dataset depends on what one is trying to accomplish. In our cases we were interested in most recent trends and therefore selected crime incidents that occurred in recent past. The offense types that we considered important were selected and others were dropped again this was an arbitrary choice. One can limit this selection of offense types for instance if one is looking for only more heinous crimes types the offense type feature would have fewer values. Venue dataset too is limited by the account type used to access the API whereas some information (nearly 5-10%) from crime incidents dataset is lost because of the unavailability of latitude and longitude co-ordinates or their incorrect values. The results obtained therefore must be looked at after accounting in all such factors.

In the end though, we could reasonably well answer the question that we posed before us as we embarked on the analysis: How safe are we when we visit one of the popular venues in the town? Merely now looking at the cluster in which this venue lies we can easily identify the two most prominent offense types in vicinity of such venue. We can further drill down to identify other offense types that are expected by looking at the distribution of offense types of the cluster. Somebody looking to start say a new restaurant in vicinity of a venue can now be more mindful of how safe the neighborhood is and what is expected.

6. Conclusion

Let us conclude the analysis by plotting the clustered venues on the map of Boston, MA where different clusters of venues are color coded indicating the similarity of the offences that are expected to occur in their vicinity. The marker for each venue shows the probabilities of two most common offences that is expected to occur around them based on average probabilities of crime incidents in the cluster. The venues are overlayed on the crime map of Boston to give more correct and exact idea of the offense that occurred in vicinity of the venue along with its place and date of occurrence.



7. References

[1]: Four Square API

[2]: GeoPy Documentation

The notebook is located at: https://nbviewer.jupyter.org/github/ab-datascientist/capstoneproject/blob/master/WK5%20Assignment.ipynb and also on GitHub