**Capstone Project – Battle of the Neighborhoods**

**Analysis of the Neighborhoods in Boston, MA**

**Abby Antoo Antony**

**Contents**

[**1. Introduction** 2](#_Toc47224850)

[**2. Data Acquisition** 3](#_Toc47224851)

[**3. Data Analysis and Modelling** 4](#_Toc47224852)

[**3.1 Data Cleaning** 4](#_Toc47224853)

[***3.1.1 Adding the District\_Names column*** 4](#_Toc47224854)

[***3.1.2 Dropping Columns*** 4](#_Toc47224855)

[**3.1.3 *Missing Values*** 4](#_Toc47224856)

[***3.1.4 Pivot Table – Adding ‘Total’ Column*** 4](#_Toc47224857)

[**3.2 Data Exploration** 5](#_Toc47224858)

[***3.2.1 Districts with the highest and lowest reported crime*** 5](#_Toc47224859)

[***3.2.2 Top 10 offense category in Charlestown*** 5](#_Toc47224860)

[***3.2.3 Dropping duplicates and using Folium to map Charlestown*** 5](#_Toc47224861)

[**3.3 K-means clustering** 5](#_Toc47224862)

[**4. Results** 6](#_Toc47224863)

[**4.1 Cluster 1** 6](#_Toc47224864)

[**4.2 Cluster 2** 6](#_Toc47224865)

[**4.3** **Cluster 3** 6](#_Toc47224866)

[**4.4 Cluster 4** 6](#_Toc47224867)

[**4.5 Cluster 5** 6](#_Toc47224868)

[**5. Discussion** 7](#_Toc47224869)

[**6. Conclusion** 7](#_Toc47224870)

[**7. Appendix** 8](#_Toc47224871)

# **1. Introduction**

I recently graduated from my master’s program which was in san Francisco and I am looking to start a new chapter in my life in Boston, Massachusetts. An individual moves an average of 11.7 times in their lifetime. In fact, in America, between 2012 and 2013, a total of 35.9 million people aged one year or older moved. There are various reasons for moving like Job Relocations, Changing Neighborhoods, Marriage or Retirement. My primary reason for moving is job hunt. Boston, MA has done a fine job of attracting young professionals in abundance over the years with their reputed educational institutions, diverse communities, and access to world class healthcare

Safety and access to necessities are the biggest concerns when thinking about moving to a new location. The goal of this project is split into two parts. First, to find out the safest neighborhood to live in Boston, MA. Second, to explore the neighborhood to determine the various venues per street and cluster them using k-means clustering. By the end of the analysis, I or any individual should be able to narrow down which street to move into in the neighborhood identified as the safest.

# **2. Data Acquisition**

The data is extracted from [Analyze Boston](https://data.boston.gov/), which is an open data hub of the city of Boston. The data set that will be used is [Crime Incidents Report](https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system). These are reports provided by the Boston Police Department to document the initial details surrounding an incident to which BPD officers respond. The dataset contains records from the new crime incident report system, which includes a reduced set of fields focused on capturing the type of incident as well as when and when it occurred. The record begins in June 2015. Since the analysis is focused on safety, the latest year i.e 2019 is considered.

***Metadata****:*

* **Incident\_num**: Internal BPD report number
* **Offense\_code**: Numerical code of offense description
* **Offense\_Code\_Group\_Description**: Internal categorization of [offense\_description]
* **Offense\_Description**: Primary descriptor of incident
* **District**: What district the crime was reported
* **Reporting\_area**: RA number associated
* **Shooting**: indicated a shooting took place
* **Occurred\_on**: Earliest date and time the incident could have taken place
* **UCR\_Part**: Universal  Crime Reporting Part number (1,2,3)
* **Street**: Street name the incident took place

# **3. Data Analysis and Modelling**

## **3.1 Data Cleaning**

This dataset consists of data from 2015 to 2020. Since the analysis requires the most recent year, the 2019 data is selected. This reduces the number of rows and columns to 98082 and 17 respectively. The cleaning process is split into 4 parts. They are as follows:

### ***3.1.1 Adding the District\_Names column***

The values in the ‘DISTRICT’ column represents the different neighborhoods in Boston but this is represented by district codes e.g C6, E18 etc. This makes it difficult for an individual who is not from Boston to identify the names of the neighborhood. Therefore, for the sake of simplicity, I added a column, ‘dist\_names’(District\_Names), with the name of the neighborhoods corresponding to their district code. (Appendix Fig 1)

### ***3.1.2 Dropping Columns***

There are 17 columns in the dataset. For this analysis, the initial goal is to identify the safest neighborhood in Boston, MA and for that 6 out of 17 columns will only be needed i.e ‘District\_Names’, ‘Street’, ‘Offense\_Category’, ‘Lat’, ‘Long’. The remaining 11 columns are dropped. (Appendix Fig 2)

### ***3.1.3* *Missing Values***

One of the main challenges in this analysis are the missing values. All the Columns have missing values but District\_Names have missing values that are represented by blanks whereas the other columns have their missing value represented as ‘NaN’. The key columns of this analysis are the District\_Names and Offense\_Category, Therefore, I converted the missing values that are blank in the District\_names column to NaN and dropped the NaN’s in both District\_Names and Offense\_Category. (Appendix Fig 3)

### ***3.1.4 Pivot Table – Adding ‘Total’ Column***

Next, I grouped the District\_Names and Offense\_Category to get the number of different offenses or crimes per district or neighborhood and renamed the column as ‘Count of offenses per dist’. Once that was done, I used the pivot\_table function to covert the data frame into a pivot table where the rows and columns are ‘District\_Names’ and ‘Offense\_Category’ respectively with the ‘Count of offense per dist’ as values. Finally, the ‘Total’ column is added to the table which represents the sum of crimes per neighborhood. (Appendix Fig 4)

## **3.2 Data Exploration**

### ***3.2.1 Districts with the highest and lowest reported crime***

Using the seaborn and matplotlib libraries, a bar graph was created to visualize the top four neighborhoods with the highest and lowest number of crimes. Based on the bar chart, Roxbury has the highest number of reported crimes with an astounding 11764 reported cases in year 2019 and Charlestown has the lowest with 1397 reported cases. Therefore, Charlestown is the safest neighborhood based on the total number of crimes per neighborhood. (Appendix Fig 5)

### ***3.2.2 Top 10 offense category in Charlestown***

Looking deeper into Charlestown, it’s clear that top 10 offenses are motor vehicle accident response, medical assistance, larceny, investigate Person, towed, simple assault, vandalism, Drug Violation and larceny from motor vehicle where motor vehicle accident response and larceny from motor vehicle are the most and least among all of them. (Appendix Fig 6)

### ***3.2.3 Dropping duplicates and using Folium to map Charlestown***

The next goal is exploring the different venues around Charlestown by street and segregate these venues into different clusters using k-means. Before that, the data needs to be cleaned a bit more. The reason for not dropping all the NaN during the data cleaning process is because dropping all the NaN would lead to having less amount of observations which would have an impact on the first part of the analysis i.e finding the safest neighborhood. Now since Charlestown is determined to be the safest neighborhood, the next step is dropping the remaining NaN as well as duplicates from the ‘Street’ column to map Charlestown. (Appendix Fig 7)

## **3.3 K-means clustering**

The final dataset contains the different streets in Charlestown along with their corresponding latitude and longitude. Now, that the final data set is ready, the Foursquare API is connected to find the various venues within a 500-meter radius. A JSON file is returned containing all the venues in each street. To get a clear picture of the data obtained from Foursquare, the data is converted into a dataframe and is assigned to a variable called Charlestown\_data.

The next step is One hot encode the Charlestown\_data to convert the categorical variables into a binary to perform k-means. The dataset is then grouped by street and the average or mean is calculated which finally leads to listing the top 10 common venues in each street.

K-means clustering is a machine learning algorithm that segregates data and clusters them based on similar traits given that the cluster sizes are predefined. In this case, the predefined cluster size is 5 and the goal is to find similar streets in Charlestown based on the various venues. This gives an individual options in the form of clusters to decide which street to move into based on the venues in those respective streets. (Appendix Fig 8)

# **4. Results**

The five clusters created by running k-means clustering consists of 138 streets in the neighborhood of Charlestown. Let’s look into each of the clusters.

## **4.1 Cluster 1**

The first cluster consists of 18 out of the 138 streets. Individuals who are current residents or potential residents will have access to a bunch of pizza places, deli shops and coffee but mainly pizza places as it appears 14 times in the 1st and 2nd most common venues. There are other venues available like music venues, tennis courts and gyms but they are not as common. (Appendix Fig 9)

## **4.2 Cluster 2**

The second cluster is the largest and consists of 60 streets out of the total 138. Individuals moving into these streets or who are currently residents have access to a lot of pubs as it appears 37 times in the 1st, 2nd, and 3rd most common venue columns. The streets also have other venues such pizza places, convenience stores, café’s, and pharmacies. The other venues that are not very common are pet stores, gyms and grocery stores to name a few. (Appendix Fig 10)

## **4.3** **Cluster 3**

The third cluster is the second largest one and consists of 34 streets. These streets are for individuals who enjoy going to history museums, national parks, going on boats & ferries and having donuts occasionally or every day. The other venues that are not very common are pizza places, gyms and grocery stores to name a few. (Appendix Fig 11)

## **4.4 Cluster 4**

The fourth cluster consists of 25 out of 138 streets and these streets are for individuals who are into fitness as well as for individuals who rely on public transportation for their daily commute or for new residents who are planning to use public transport as their main source of transportation. Other venues are donut shops, cafe's, parks, and liquor stores to name a few. (Appendix Fig 12)

## **4.5 Cluster 5**

The fifth cluster consists of just 1 street out of the 138 making this an outlier. The most 3 most common venues are taxi stands, restaurants and yoga studios. (Appendix Fig 13)

# 

# **5. Discussion**

The goal of this analysis/project was to help me to look into the different neighborhoods of Boston, MA to decide which one to move to but this is an analysis/project that can assist any individual who are planning to move into the city of Boston. The analysis showed that Charlestown was the safest neighborhood having 1397 reported crimes compared to an astounding 11764 reported crimes in Roxbury in the year 2019.

K-means clustering grouped all streets in the neighborhood of Charlestown into 5 clusters with similar venues. Based on the analysis, I would personally look into the streets in cluster 4 because I am a person who enjoys fitness as well as uses public transport as a daily source of transportation. For individuals who like to eat out a lot and have fun with access to convenient stores and pharmacies, clusters 1 and 2 are for you. Individuals who enjoy hikes, history, boats, ferries and bit of surfing, cluster 3 is for you.

# **6. Conclusion**

Moving into a new city or country has so many factors to it. This analysis is an indicator of how technology can be utilized to gain additional insights on the city or country one is planning on moving to. Couple lines of code narrowed down the safest neighborhood as well the venues around that neighborhood making the whole process of moving easier for an individual. They could use their time to look into the other factors that were not taken into consideration like the cost of living and housing, which deserves their own analysis. But as a start, anyone who is planning to move to the city of Boston, Charlestown is the place to be.

# **7. Appendix**

Fig 1:

A screenshot of a cell phone

Description automatically generated

Fig 2:

![A screenshot of a cell phone

Description automatically generated]()

Fig 3:

A screenshot of a cell phone

Description automatically generated

Fig 4:

A screenshot of a cell phone

Description automatically generated

Fig 5:

A screenshot of a cell phone

Description automatically generated

Fig 6:

A screenshot of a cell phone

Description automatically generated

Fig 7:

![A picture containing map, text

Description automatically generated]()

Fig 8

A picture containing text, map

Description automatically generated

Fig 9:



Fig 10:

![A screenshot of a cell phone

Description automatically generated]()

Fig 11:

![A receipt on a black background

Description automatically generated]()

Fig 12:

![A screenshot of a cell phone

Description automatically generated]()

Fig 13:

