Data Analysis on Choosing District for New Residential Project

Tse Chun Ho

August 9, 2021

1. Introduction

1.1 Background

Equity Residential is one of the greatest real estate investment trusts in America, which invests in high class apartments. In this project, we assume that they would like to start a new residential project in Boston, one of the most populous cities in the United States. Assuming I am one of the data scientists in Equity Residential and in-charge of analysing data to choose the best place for a new residential project. The management of the company has two main requirements for selecting a district. The first requirement is the selected district must be safe to protect the resident. The second requirement is the environment of the selecting district needs to be similar to Downtown Boston since the company has had a huge success on the Downtown Boston residential project.

1.2 Problem

This project aims to choose a potential district in Boston for a new residential project. Therefore, this project will also analyse crime records in Boston in order to enhance new residential project security.

1.3 Interest

Obviously, the management level of Equity Residential will be very interested in this project. Boston residents may also be interested in this project since they will know which district is safer.

2. Data acquisition and cleaning

2.1 Data Sources

One of the data sources is the crime record of Boston from 2015 to 2018. The data can be downloaded from the official site of the Boston government. The following table is the metadata of the data:

Column Name	Data Type	Description
INCIDENT_NUMBER	String	Internal BPD report number
offense_code	String	Numerical code of offense description
Offense_Code_Group_Description	String	Internal categorization of offense_description
Offense_Description	String	Primary descriptor of incident
district	String	What district the crime was reported in
reporting_area	String	RA number associated with the where the crime was reported from
shooting	String	Indicated a shooting took place
occurred_on	Datetime	Earliest date and time the incident could have taken place
UCR_Part	String	Universal Crime Reporting Part number (1,2, 3)
Street	String	Street name the incident took place
Location	String	Latitude and longtitude of the Street

table 1 : Column information of the crime record

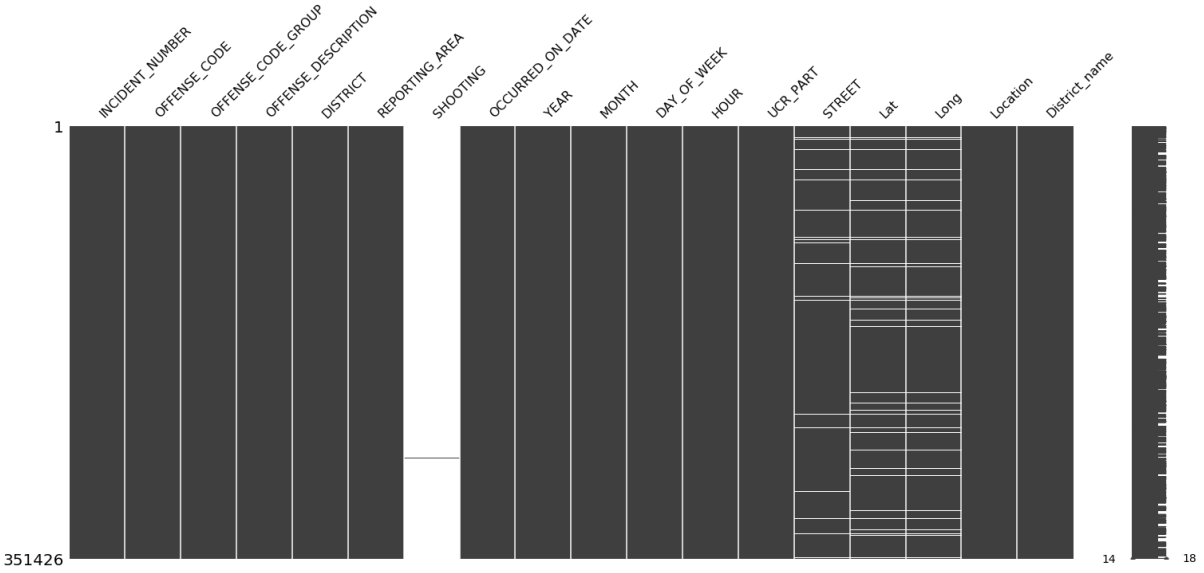
Second data the project will be used for is from the foursquare database. Foursquare API will be used to get the nearby venues by district name. The detailed foursquare API document can be found here.

2.2 Data Cleaning

First of all, the crime records downloaded on the Boston government website are saved year by year. The first step is to combine all records into one dataframe.

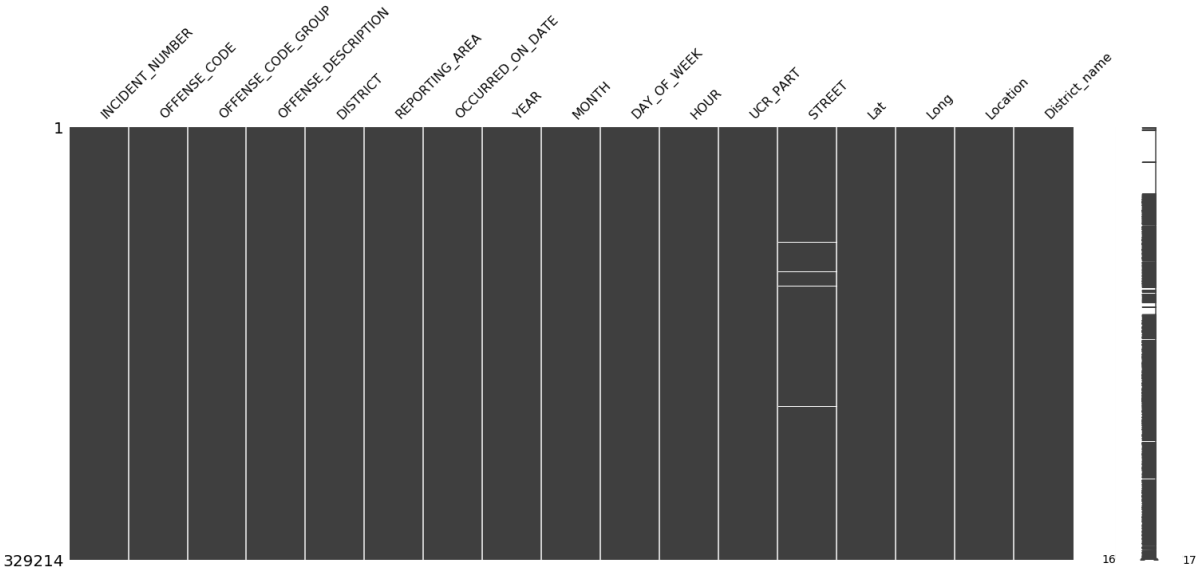
From the dataframe, the district column is not informative enough. The columns only saved the district code instead of the district name. In order to create a new column to save the corresponding district name, I have found the corresponding district name from Boston's official website and recorded it in a csv format.

Therefore, some of the columns in the dataframe may not be useful at all. Missingno is one of the useful python libraries to check the quality of the dataframe. It can generate a graph to visualize the nan value in the dataframe.



Picture 1: Visualizing the nan value in dataframe

From picture 1, we can observe that the “Shooting” column is nearly all nan. The “Shooting” column will be dropped since it does not imply any information. The second observation is that some of the location information is missing. The row that missing location data will be dropped since location data is extremely important for further analysis.



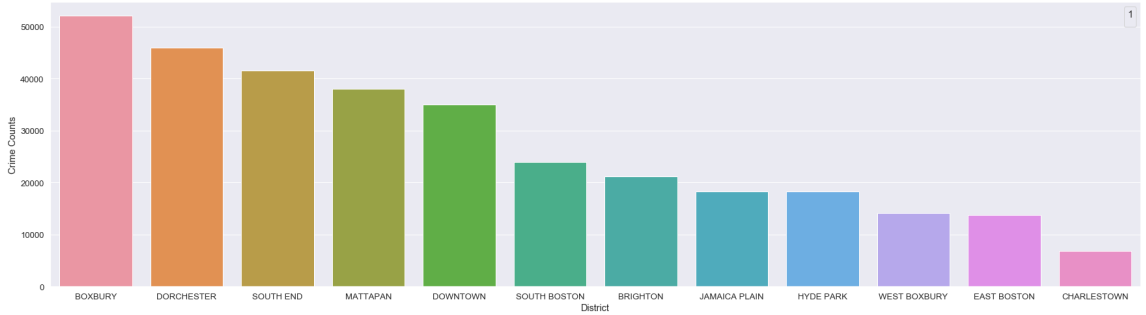
Picture 2 : Visualizing the dataframe after dropping the unused part

After dropping the shooting column and some location-missing row, the dataframe has a better condition. Only
Street column remains some NaN value, but it is fine since Lat and Long can be used for location analysis.

Since the occured time of crimes will also be analysed, I have converted the date, day into numeric number columns called dayofweek, quarter, dayofyear, dayofmonth and weekofyear, which helps to analyse the distribution between the time and crime.

3. Exploratory Data Analysis

3.1 Crime Distribution grouped by District and UCR Part

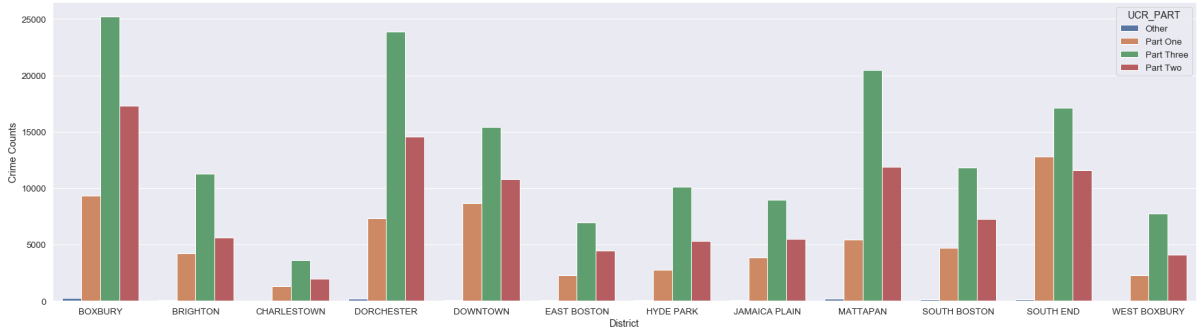


Picture 3: Barchart on the crime count by district

From picture 3, we can observe that Boxbury, Dorchester and South End are the highest crime rate districts in Boston. These three districts have over 40000 counts from 2015 to 2018. In contrast, West Boxbury, East Boston and Charlestown are the lowest districts which are having the lowest crime count.

Based on the barchart, we can classify all districts into high crime rate groups, middle crime rate groups and low crime groups. Boxbury, Dorchester, South End and Mattapan are defined as high crime rate groups. Downtown, South Boston, Brighton and Jamaica Plain are grouped as middle crime rate groups. Lastly, Hyde Park, West Boxbury, East Boston and Charlestown are defined as low crime rate groups.

For the UCR part, it may need some explanations for what it implies. UCR means uniform crime reports, which are published by the FBI. UCR part one usually has two categories: violent and property crimes. All the crimes grouped in UCR part one are considered as more serious. For part two, some less-serious crimes are grouped, such as forgery and counterfeiting, disorderly conduct, driving under the influence, drug offenses, fraud and gambling etc.



Picture 4: Barchart to show the crime counts grouped by district and UCRpart

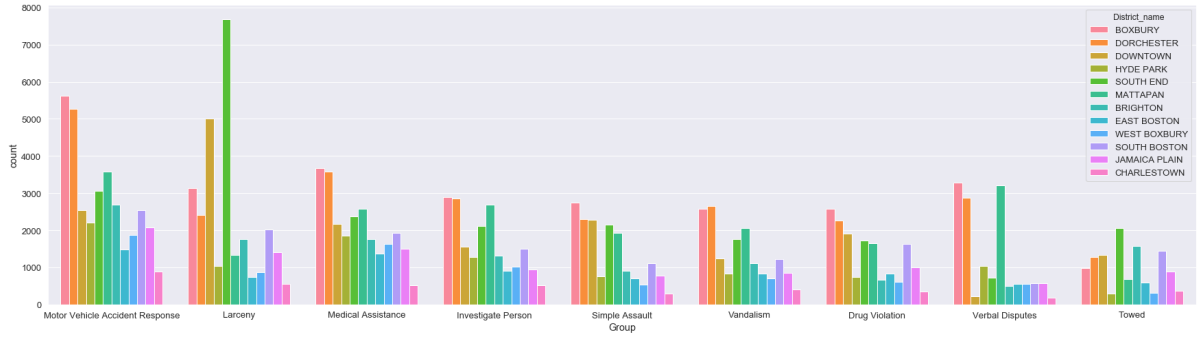
From picture 4, we can observe three trends.

First, Boxbury, Mattapan and Dorchester have a higher crime rate than other districts. From the barchart, these three districts are the top three districts in UCR two and UCR three. It means that the residents may encounter more minor crime in these districts, which is not a good place for us to develop a new residential project here.

Second, South end and Downtown have a higher UCR part two crime rate than the remaining districts. As explained before, UCR part one crime is the most serious crime category. The potential resident will avoid choosing this kind of danger district as their home, it may affect the sales of the new residential project.

Third, Charlestown and East Boston have a competitively low crime rate in Boston. From the graph, it shows that the UCR part one, two and three in Charlestown are lower than 5000, which is the lowest crime rate district. Also, East Boston had the second lowest crime rate in Boston. These two districts may be the potential districts for the new residential project.

3.2 Crime Category count group by district



Picture 5 : Barchart on different crime category group by district

In this section, we go deep to study different crime category counts in different districts. In picture 5, the top 10 crime categories are listed on the chart and separated by different districts. The top 10 of crime categories are motor vehicle accident response, larceny, medical assistance, investigating person, simple assault group, vandalism, drug violation, verbal disputes and towing.

Within these ten categories, some crimes will highly affect our resident living experience.

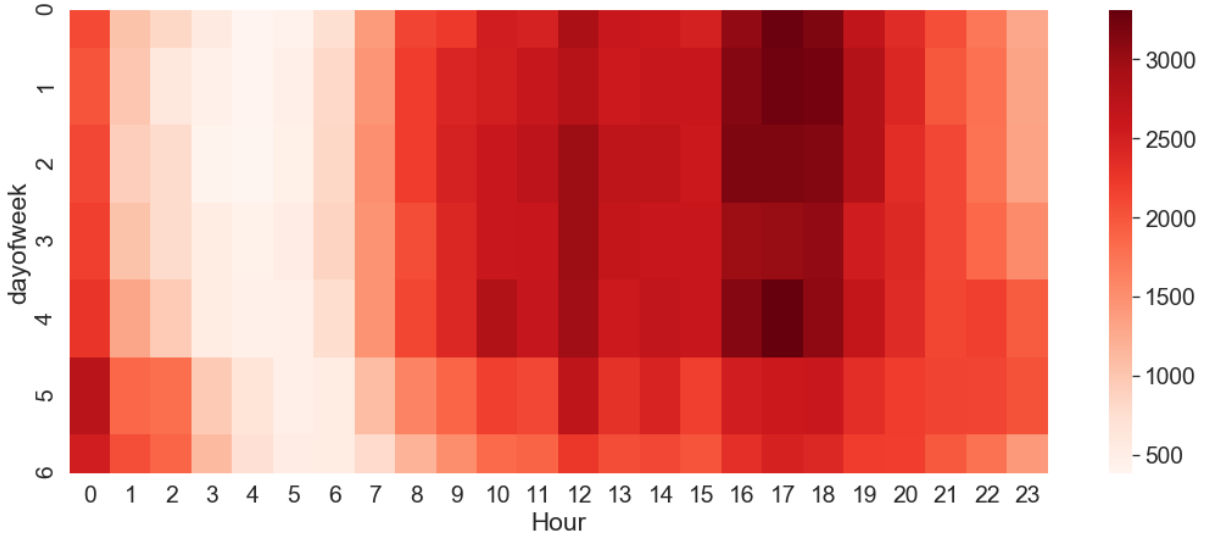
First, larceny will be one of the crime categories that need to be concerned. From picture 5, larceny placed second in all crime categories. which is a frequently-occured crime in Boston. It will be a big advantage if the residential project is placed in competitively low larceny rate places. Boxbury, South End and Downtown are the top three larceny rate districts. To ensure the residents have a lower opportunity to be pickpocketed, the new residential project should avoid choosing these three districts.

The second crime category that should be concerned is simple assault. As a high class apartment developer, resident safety will be an indispensable factor for us to consider the location. Based on this factor, a higher violence rate district will not be considered as a potential district for new residential projects. From picture 5, Boxbury, Dorchester and Downtown are the top three districts having the highest simple assault rate, these three districts may not be good to be chosen.

Therefore, Vandalism is also a kind of crime that needs to be concerned. Vandalism means an action to deliberate destruction of or damage to public or private property. This kind of crime may highly affect the cityscape. As a luxurious apartment brand, our customers are expected to live in a better cityscape district. From picture 5, Boxbury, Dorchester and Mattapan are the top three districts on the vandalism crime count; these three districts are also not recommended to be chosen.

The last crime category that needs to concern is drug violation. If the district is having a high drug violation crime rate, there may be more addicts appearing on the street. This kind of situation may be harmful to our residents. According to the result from picture 5, Boxbury, Dorchester and Downtown have the highest rate in drug violation, which are not recommended for choosing too.

3.3 Data Visualization on Crime Occured Time



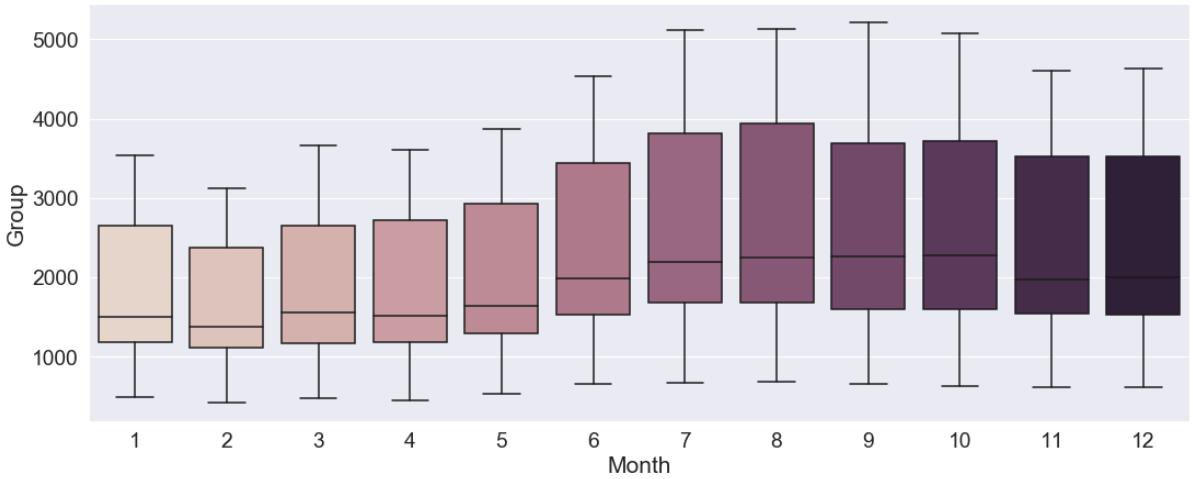
Picture 6: Heatmap about the occured time of crime

After discussing the statistics on crime categories and districts, this part will focus on the occured time of the crime. In picture 6, heatmap will be used to understand the distribution on the weekday and time of the crime.

From the above heatmap, we observe that most of the crimes happened from 16:00 to 18:00 on Monday to Friday. The graphs showing these time periods are having a deeper color, which is above 3000 counts in each time slot.

Therefore, 9am to 6pm and midnight have a higher crime rate than other timeslot. From picture 5, these timeslot are having around 1500-3000+ counts, which means that most of the crimes are happening within these timeslot.

After analysing the occured time, our company can use this kind of data to plan the security arrangement for the new residential project. For example, we can arrange a security guard to patrol from 4pm to 6pm since it is the most unsafe timeslot.

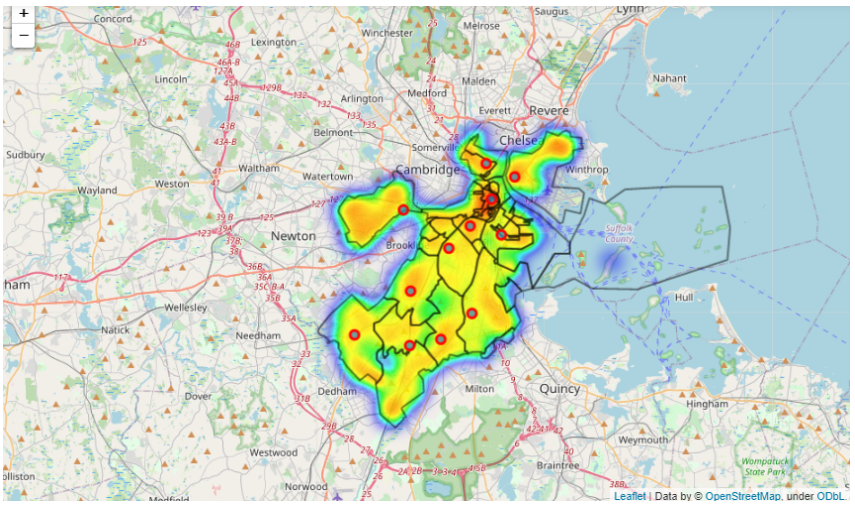


Picture 7: Box-and-whisker diagram on crime occurred month

In this part, crime occurring month is used to analyse the pattern. From picture 7, we can observe that July, August, September and October are slightly higher than other months. The maximum of these four months are all over 5000 counts. Also, August will be the highest crime rate month. According to the graph, the upper quartile and lower quartile of August are higher than other months. And the median of August is nearly highest.

After we have gathered this information, the new residential project can plan to add some security manpower in August in order to strengthen the security for our apartment.

3.4 Crime Heatmap on Boston



Picture 8 : Crime heat map generated by folium

In order to analyse the crime location distribution, a heat map will be a good way to see the safety level of the district or even street. In this part, folium will be used to generate the heat map widget. The latitude and longitude will be used to create a heat map layer. The widget can also be zoomed in to a particular district to check the safer zone.

4 K-means Clustering

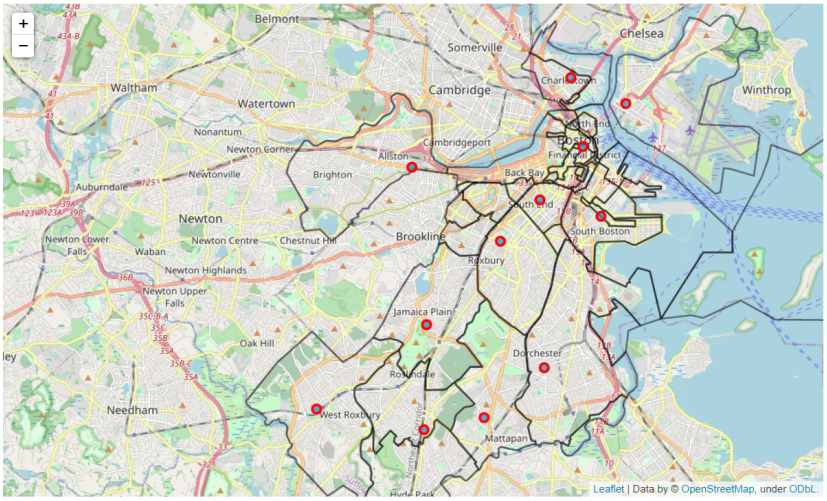
4.1 Data Preparation on K-mean Clustering

One of the requirements of the apartment location is the district needs to have a venue similar to Downtown Boston. Because of that, I need to gather information on the venues in different districts. Foursquare API could help to gather this kind of information.

K-means clustering algorithm will be used to partition the districts into different clusters. If the district is in the same cluster, it implies that this district is having similar venues as Downtown Boston, the district can be a potential district for the next residential project.

One of the parameters needed for foursquare API are the latitude and longitude of the district. For getting this information, a python library - geocoder is used. The geocoder will return the coordinate if you passed a place name to it. However, some of the coordinates are not accurate enough. Because of that, I have used google maps to check and correct the latitude and longitude.

To visualize the location of the district, I have used folium to plot a map and add some location marks.



Picture 9 : Capture from Folium map widget

After confirming the latitude and longitude, foursquare API can be called to get the district's venues.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	BOXBURY	42.330304	-71.089469	Dudley Café	42.329866	-71.083620	Café
1	BOXBURY	42.330304	-71.089469	Southwest Corridor Park	42.331830	-71.094111	Playground
2	BOXBURY	42.330304	-71.089469	Madison Field	42.332222	-71.086956	Soccer Field
3	BOXBURY	42.330304	-71.089469	Joe's Famous Steak & Cheese	42.328800	-71.083908	American Restaurant
4	BOXBURY	42.330304	-71.089469	Domino's Pizza	42.331238	-71.095335	Pizza Place

Picture 10 : Top 5 row in the venues dataframe got from Foursquare API

In the venues dataframe, the content of the venues are too rich, lots of columns will not be used. The aim of this part is to find the similar districts which have similar venues, all we need to use is the 'venue category' to identify the similarity.

In order to analyse the venue category more easily, one hot encoding will be used.

	Neighborhood	Accessories Store	African Restaurant	American Restaurant	Arepa Restaurant	Art Gallery	Asian Restaurant	Athletics & Sports	Automotive Shop	BBQ Joint	...
0	BOXBURY	0	0	0	0	0	0	0	0	0	...
1	BOXBURY	0	0	0	0	0	0	0	0	0	...
2	BOXBURY	0	0	0	0	0	0	0	0	0	...
3	BOXBURY	0	0	1	0	0	0	0	0	0	...
4	BOXBURY	0	0	0	0	0	0	0	0	0	...

Picture 11: Capture of first five rows of the dataframe after applying one hot encoding

After applying one hot encoding, all the columns will become the venue category. If the venue category is american restaurant, the american restaurant column will be 1.

Since there are too many rows in the dataframe, group by is a good method to integrate these rows to a more visible format.

	Neighborhood	Accessories Store	African Restaurant	American Restaurant	Arepa Restaurant	Art Gallery	Asian Restaurant	Athletics & Sports	Automotive Shop
0	BOXBURY	0.000000	0.076923	0.076923	0.000000	0.000000	0.000000	0.000	0.000000
1	BRIGHTON	0.000000	0.000000	0.000000	0.000000	0.000000	0.032258	0.000	0.032258
2	CHARLESTOWN	0.000000	0.000000	0.025000	0.000000	0.000000	0.000000	0.025	0.000000
3	DORCHESTER	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000	0.000000
4	DOWNTOWN	0.000000	0.000000	0.030000	0.000000	0.000000	0.020000	0.000	0.000000
5	EAST BOSTON	0.000000	0.000000	0.024390	0.000000	0.048780	0.000000	0.000	0.000000
6	HYDE PARK	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000	0.000000
7	JAMAICA PLAIN	0.000000	0.000000	0.000000	0.000000	0.047619	0.000000	0.000	0.000000
8	MATTAPAN	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000	0.000000
9	SOUTH BOSTON	0.000000	0.000000	0.025641	0.000000	0.000000	0.000000	0.000	0.000000
10	SOUTH END	0.013158	0.000000	0.039474	0.013158	0.013158	0.013158	0.000	0.000000
11	WEST BOXBURY	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000	0.083333

Picture 12: Capture of the dataframe after grouping all the rows by neighborhood

As shown from picture 12, the data has been normalized and ready to apply k-mean clustering. The result of k-mean clustering will be shown in the result session.

4.2 Approaches in K-means Clustering

As the venue dataframe has been ready, K-mean clustering can be performed to group the districts.

Using a random number to be K may not be an accurate way to cluster the district. In order to increase the accuracy, I will run the K-means clustering for ten times with different K values. After running 10 times K-means clustering, there will be ten clustering results. If the cluster group is the same as downtown, I will change the value to one. Otherwise, the value will be changed to zero.

district	BOXBURY	DORCHESTER	HYDE PARK	SOUTH END	MATTAPAN	BRIGHTON	EAST BOSTON	WEST BOXBURY	SOUTH BOSTON	JAMAICA PLAIN	CHARLESTOWN
Cluster Labels (k=11)	0	0	0	0	0	0	0	0	0	0	0
Cluster Labels (k=10)	0	0	0	0	0	0	0	0	1	0	0
Cluster Labels (k=9)	0	0	0	0	0	0	0	0	1	0	0
Cluster Labels (k=8)	0	0	0	1	1	0	0	0	1	1	0
Cluster Labels (k=7)	0	1	0	1	1	0	0	0	1	1	0
Cluster Labels (k=6)	0	1	0	1	1	0	1	0	1	1	0
Cluster Labels (k=5)	0	1	0	1	1	0	1	0	1	1	0
Cluster Labels (k=4)	1	1	0	1	1	1	1	0	1	1	0
Cluster Labels (k=3)	1	1	0	1	1	1	1	0	1	1	1
Cluster Labels (k=2)	1	1	1	1	1	1	1	0	1	1	1

Picture 13: Capture of the result dataframe

Based on the above result, we can sum up the column separately to get the similarity level to Downtown since the number represents the number of times that the district is in the same cluster with Downtown.

5 Result and Conclusion

5.1 Result

Based on the requirements of the new residential project location mentioned above, our result section should discuss two parts. First part of the result session will conclude the findings on the crime rate data. Second part of the second session will discuss the similarity with Downtown by the K-means clustering algorithm. Lastly, we will conclude the final decision for the location.

Result of Crime Record Data

Based on the part three discussion, I have used data visualization to study the crime counts, UCR part, crime category and the crime occured time. Districts we have defined as high, middle, low crime rate district groups in session 3.1 . High crime rate district groups are not suggested to choose as the location. For the normal crime rate group and low crime rate group, the crime counts are similar instead of Downtown. So, these districts in these two groups may be our potential district for the new residential project.

Therefore, according to the UCR part analysis in session 3.1 and crime category analysis in session 3.2 , Boxbury, Dorchester, Mattapan and South End have higher crime counts in some particular crime categories that highly affect the living quality.

In session 3.3 and 3.4, we can conclude that more crime is happening at 4:00pm to 6:00pm on weekdays. Also, July, August and September are having a higher crime count than other months. This kind of information would be helpful when the new residential project is going to plan the security planning.

Result of K-Means Clustering

district	Similarity with Downtown(%)
SOUTH BOSTON	90.0
SOUTH END	70.0
MATTAPAN	70.0
JAMAICA PLAIN	70.0
DORCHESTER	60.0
EAST BOSTON	50.0
BOXBURY	30.0
BRIGHTON	30.0
CHARLESTOWN	20.0
HYDE PARK	10.0
WEST BOXBURY	0.0

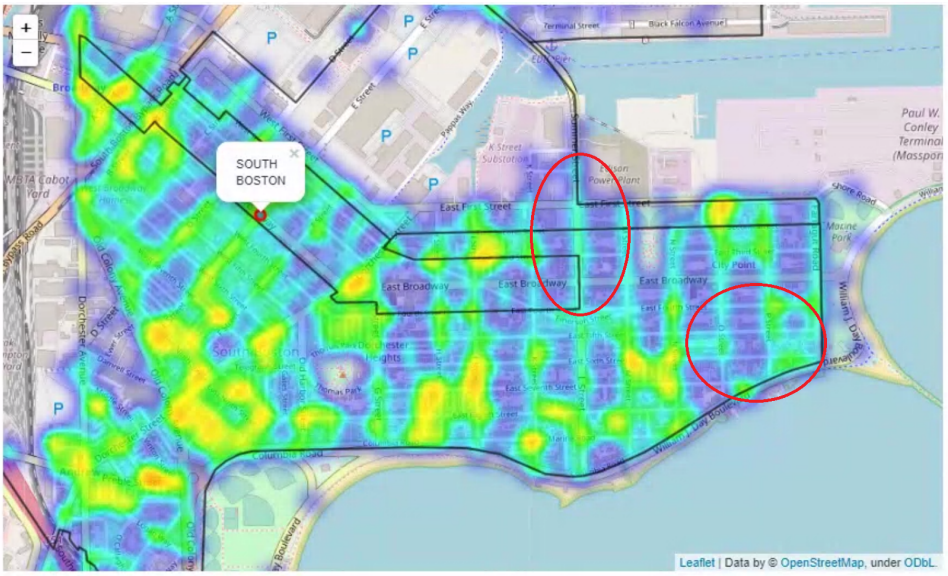
Picture 14: K-means Clustering Result on the Similarity with Downtown

As mentioned in session 4.2, the percentage of the similarity is calculated by numbers of time having the same cluster with Downtown divided by numbers of times we did for k-means clustering. The result shows that South Boston is the most similar to Downtown, 9 out of 10

times that are clustering the same group with Downtown. South End, Mattapan and Jamaica Plain also have 70% similarity to Downtown, but unfortunately, South End and Mattapan may not be good because some particular crime categories that affect the living quality are high.

5.2 Conclusion

After all the analysis I have done, South Boston will be chosen as the location of a new residential project. South Boston has the highest similarity with Downtown, which fits one of the requirements. Also, the crime counts of South Boston stays average among all the districts, and the crime counts of some particular crime categories also stays average than other districts, which fits the other requirement for the new residential project.



Picture 15: Crime heat map in South Boston

From the above heat map, we can observe the two red circles have a lower crime rate than other streets. These two locations can be the location of our new apartment.