Crime types and Venues Data Analysis of Halifax

Published January 21, 2021

1. Introduction

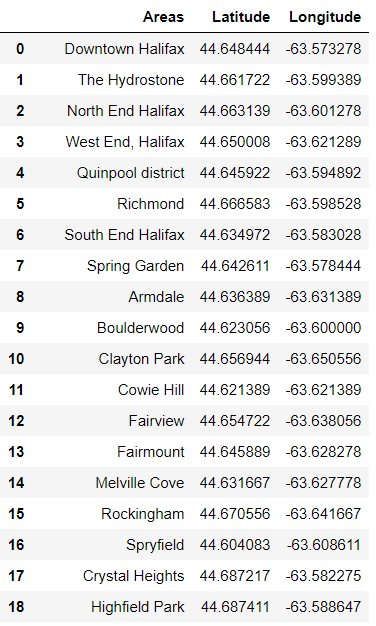
Halifax is a growing city in Nova Scotia, Canada. The Halifax regional municipality or the city of Halifax and the surrounding areas has a population of 403,100. The Halifax regional municipality is split into 48 neighborhoods according to Wikipedia (1). I currently live in this city and decided to base my Capstone project on finding out if the venues in certain neighborhoods can be used in addition to the types of crimes that take place in the area surrounding the neighborhood to build clusters in the city where certain avenues occur with certain types of crimes. Therefore, enabling me to able to name these clusters based on these characteristics such as, for example, a cluster with lots of theft of vehicles and has a lot of parking lots, or a cluster with a lot of assaults and a high number of bar or nightlife activity.

This type of clustering will enable policy makers, for instance, in the police department to focus resources looking for certain crimes in certain neighborhoods based on what venues exist in the area.

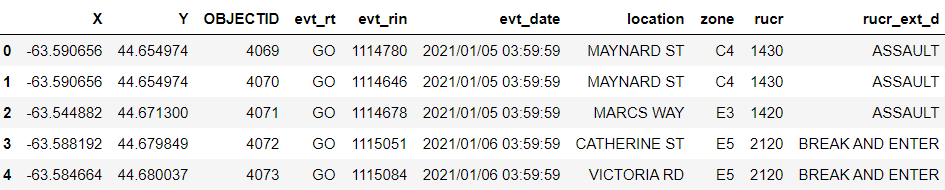
1. Data description

I used the Wikipedia webpage to load the different neighborhoods in the Halifax regional municipality area. I used the Foursquare API to get the most common venues of a given neighborhood in the Halifax area. I used the Halifax website which shares data to obtain the dataset regarding the different locations different crimes take place in the Halifax area and what types of crimes they are. I also used this website to obtain a geojson file showing the boundaries of the different community areas in the Halifax area, which was useful for building a choropleth map later on. Of course, I cleaned up this data, removing null values, and enabling the data to be combined between the different datasets.

1. Methodology
2. First step was to save the dataframe for neighborhoods

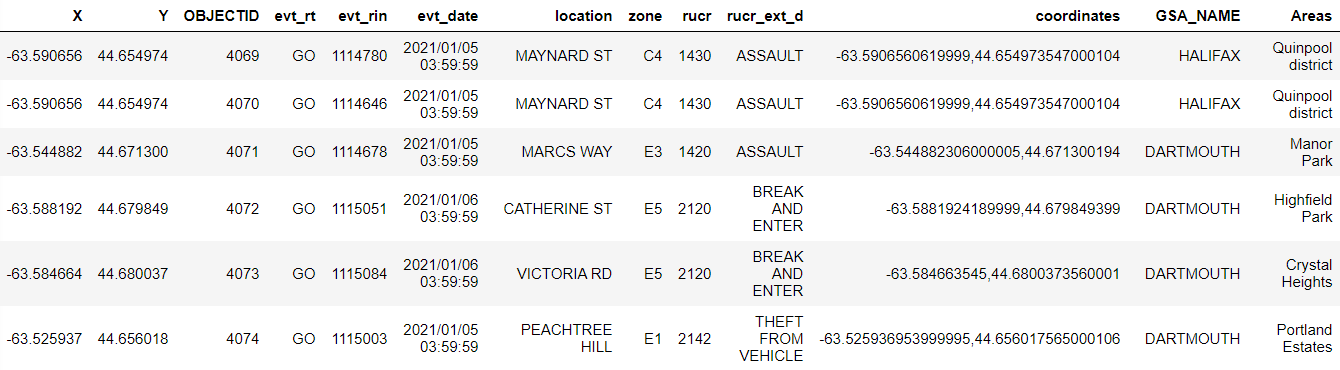


1. Next I saved the crime data into a dataframe too

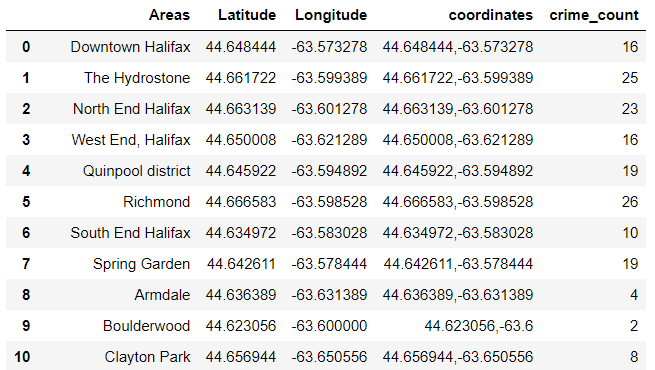


rucr\_ext\_d is what I was interested in and details the type of crime committed along with the X and Y coordinates.

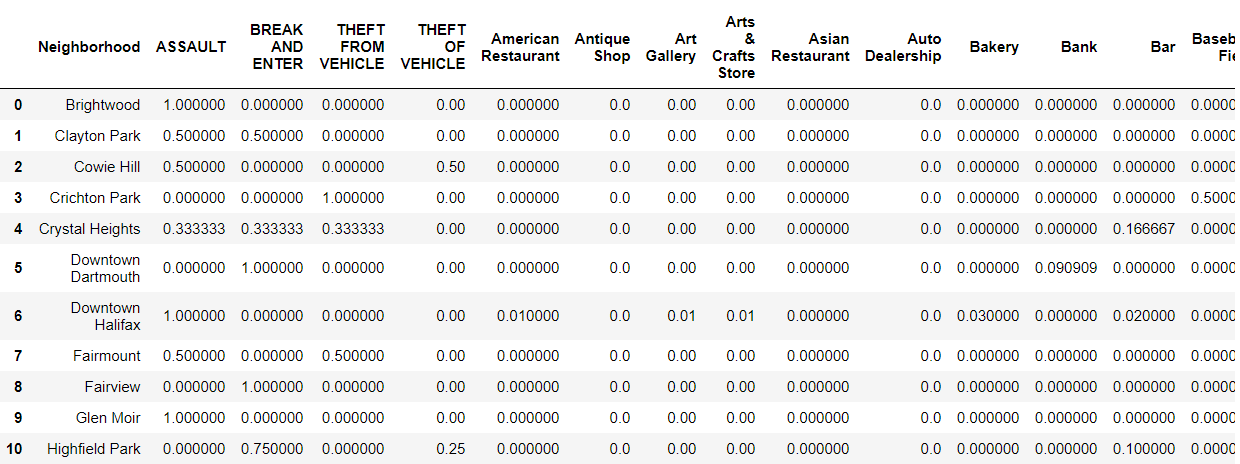
1. Next, I read the Json file with community boundaries of Halifax Area and made a function to check if coordinate in crime df is in one of the community areas. If yes then it adds the community name in GSA\_NAME column to the crime dataframe
2. I also made another function to check if the coordinate in crime is in within 5000 m radius of the neighborhoods in neighborhoods dataframe of step 1. If yes then it checks which of the Areas is closest to it and adds that Areas name to the Areas column in the crime dataframe
3. After the above 2 steps the crime dataframe becomes:



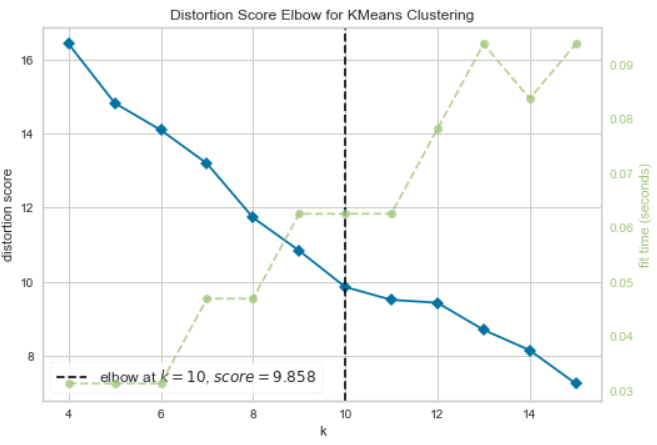
1. Next, I made a function that adds to the crime count columns in neighborhoods dataframe from step 1 for each crime in the crime dataframe that falls within 2000 meters of Areas:



1. I then used the foursquare API to get all the nearby venue categories (within 500 meters) from each Areas from the neighborhood dataframe.
2. I used this data to end up with the halifax\_grouped dataframe which contains the neighborhoods as the index and the venue categories in addition to the crime type categories as columns with the data showing the ratio for each category out of the total number of other categories in each neighborhood.

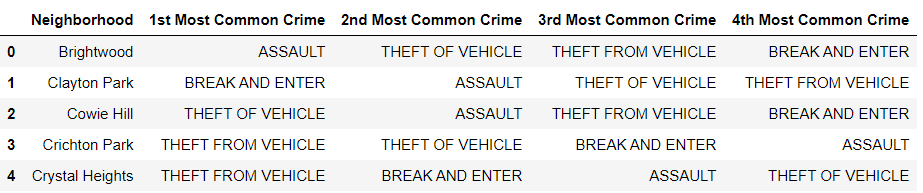


1. I then used the knn clustering algorithm elbow method to find the best k value to used for the knn clustering algorithm. The halifax\_grouped dataframe from the previous step was used for the model. It was found to be 10 clusters:



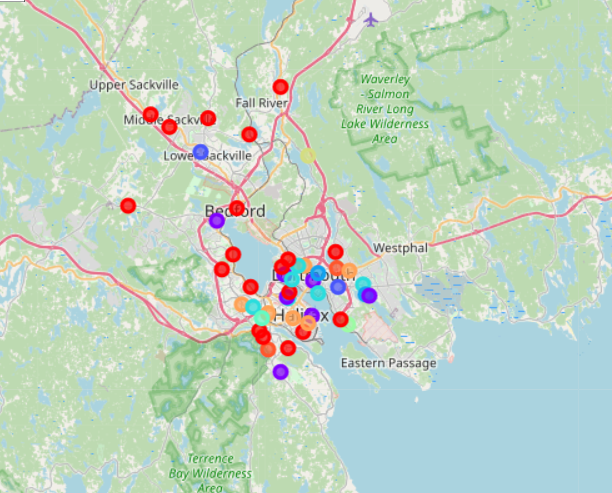
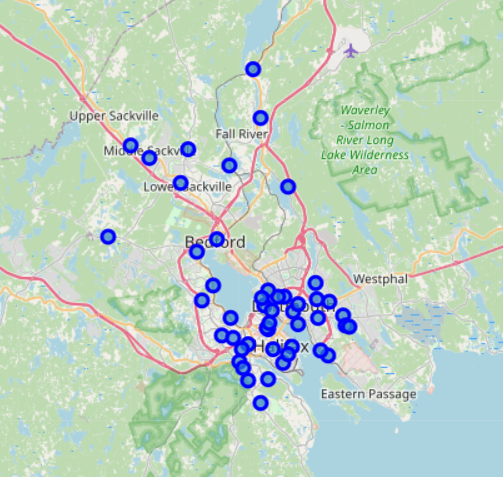
1. I also built 2 dataframes showing the most common venue categories in each neighborhood and the other showed the most common crime categories in each neighborhood to better understand the data.



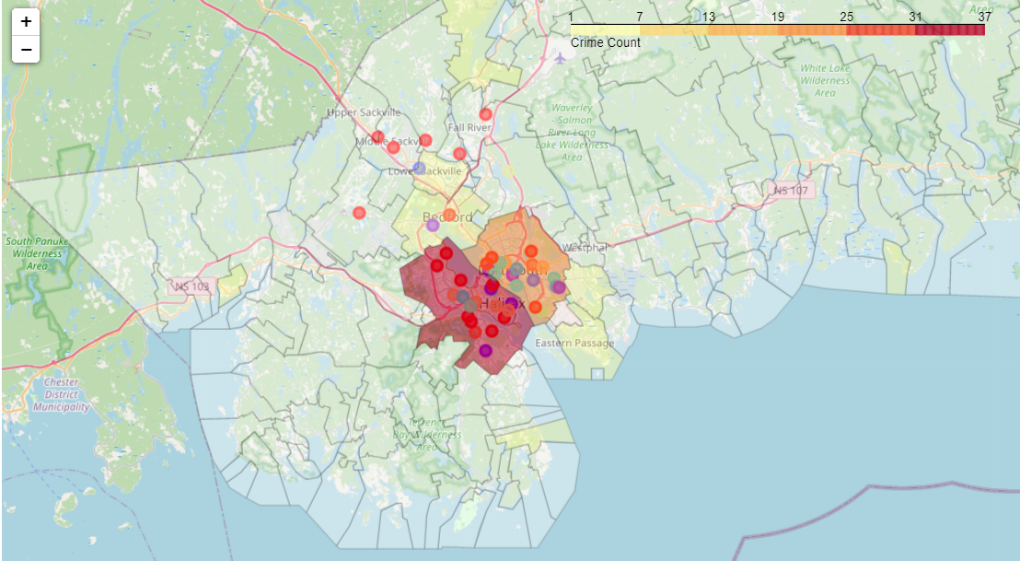


1. Results:

As a result, I obtained 10 different clusters using the kNN algorithm and I plotted these clusters on the map to better understand where these clusters were located. Below I present the map with the neighborhoods before they are identified to which cluster they belong to and beside it, a map with clusters identified.



Also, using the geojson file and GSA\_NAME column in the crime dataframe, I was able to map choropleth showing the community boundaries superimposed on these clusters with crime count for each community presented.



1. Discussion

Ten clusters were identified. The first cluster had a high incidence of theft from vehicles and theft of vehicles and the venue categories were mostly restaurants and tourist attractions. Likely with a high number of cars in parking lots. Cluster 2 was mostly areas with high assault rates. The most common venue categories were areas with sport related activities (hockey, skating etc.). The remaining clusters fell along those lines or well less defined.

1. Conclusion

To better understand this relationship and build a better model that can help the police department to focus resources by linking crime types to venue categories around a neighborhood, a bigger dataset will need to be used of crime data. I only had data pertaining to crimes committed for the month of January of 2021. Also, the city of Halifax has a smaller city with fewer venues and venues categories. It would be interesting to apply this project to a region with a larger population, greater number of crimes, and a greater number of venues.

Mahmoud Atwi

References

1. <https://en.wikipedia.org/wiki/Communities_in_the_Halifax_Regional_Municipality>
2. https://catalogue-hrm.opendata.arcgis.com/