COURSERA CAPSTONE

IBM Applied Data Science Capstone

ABSTRACT

Finding and Clustering Neighborhoods in Noida based on the availability of necessary/luxury commodities

Yash Singhal

June 2020

Introduction

Background

Noida is an organized and developed city in the National Capital Region of India, adjacent to the Capital, Delhi. It is composed of multiple parks, shopping complexes, offices, residential complexes, schools, local markets and general recreational spots. With a population of 600 thousand people, it is one of the large cities of India and plays an important role in the Region's development as it is home to multiple multinationals such as Samsung, Adobe, CSC, HP, HCL and many more. Being a planned city, Noida is divided into multiple sectors, with different sectors having different group of amenities.

Business Problem

I saw during the Current CoronaVirus Pandemic that groups of residential sectors are and were locked down to contain the spread of the Virus. In some cities, areas are being boxed off and "fortified" to contain the spread of the Virus. This can work well in the urban areas where people are normally in close contact but unfortunately in these zones, at this moment, there is a lack of shops and venues to fulfill the needs and desires of the residents. This presented the problem that there were sectors/groups of residential sectors without certain types of shops or amenities present, forcing the people to go out of their sectors to get said amenities. The Government and the Noida Authority would definitely want to know these locations so they could better develop these sectors and promote the building of the shops which are not there.

Idea

My idea is to cluster the sectors in Noida based on the availability of different types of shops/places in them. I have made a list of sectors in Noida and stored them in a CSV file. I would be using the Python Geocode to get the coordinates(Latitudes and Longitudes) of the sectors. Then I would find the nearby places using the Foursquare API. Finally, I would cluster the sectors and try to analyse and identify sectors that lack the basic amenities or different types of shops/venues.

Target Audience/Stakeholders

My Target audience is the Indian Government and the Noida Authority who would want to know which sectors need more development and work accordingly. With this data, they can make better decisions on government funding and planning of the residential sectors. Also, it would enable them to get more investment in these residential sectors, besides making the 'self-sustainment revolution' a far more likely possibility.

Additionally, other people who are looking to buy a new house or open a shop can look into this data to choose a location where more amenities are available or choose a new store location where existing stores are less in number. It also allows the person to find a new place to live or shop, which they might have never considered without this data.

Data

Data Needed

1. List of Residential Sectors in Noida
2. GeoLocation of Residential Sectors in Noida
3. Places near each sector within a kilometer from the sector's location(This is taken from the GeoLocation of the Sector which I have seen is generally at the center of the Sector)

Sources of Data

For the List of the Sectors in Residential Sectors in Noida, I am referring to the site: https://www.myloancare.in/noida-circle-rate-revised/, from which I would be making a list with the sector names and Noida at the end in the format: ***sector 50, noida***. To make this list, I will use pandas to get the list from the site and then I will be data wrangling to convert it into the needed format. Then I will get the geographical coordinates of the sectors using Python Geopy.Geocoder package which will give me the latitude and longitude coordinates of the sectors.

Finally, I would be using the Foursquare Places API to get the venue data for those sectors. Foursquare is a social location service that allows users to explore the world around them. The Foursquare API allows application developers to interact with the Foursquare platform. Foursquare API will provide many categories of the venue data, which I then use to count and find the availability of each of the categories of the venue in a particular sector. I would then be using my data science skills to cluster the sectors using K-Means Clustering, identifying the different sectors that the government needs to deal with.

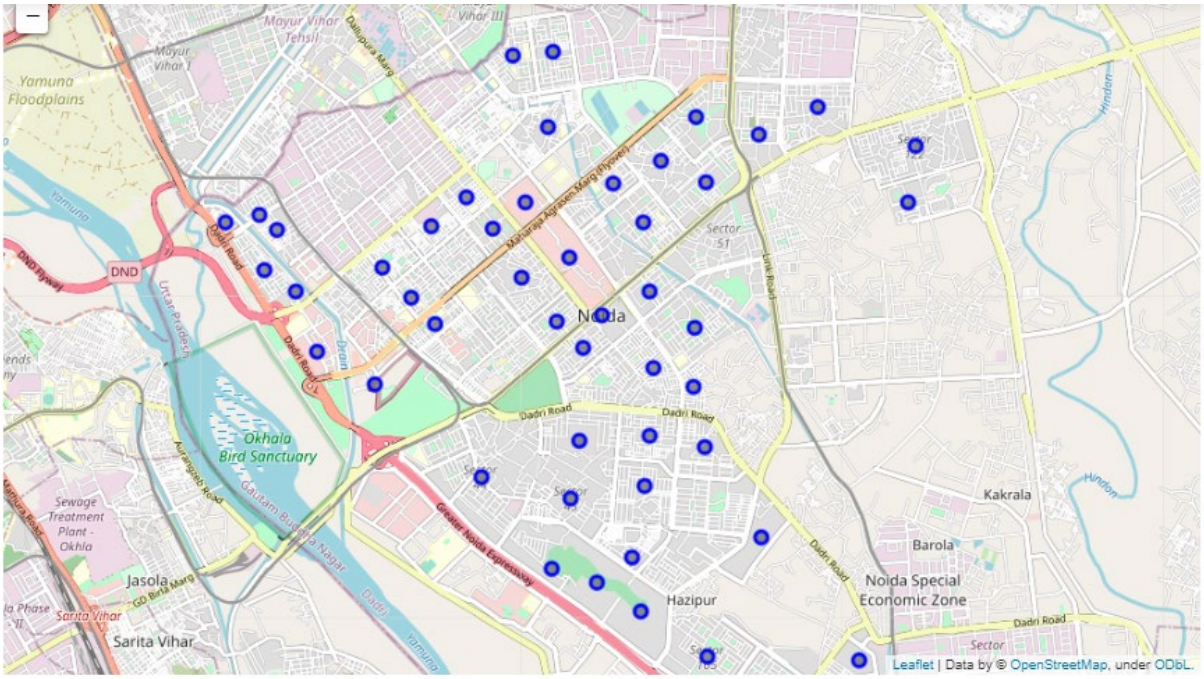
Methodology

Data Acquisition and Preprocessing

Sectors
Sector 11, 12, 16, 16A, 16B, 22, 24, 42, 43, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 104, 107, 110, 115, 117, 118, 119, 120, 121, 130, 133, 143, 143B, 144, 150, 151, 168
Sector 14, 14A, 15A, 17, 25A, 30, 32, 35, 36, 38A, 39, 44, 50, 51, 52, 92, 93, 93A, 93B, 96, 97, 98
Sector 15, 19, 20, 21, 23, 25, 26, 27, 28, 29, 31, 33, 34, 37, 38, 40, 41, 45, 46, 47, 48, 49, 53, 55, 56, 61, 62, 82, 99, 100, 105, 108, 122, 128, 129, 131, 134, 135, 137
Sector 63A, 86, 112, 113, 116
Sector 102, 158, 162

Using the methods explained in the sources of data, I firstly brought in the list of sectors in Noida. I had to clean the as it was not in a list on the site but was columns each containing multiple sector numbers as shown above. In order to process this data, I first replace the word 'Sector' in each column and then join them all in a single string. Finally I separate them in a list based on the comma. After getting my list of sectors, I use Geopy's geocoder Nominatim to find the coordinates of each sector which I then again store in a DataFrame.

I then map these points to visualize them as seen below:



Next, I connected to the Foursquare API and got all the venues in a 1000 meter radius. In order to analyze this data, I then group the data by sector and see a count of different categories. I would be later using this count data in my analysis.

Data Encoding and Clustering

In order to cluster my data, I would first have to one hot encode it to get a table of row of 0s with a 1 for each category. After one hot encoding it, I group the one hot encoded data by the sector and put the mean each category of venue for each category. Additionally, I add the number of different categories of venues in a locality to the dataframe, giving it an important piece of information. I normalize this venue data by dividing it by 25 which is a bit over the max categories in a sector, so this value doesn't overpower the mean value of other different types of venues in the region.(I tries without normalization and the data was too skewed towards the number of categories in a sector)

Finally, I use K-means Clustering to cluster the sectors into 3 categories that I believe are:

1. Fully Self Sufficient - Have low or none dependence on other sectors

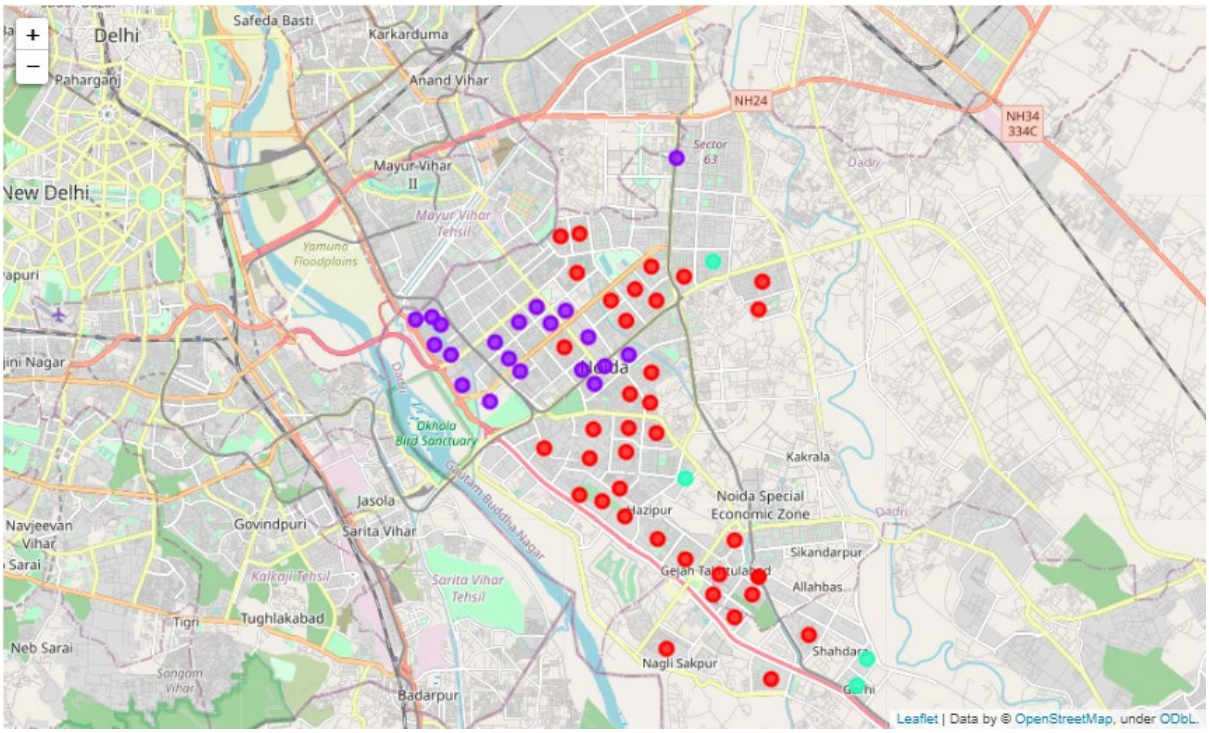
2. Partly Self Sufficient - Have partial dependence on other sectors

3. Barely Self Sufficient - Have high dependence on other sectors

I use K-Means clustering because I need to cluster the sectors into different cluster so I can recommend the sectors that need the most and least work.

Results

Analysis



On the above map, we can see the different clusters that are formed. Seeing the number of venue categories in each cluster mainly, I derived that the Blue cluster is the Highly Self Sufficient cluster, Red cluster is the Partly Self Sufficient cluster, and Green cluster is the Barely Self Sufficient cluster.

Discussion

Based on the above analysis and my knowledge of Noida, I can say that the Highly Self Sufficient sectors are those near the malls and shopping complexes and have a large variety and quantity of shops. The Partly self sufficient ones including mine are mostly composed of residential complexes and have a smaller number of shops. The Barely self sufficient ones are the ones on the outskirts of Noida and are recently formed, therefore don't have many shops yet.

Conclusion

In conclusion, I was able to find the sectors that need the most and least work in terms of opening of different types and number of shops and venues in Noida especially during a crisis such as the CoronaVirus Pandemic where sectors are sealed off. I would recommend the government to provide more resources to the Barely Self Sufficient sectors while developing shops there in the long term. For the Partly Self Sufficient Sectors, I would recommend the government to slowly increase the categories of venues available there and provide short term resources where needed.