IBM Applied Data Science Capstone

Best Stationery Shop Locations

5/7/2020

Project by:
PRASFUR TIWARI
prasfur9@gmail.com

INDEX

INTRODUCTION

"**Today, India has almost 15 lakh schools and is one of the biggest education systems of the world**" says a study conducted by Alok Kumar, adviser, NITI Aayog and Seema Bansal, director, social impact, Boston Consulting group.

With a population of over 1.38 billion, India is the second largest populated country in the whole world, just lagging behind China. Out of these, India counts for more than **315 million** students. Each student needs education, for which nearly every one of them attends a school.

Seeing the current scenario of India, there are innumerous schools; nearly one in every street, which ultimately increment the number of students. As we all would be familiar with the fact that there is a lot of homework and project work allotted to the students during school. To gain utter perfection in their work, the students require several stationery stuffs like pen, pencils, charts, project files, etc. This makes the business of a stationery shop much favorable to flourish and prosper.

Considering the school students as the target audience, this project aims to find the schools around a given location so that a business person, aiming to start a stationery shop, can receive assistance to pursue a profitable business.

Business Problem: For a business person to run a stationery shop business, it will be utterly problematic for his business to acquire profits if he sets his shop in a location with absence of school and students. Hence, this project aims to collect data about a particular location where many schools are located, so that his shop can attract more and more students.

This idea can solve the problems faced by students and make the business profitable as students require a lot of stationery stuff for their educational requirements.

DATA ANALYSIS

Data Collection/Gathering: The data used was the location of schools that was acquired using the **foursquare** website. To gather the data, foursquare API was used along with the foursquare credentials Client ID and Client Secret. A 'search' query was made in the **Jupyter notebook** with **Python kernel**, so as to search the schools. Through the website, coordinates of schools in a particular location were generated and processed.

To be specific about a region, **Kanpur** city was selected as the target area for project.

Some factual data to support this report was taken from here.

Data Preprocessing and Wrangling: Using the modules of python, only valid and usable data was selected from the JSON file generated by foursquare and data-frame was created using **'Pandas'**. Since the project required only the locations, the '**venues**' section under the '**response**' section was selected.

The generated data-frame still had numerous of unwanted data inside the 'venues' section, which needed to be filtered and cleaned. Several sections, like 'id', 'categories' etc. had irrelevant data and null values too.

	id	name	categories	referralId	hasPerk	location.address	location.lat	location.lng
0	4c53c90ab3b09c74cc1b1eb4	Gumti No. 5	[{'id': '4bf58dd8d48988d12b951735', 'name': 'B...	v-1588836227	False	Gumti	26.467596	80.312690
1	5303a0cb11d2eac30b61e9c6	HDFC Bank	[{'id': '4bf58dd8d48988d10a951735', 'name': 'B...	v-1588836227	False	110 / 189, R. K. Nagar G. T. Road	26.468362	80.317648
2	501ff6e3f470fc90bfb4f394	Fastrack Store	[{'id': '4bf58dd8d48988d103951735', 'name': 'C...	v-1588836227	False	Ground Floor ,Next To Wot, Z Square Mall	26.463888	80.332890
3	4e4287276365c15e6fa844e6	Reliance Digital	[{'id': '4bf58dd8d48988d122951735', 'name': 'E...	v-1588836227	False	Z-square Mall - Property NO 16/113, The Mall	26.458471	80.319865
4	5303a0c711d2eac30b617dc0	HDFC Bank	[{'id': '4bf58dd8d48988d10a951735', 'name': 'B...	v-1588836227	False	124/248, C Blk	26.451347	80.309307

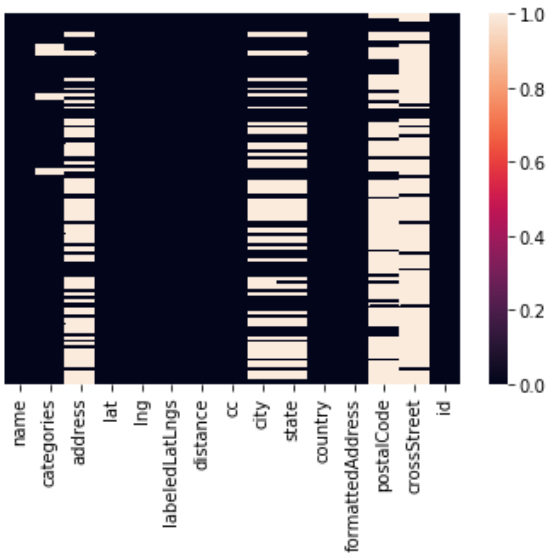
Data Filtering: Since one can work on a dataset if it is understandable. Hence, the dataset was filtered by keeping only the relevant data.

The ‘category’ column had a set of values like ‘id’, ‘name’ etc. and some of them were actually redundant for our dataset. Hence, only the ‘name’ of the ‘category’ was filtered, which gave some detail of the location.

	name	categories	address	lat	lng	labeledLatLngs	distance	cc	city	state	country	formattedAddress
0	Gumti No. 5	Bus Line	Gumti	26.467596	80.312690	[{'label': 'display', 'lat': 26.46759554100716...	1170	IN	Kānpur	Uttar Pradesh	India	[Gumti, Kānpur, Uttar Pradesh, India]
1	HDFC Bank	Bank	110 / 189, R. K. Nagar G. T. Road	26.468362	80.317648	[{'label': 'display', 'lat': 26.468362, 'lng':...	924	IN	Kānpur	Uttar Pradesh	India	[110 / 189, R. K. Nagar G. T. Road, Kānpur, Ut...
2	Fastrack Store	Clothing Store	Ground Floor ,Next To Wot, Z Square Mall	26.463888	80.332890	[{'label': 'display', 'lat': 26.463888, 'lng':...	1157	IN	Kānpur	Uttar Pradesh	India	[Ground Floor ,Next To Wot, Z Square Mall, Kān...
3	Reliance Digital	Electronics Store	Z-square Mall - Property NO 16/113, The Mall	26.458471	80.319865	[{'label': 'display', 'lat': 26.458471268671 19...	330	IN	Kānpur	Uttar Pradesh	India	[Z-square Mall - Property NO 16/113, The Mall ...
4	HDFC Bank	Bank	124/248, C Blk	26.451347	80.309307	[{'label': 'display', 'lat': 26.451347, 'lng':...	1635	IN	Kanpur Nagar	Uttar Pradesh	India	[124/248, C Blk (Govind Nagar), Kanpur Nagar 2...
...	...	...	...	...	...	...	...	...	...	...	...	...

Data Cleaning: To fit the model, one needs to get rid of the null values. Hence, the firstly, the columns with null, none or NaN values, were identified. A count of the null values from each column was taken and also, a Heatmap was generated to check the uncertainties in the dataset.

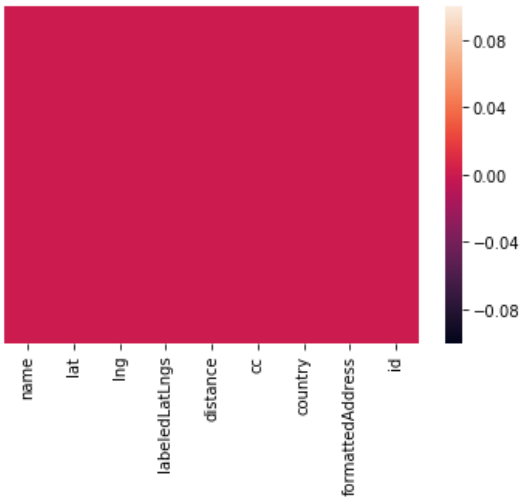
```
name                0
categories          8
address            67
lat                 0
lng                 0
labeledLatLngs      0
distance            0
cc                  0
city               54
state              53
country             0
formattedAddress    0
postalCode         87
crossStreet       101
id                  0
dtype: int64
```



The cleaning of data was done by removing the columns with NaN or null values. The columns were dropped keeping in mind whether they were really useful for analysis purpose or not.

After removing the columns, the Heatmap was regenerated for verification.

```
name                0
lat                 0
lng                 0
labeledLatLngs      0
distance            0
cc                  0
country             0
formattedAddress    0
id                  0
dtype: int64
```



We know that for clustering purpose, we need some valid data so as to generate relationship between the rows and make clusters. Hence, some more columns, unusable for clustering purpose, were also removed. These include ‘name’, ‘country’, ‘postal code’ etc.

Data Standardization: Standardization of data is required to keep the data in a suitable range and avoid large variations among data. This helps to fit the model in a more perfect way.

To do this, the scikit-learn module of python was used and standard values for each column were obtained in the form of an array.

```
array([[ 1.82301382e-01, -6.03723211e-01, -2.39196794e-01],
       [ 2.97747002e-01,  9.19201179e-04, -8.65179892e-01],
       [-3.76136025e-01,  1.85980092e+00, -2.72277201e-01],
       [-1.19201517e+00,  2.71269279e-01, -2.37670006e+00],
       [-2.26508709e+00, -1.01633132e+00,  9.44063941e-01],
       [-3.55822226e-01, -1.76226059e-01, -1.55477916e+00],
       [-1.64505825e-02,  2.74470684e-01, -1.62602927e+00],
       [-2.00819806e+00, -1.02050641e+00,  6.53974212e-01],
       [-3.30541050e-01,  1.89089637e+00, -1.85759212e-01],
       [ 9.27520123e-01, -4.08643364e-01,  5.82724104e-01],
       [ 3.15601245e-01, -5.67160913e-01, -1.29776984e-01],
       [ 7.91224833e-02, -8.66472564e-02, -1.12473386e+00],
       [-2.01264454e+00, -1.06023615e+00,  7.25224321e-01],
       [ 9.77224579e-01, -1.50533202e-01,  4.32589946e-01],
       [-2.26508709e+00, -1.01633132e+00,  9.44063941e-01],
       [ 1.90322737e+00,  4.59358801e-01,  1.91357435e+00],
       [-1.11102098e+00,  3.73378509e-01, -2.61589685e+00],
       [ 1.64912799e+00, -3.22756962e-01,  1.74053837e+00],
       [ 9.98591375e-01,  1.58366800e-01,  2.85000435e-01],
       [-2.07635766e+00, -1.10011637e+00,  8.62635245e-01],
       [-2.57175357e+00, -1.14670427e+00,  1.53187734e+00],
       [-3.54590846e-02, -5.15772352e-01, -6.33617038e-01],
       [ 8.30647796e-01, -2.85437857e-01,  3.00268315e-01],
       [ 7.55336774e-01,  3.36547235e-01, -2.26473560e-01],
       [-5.04356731e-01,  2.59472063e+00,  1.17562679e+00],
       [-5.12062827e-01,  2.55567688e-01, -2.43777158e+00],
       [ 5.50977224e-01,  2.42923034e-01, -5.75090163e-01],
       [-4.72890916e-01, -1.09143902e+00,  1.60312745e-01],
```

MODEL FITTING

Clustering: This is a machine learning techniques which is used to make clusters based on the similarity of the data values.

The clustering process was started and the ‘k-means’ clustering algorithm was used. In this algorithm, the value of ‘k’ signifies the number of clusters one wishes to generate. To keep enough number of schools in the clusters, k was chosen to be 4. The model was the fitted and the labels were generated in the form of an array.

Since there were 4 clusters, the labels ranged from 0 to 3. These labels were concatenated with the data-frame so that the rows and their assigned labels can be kept and viewed together.

The data-frame then looked like:

	lat	lng	distance	Labels
0	26.467596	80.312690	1170	3
1	26.468362	80.317648	924	2
2	26.463888	80.332890	1157	1
3	26.458471	80.319865	330	2
4	26.451347	80.309307	1635	0

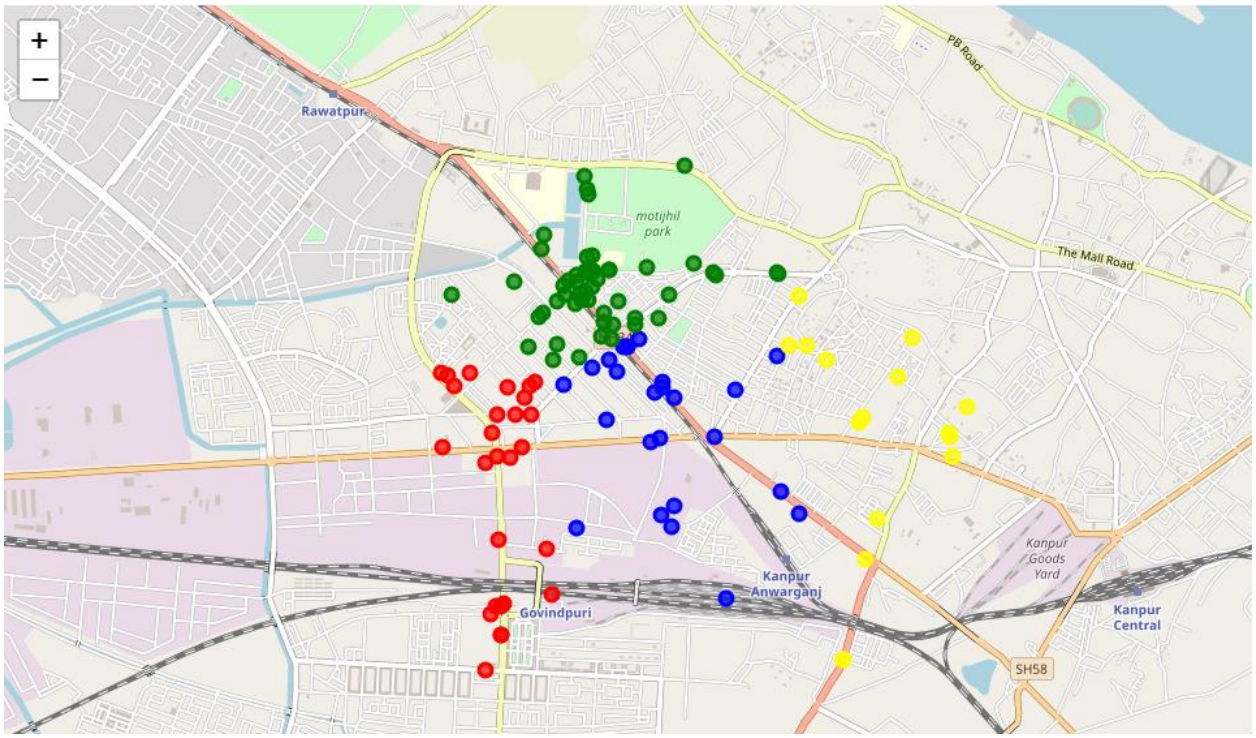
Color Coding: To view the clusters separately and help visualization, each label was assigned with a color code.

The color codes used in this project are:

Labels	Color Code
0	Red
1	Yellow
2	Blue
3	Green

DATA VISUALIZATION

Visualizing Clusters: All the 4 clusters were visualized on a world map centered on Kanpur. The color coding was applied while visualizing for differentiating between the clusters.

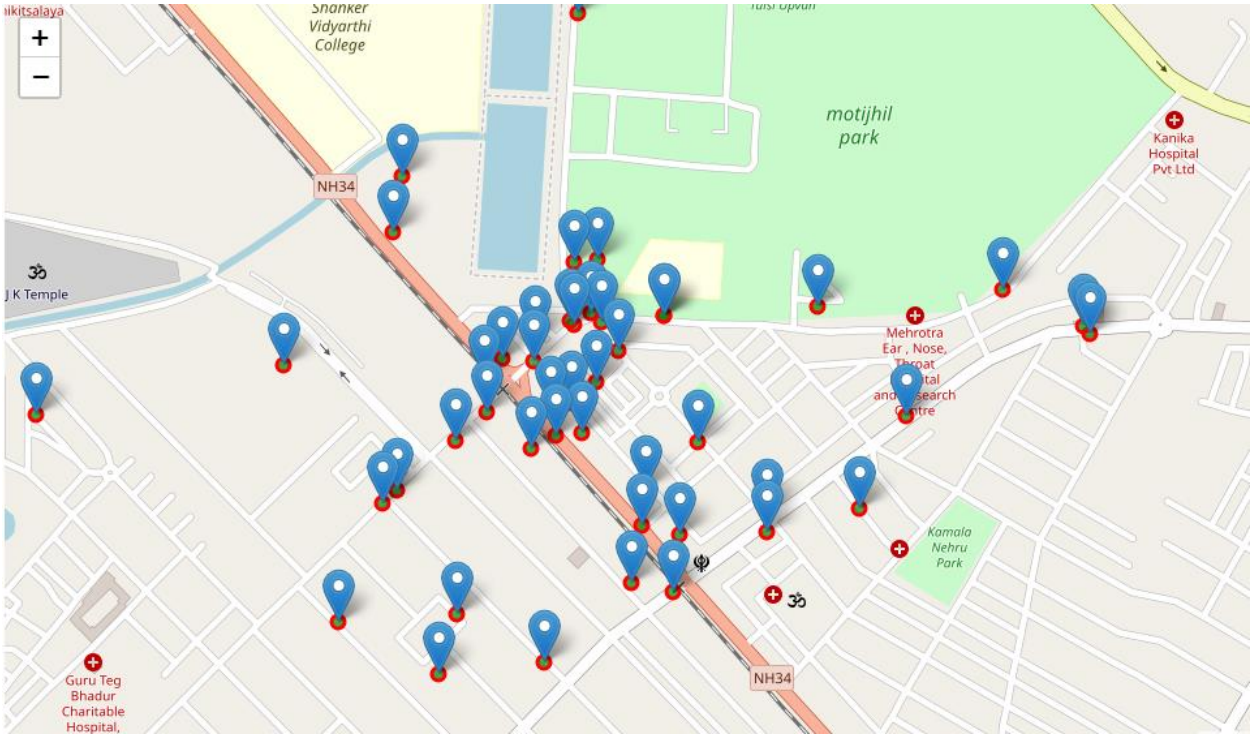


Partitioning Dataset: Here, dataset was separated into 4 partitions to make it easy to be understood.

The count of schools in each cluster was taken so as to get the best cluster for the stationery shop. The count was:

Cluster	Schools
1	27
2	16
3	25
4	51

Visualizing target cluster: In the final cell, the 4th cluster was visualized as it has most number of schools. It was displayed along with pop-ups displaying the name of the places so that the business person can easily know about the schools and landmarks of the area.



Observation: It was noted that the 4th cluster had 51 schools. It's logical to assume that more and more schools will attract more students. Hence, the areas inside the 4th cluster qualify to become the best suited place for a stationery shop business.

Project URL:

https://nbviewer.jupyter.org/github/Prasfur/Coursera_Capstone/blob/master/Applied%20Data%20Science%20Capstone.ipynb