BEST STATIONERY SHOP LOCATION

Project by:
PRASFUR TIWARI

Introduction

- India's population is around 1.38 billion and is a hub of nearly **315 students**.
- There are **15 lakh schools** in India and is one of the biggest education systems in the world.
- Every student needs education, and hence, nearly all of them attend schools.

Business Problem

- With the increase in number of students, the number of schools will ultimately rise.
- Every student is allotted with some homework or project work which requires a lot of stationery stuff.
- This generates a need for a stationery shop to aid the students with their educational work.

Solution

- The solution is to set up a stationery shop in a region where many schools are located.
- This problem can be solved using Data Science.
- The locations of schools can be gathered and clustered in different groups so as to get a suitable location for the shop.

Data Gathering/Collection

- Using the website **Foursquare**, the data was gathered by generating a search query on **Jupyter notebook** with **Python kernel**.
- The necessary foursquare credentials, i.e. the **client id** and **client secret** were passed in the query.
- To be specific, we aim to find the schools in **Kanpur** city, Uttar Pradesh.

Data Preprocessing and Wrangling

- Using the ***pandas*** module of python, only useful and relevant data required for the analysis was filtered.
- We extract on the **'venues'** data under the **'response'** section.
- This dataset still contains irrelevant data and needs to be filtered and cleaned.

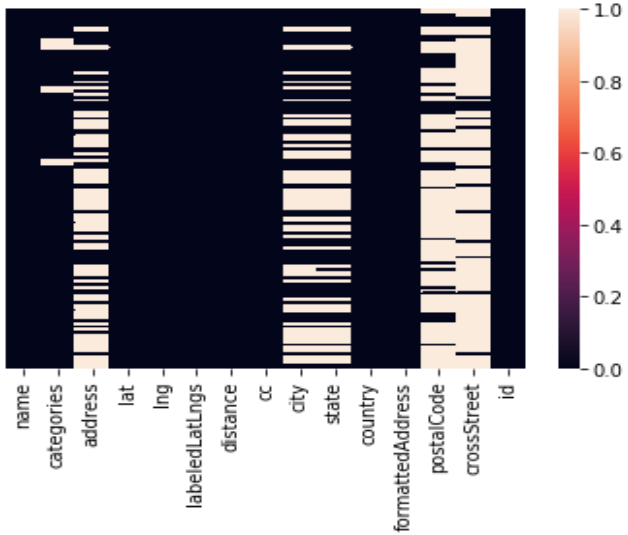
Data Filtering

- 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.
- This make data more a bit more readable.

Data Cleaning

- Null, none and NaN values are detected and irrelevant columns are removed.
- The columns are removed keeping the fact in mind that necessary data isn't removed.

```
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
```



Data Standardization

- The ***scikit-learn*** module is used so as to standardize the data.
- An array of standard values is returned by the scikit-learn module.
- The data is standardized to get a suitable range of data for model fitting.

Model Fitting

- The ***k-means*** algorithm is selected to perform **clustering.**
- The value of 'k' is selected as **4** for this project.
- The model is fitted using the scikit-learn module and labels for each data values are generated, ranging from 0 to 3.

	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

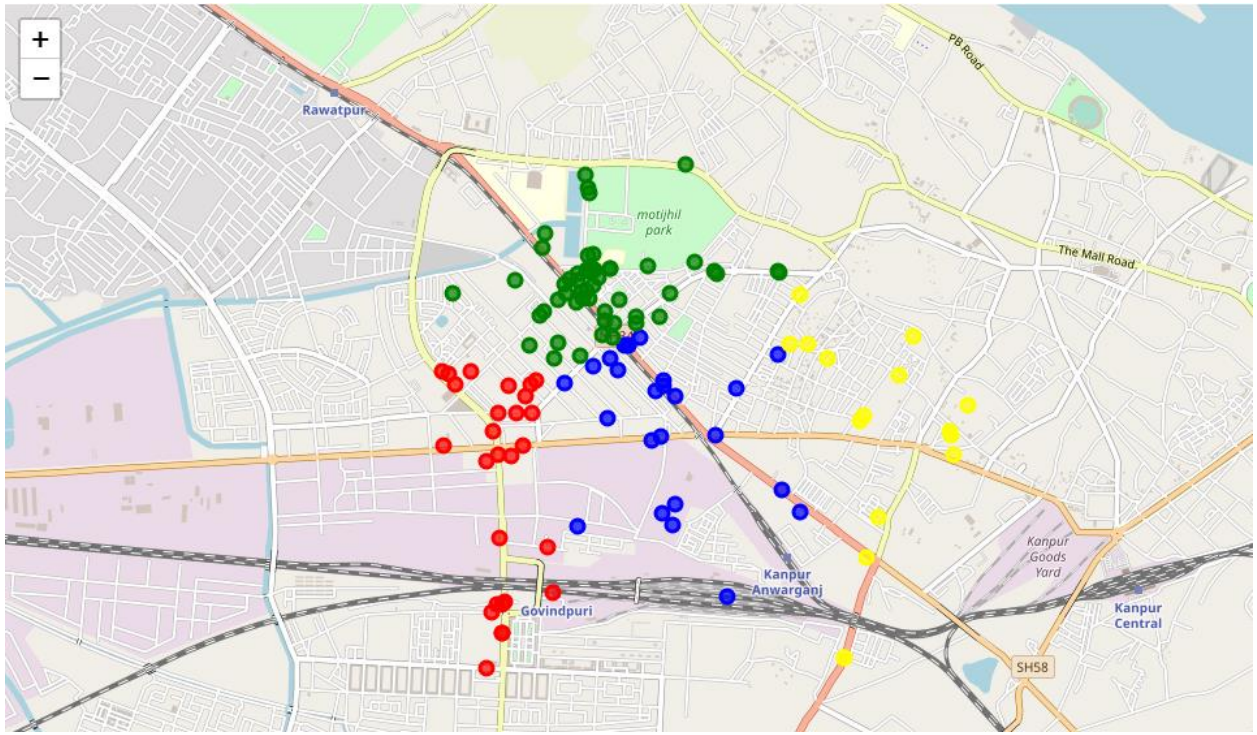
Color Coding

- To view the clusters separately and help visualization, each label was assigned with a color code.
- The color codes used in this project are:

Label	Color Code
1	**Red**
2	**Yellow**
3	**Blue**
4	**Green**

Data Visualization

- The clusters, based on their color codes, are visualized on a world map centered on Kanpur.



Finding Count

- The number of schools in each cluster is noted so as to find the best suited cluster or location for the stationery shop.

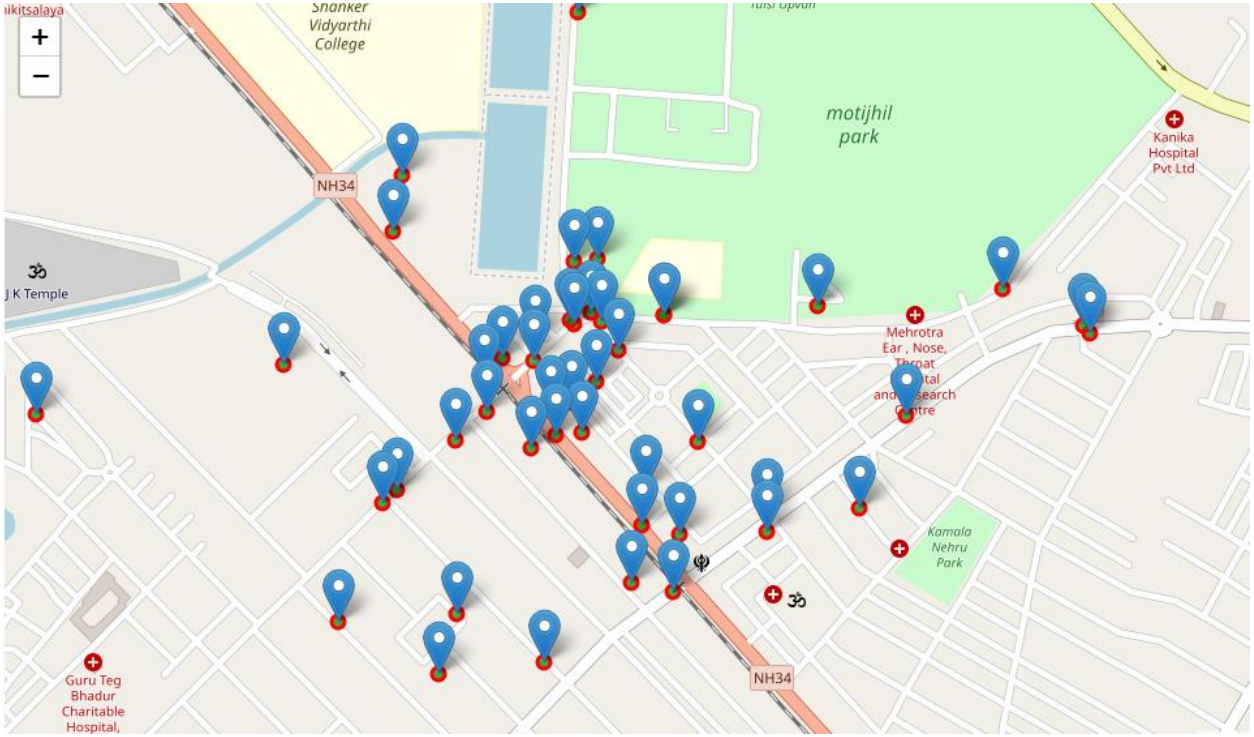
- It comes out to be:

Cluster	Schools
1	27
2	16
3	25
4	**51**

Visualizing Target Cluster

- Since the 4th cluster has the most number of schools, it qualifies to be the best location for the shop.

- Visualizing 4th cluster with pop-ups to get a better idea 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