2021

Best Staying location for Tourists

Divyansh Tiwari

https://linkedin.com/in/dt03

5/1/2021

Index

1. Introduction

2. Data Analysis
   2.1. Data Gathering
   2.2. Data Preprocessing
   2.3. Data Filtering
   2.4. Data Cleaning

3. Model Fitting
   3.1.Clustering

4. Visualization
   4.1. Visualizing Clusters

5. Observations and Conclusions

Introduction

We all love to visit and explore new places. Many a times our vacations are very short, because we decided to take a short break from our regular schedule to refresh ourselves. Or we have a tight budget for a vacation that is compulsory for us to go to. In both cases we'll want efficient use of our money and time at our destination place.

But what if your hotel that you booked is too far away from where all the main tourist attractions are? Then this is not only going to waste your money in travelling to and fro from hotel and other intercity travels but is also going to waste a lot of time. This will be a even bigger issue in cities like Mumbai where due to high traffic at peak hours even travelling short distances take a long time. And when you are on a vacation you won't want to spend most of your time in a cab amidst the city with no scenic view outside your window.

This project aims to find the best location for a tourist visiting Mumbai with respect to all the "Popular among visitor" places and Tourist attractions so that he can use his time and money efficiently. This saves a lot of his research time before the trip too by showing him the main tourist attractions and considerably reducing the locations he should look for staying.

Data Analysis

Data 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 "Popular among visitor" Places in Mumbai.

The API responded with JSON data of the main tourist attractions in Mumbai, and their other details like coordinates, Address etc.

Data Preprocessing

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 'items' [which was under 0^{th} index of 'group' section in the 'response'] section was selected.

Data Filtering

Only relevant and useful data was taken from the venues section of the JSON data.

Fields like 'Name', 'formattedAddress', 'lat', 'lng' , 'Categories' were kept and all other were discarded. Further only 'name' from 'categories' section was kept as the category of the location. This entire data was stored in a pandas DataFrame.

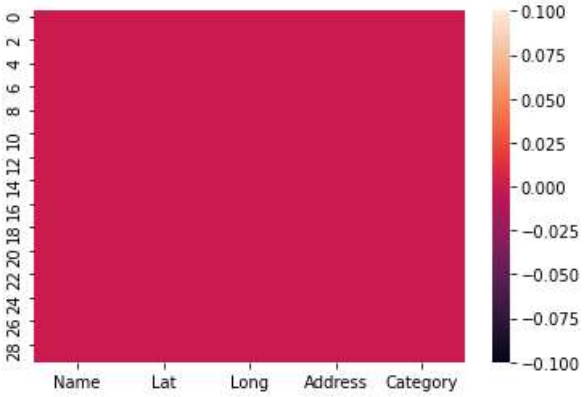
Out[3]:

	Name	Lat	Long	Address	Category
0	Taj Mahal Palace & Tower	18.922306	72.833578	Apollo Bunder (P. Jetha Singh Ramchandani Marg...	Hotel
1	Prithvi Theatre	19.106157	72.825810	Church Road (Juhu), Mumbai 400049, Mahārāshtra...	Theater
2	Jogger's Park	19.059728	72.822055	Carter Road (Bandra West), Mumbai 400 050, Mah...	Park
3	Nariman Point	18.929183	72.822232	Nariman Point (Dorabji Tata Road), Mumbai 4000...	Scenic Lookout
4	Starbucks Coffee Capital	19.063457	72.861576	The Capital, India	Coffee Shop

Data Cleaning

To fit the model, one needs to get rid of the null values. Hence, first we need to identify the columns with null, none or NaN values. A count of the null values from each column was taken and also, a Heatmap was generated to check the uncertainties in the dataset.

Out[6]: <AxesSubplot:>



Out[7]:
```
Name        0
Lat         0
Long        0
Address     0
Category    0
dtype: int64
```

We don’t have any null values in our Dataset. Hence we proceed with Clustering the places together.

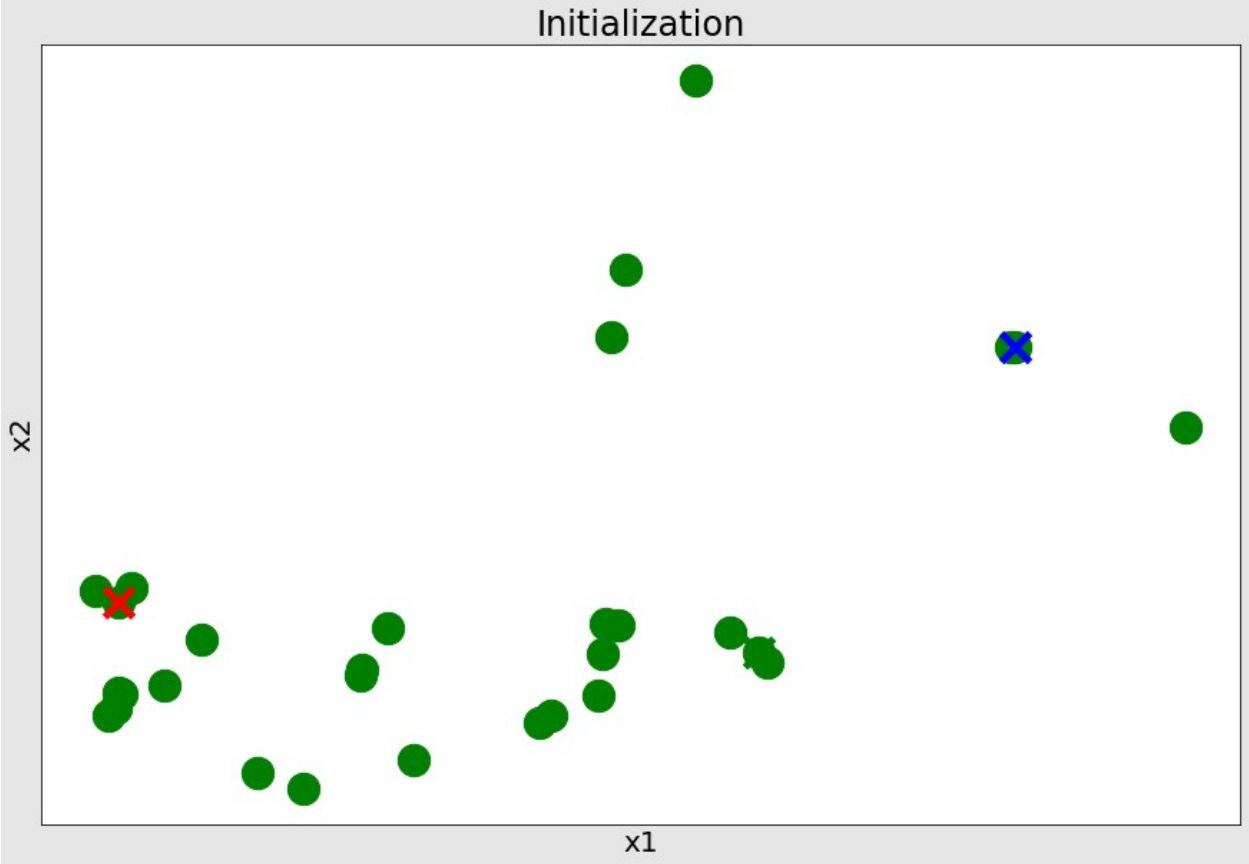
Model Fitting

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 places to visit in the clusters, k was chosen to be 3. The model was the fitted and the labels were generated in the form of an array. Since there were 3 clusters, the labels ranged from 0 to 2. Each cluster was assigned a different color (red, blue, green).

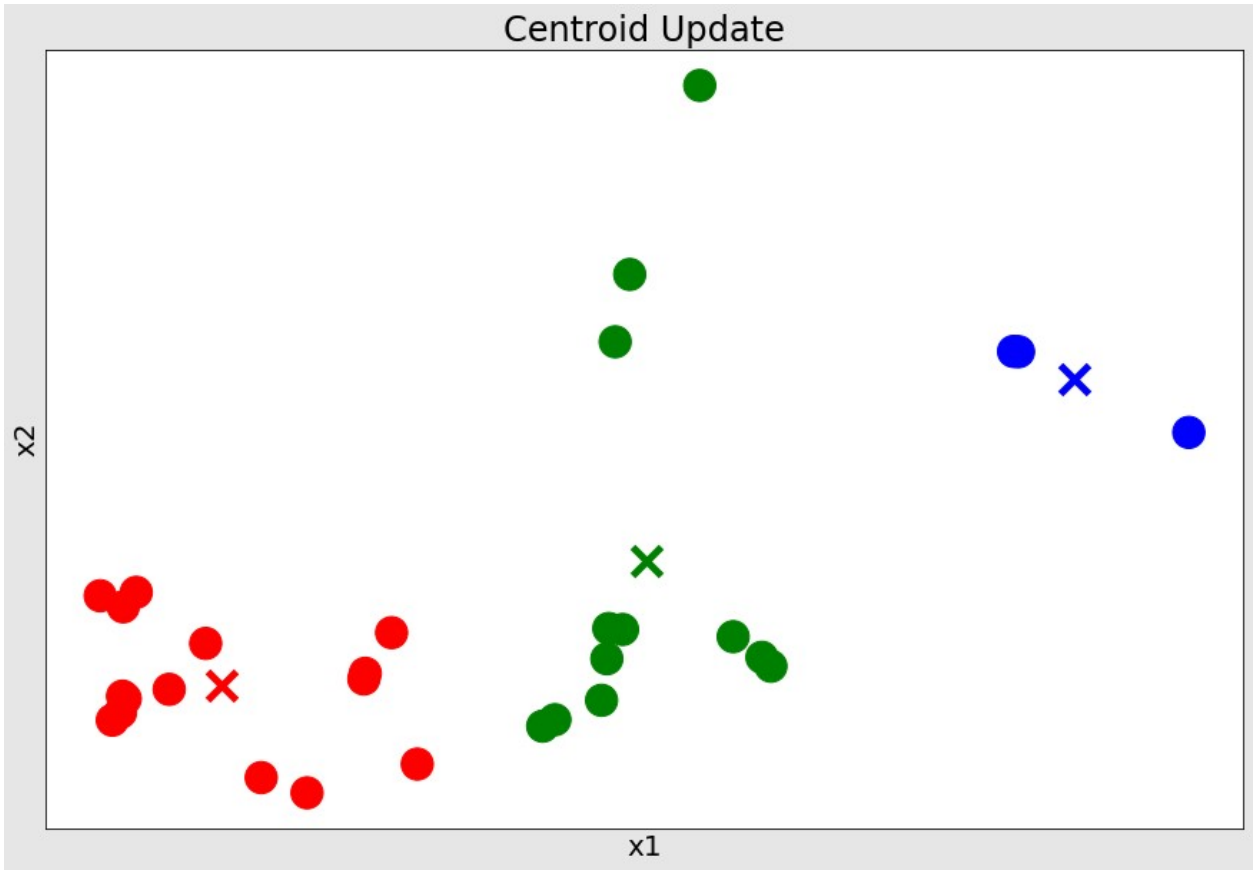
Initial:



Here x1 and x2 are the latitude and longitude plotted against the x and y axis respectively. The 3 X represent the 3 initial places and the green circles represent the places plotted by their lat and long.

Final Clusters:

The clusters were color coded. The places belonging to Cluster 1 were coded as red, cluster 2 as blue and cluster 3 as green.

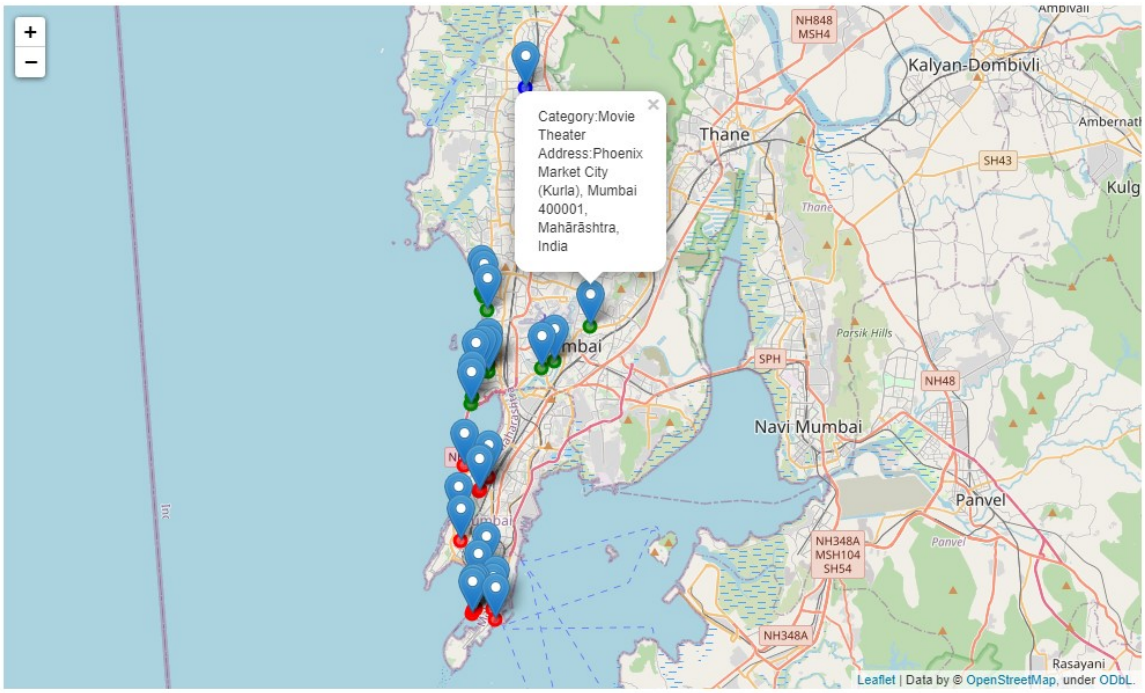

Centroid Update
x2
x1

Data Visualization

Visualizing Clusters

All the clusters were visualized on a world map centered on Mumbai. The color coding was applied while visualizing for differentiating between the clusters. Folium Library was used to visualize the clusters.

Out[78]:



Proper labels were attached to the plotting so that the user can get the category of the place and the address.

Observation and Conclusion

We observed most of the tourist places are centered around south east of Mumbai and hence the visitor should look for a hotel in that area.

Link: https://nbviewer.jupyter.org/github/DivyT-03/BestStayingPlace/blob/main/BestStayPlace.ipynb