Exploring venues in Bangalore,India

Supragya Bajpai

May 15, 2020

1. Introduction

1.1 Background

Bangalore also known as the “The Silicon Valley Of India” is a hotspot of all the HQ of tech Giants in India.Due to this reason there are a lot of freelancers who visit the city frequently.With this project I am trying to come up with a solution for these people to explore the venues present nearby.

1.2 Problem

Data that might contribute to exploring the different venues in bangalore,with the help of data science we will come with a solution to predict the most visited venues in a cluster.

1.3 Interest

All the freelancers who are visiting the city can explore these venues..

2. Data acquisition and cleaning

2.1 Data sources

I have collected the data of area wise postal code of bangalore though a India TV website and downloaded a csv file from the internet having latitudes and longitudes of all the postal codes of india .

Finally I have used Foursquare API for collecting all the nearby venues data.

2.2 Data cleaning

Data downloaded or scraped from multiple sources were combined into one table. There were a lot of missing values, because of lack of record keeping.There are several problems with the First datasets.It contained all the postal code of the neighborhood of india, I cleaned it out and formed a new table having only required data.After fixing these problems, I checked for outliers in the data. I found there were some extreme outliers, mostly caused by some types of small sample size problem and then finally created a table by merging the two data sets .

	IN	Postal Code	Kakana	Andaman & Nicobar Islands	01	Nicobar	638	Carnicobar	Unnamed: 8	lat	lng	4
0	IN	744301	Sawai	Andaman & Nicobar Islands	1	Nicobar	638.0	Carnicobar	NaN	7.5166	93.6031	4
1	IN	744301	Lapathy	Andaman & Nicobar Islands	1	Nicobar	638.0	Carnicobar	NaN	9.1833	92.7667	3
2	IN	744301	Mus	Andaman & Nicobar Islands	1	Nicobar	638.0	Carnicobar	NaN	9.2333	92.7833	4
3	IN	744301	Carnicobar	Andaman & Nicobar Islands	1	Nicobar	638.0	Carnicobar	NaN	9.1833	92.7667	3
4	IN	744302	Campbelbay	Andaman & Nicobar Islands	1	Nicobar	638.0	Nancowrie	NaN	9.1833	92.7667	1

	Neighborhood	Postal Code	lat	lng
0	Bangalore International Airport	560300	13.0082	77.5293
1	Benson Town	560046	13.2257	77.5750
2	C.V.Raman Nagar	560093	13.1077	77.5810
3	Domlur	560071	13.2257	77.5750
4	Fraser Town	560005	12.9910	77.5843

Final DATASET

2.3 Feature selection

I had to drop multiple columns from the dataset like state and Div from the dataset to make it more efficient in Data Analysis

2.4 Algorithm

I have used K_means algorithm which is one the best clustering algorithm which clusters different data points which are near to the centroid.

K-Means Clustering

In [73]:
```
kclusters = 5
from sklearn.cluster import KMeans

bang_grouped_clustering = bang_grouped.drop('Neighborhoods', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(bang_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:100]
```

Out[73]:
```
array([0, 0, 1, 0, 4, 1, 0, 1, 0, 0, 1, 1, 1, 4, 4, 0, 1, 3, 4, 0, 1, 1,
       0, 1, 1, 0, 1, 0, 2, 4, 4, 0, 0], dtype=int32)
```

In [74]:
```
neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)
```

In [75]:
```
bang_merged =df
# merge toronto_grouped with toronto_data to add latitude/longitude for each neighborhood
bang_merged = bang_merged.join(neighborhoods_venues_sorted.set_index('Neighborhood'), on='Neighborhood')
bang_merged
```

Out[75]:

	Neighborhood	Postal Code	lat	lng	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	Bangalore International Airport	560300	13.0082	77.5293	1	Indian Restaurant	Ice Cream Shop	Coffee Shop	Italian Restaurant	Multiplex	Fast Food Restauran
1	Benson Town	560046	13.2257	77.5750	0	Resort	Cupcake Shop	Light Rail Station	Vegetarian / Vegan Restaurant	Flea Market	Cricket Ground
2	C.V.Raman Nagar	560093	13.1077	77.5810	4	Café	Indian Restaurant	Chinese Restaurant	Ice Cream Shop	Pizza Place	Smoke Shop
3	Domlur	560071	13.2257	77.5750	0	Resort	Cupcake	Light Rail	Vegetarian	Flea	Cricket

3. Conclusions

In this study, I analyzed the most common venues of a particular area so that our explorer can visit them if he is willing to.

:

	Neighborhood	Postal Code	lat	lng	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	Bangalore International Airport	560300	13.0082	77.5293	1	Indian Restaurant	Ice Cream Shop	Coffee Shop	Italian Restaurant	Multiplex	Fast Food Restauran
1	Benson Town	560046	13.2257	77.5750	0	Resort	Cupcake Shop	Light Rail Station	Vegetarian / Vegan Restaurant	Flea Market	Cricket Ground
2	C.V.Raman Nagar	560093	13.1077	77.5810	4	Café	Indian Restaurant	Chinese Restaurant	Ice Cream Shop	Pizza Place	Smoke Shop
3	Domlur	560071	13.2257	77.5750	0	Resort	Cupcake Shop	Light Rail Station	Vegetarian / Vegan Restaurant	Flea Market	Cricket Ground
4	Fraser Town	560005	12.9910	77.5843	1	Indian Restaurant	Hotel	Ice Cream Shop	Pub	Lounge	Bakery
5	G.K.V.K.	560065	12.9881	77.5052	1	Indian Restaurant	Ice Cream Shop	Café	Pizza Place	Seafood Restaurant	Gym
6	H.K.P. Road	560051	12.9979	77.5861	1	Indian Restaurant	Hotel	Ice Cream Shop	Pub	Lounge	Burger Joint
7	Indiranagar Bangalore	560038	13.2257	77.5750	0	Resort	Cupcake Shop	Light Rail Station	Vegetarian / Vegan Restaurant	Flea Market	Cricket Ground
8	Maruthi Sevanagar	560033	13.0108	77.7494	1	Hotel	Indian Restaurant	Café	Pizza Place	Restaurant	Ice Cream Shop
9	Sadashivanagar	560080	13.2257	77.5750	0	Resort	Cupcake Shop	Light Rail Station	Vegetarian / Vegan Restaurant	Flea Market	Cricket Ground