ANALYSING BENGALURU NEIGHBORHOODS FOR STARTING A RESTAURANT BUSINESS

DATA AQUISTION AND CLEANING

Introduction:

I had been searching for some data suitable for analyzing neighborhoods of a city which I personally knew, but there wasn't a lot of data available for Indian cities suitable for the project. I found this kaggle dataset interesting and suitable for our foursquare analysis. It has the restaurant data from the city of Bengaluru(Banglore) scraped from the website of Zomato - the top restaurant review and food delivery service.

Data Acquisition:

The data used in the project are acquired from three sources:

- Restaurant data from the kaggle dataset mentioned above
- Geographical coordinates obtained using geopy package
- Venues and interests for the neighborhood locations using Foursquare API

The restaurant dataset from kaggle is in comma separated value format and is of size 547 MB. It contains the following fields:

Field name	**No of unique values**	**Description**
url	51,717	url of the restaurant's page in the zomato website
address	11,495	restaurant address as listed in the zomato website
name	8,792	name of the restaurant
online order	2	whether online ordering is available in the restaurant or not
book table	2	table book option available or not
rate	64	overall rating of the restaurant out of 5
votes	2328	total number of votes for rating
phone	64	restaurant phone number
location	93	neighborhood in which the restaurant is located

rest type	93	restaurant category (like quickbites, bakery etc.)
dish liked	5271	dishes people liked in the restaurant
cuisines	2723	cuisine types available in the restaurant
approx cost (for two people)	70	approximate cost of meal for two people
reviews list	22513	list of tuples containing customer reviews and ratings for the restaurant
menu item	9098	menu items available in the restaurant
listed in(type)	7	type of cuisine in which the restaurant is listed
listed in(city)	30	neighborhood in which the restaurant is listed

Data Organizing and Cleaning:

The restaurant dataset in CSV format is imported into a pandas data frame and is checked for duplicates and null values. The irrelevant fields from the dataset like the url field is removed. The ambiguous data from certain fields are corrected. The fields review list and menu item are list of tuple data which has to be converted into separate lists. The data is checked for other inconsistencies.

Once the data is cleaned latitude and longitude coordinates for each neighborhood in the dataset is obtained by calling the geocoder in the geopy package. After getting the coordinate values it is used to get the venue details from the Foursquare API.

Now the dataset is ready for exploratory data analysis and running the clustering model.