Identifying the best hotels of Chennai using Zomato API

Cyril Bastin M

August 1, 2020

1. Introduction

1.1. Background

Zomato is the most commonly used service for hotels in Chennai, Tamil Nadu. It has an API called "Zomato API"[1]. Zomato provides details of hotels and provide the following information: Search for restaurants by name, cuisine, or location, Display detailed information including ratings, location and cuisine, Zomato Foodie Index to show great areas to dine in a city.

Chennai, on the Bay of Bengal in eastern India, is the capital of the state of Tamil Nadu, India. The city is home to Fort St. George, built in 1644 and now a museum showcasing the city's roots as a British military garrison and East India Company trading outpost, when it was called Madras. It is one of the largest cultural, economic and educational centres of south India.The city together with the adjoining regions constitutes the Chennai Metropolitan Area, which is the 36th-largest urban area by population in the world Chennai is among the most-visited Indian cities by foreign tourists. It was ranked the 43rd-most visited city in the world for the year 2015. The Quality of Living Survey rated Chennai as the safest city in India. Chennai attracts 45 percent of health tourists visiting India, and 30 to 40 percent of domestic health tourists. As such, it is termed "India's health capital". Chennai has the fifth-largest urban economy of India.

Chennai was added to the UNESCO Creative Cities Network (UCCN) list for its rich musical tradition. Chennai is nicknamed "The Detroit of India", with more than one-third of India's automobile industry being based in the city. Home to the Tamil film industry, Chennai is also known as a major film production centre[2].

1.2. Problem

Chennai has the shortage of database when it comes to details about hotels. Although popular services like foursquare has a database of hotels across the world, it has the count of hotels less than 10. This project identifies a best API for hotels across India and the ways in which data is retrieved and analyzed.

2. Data Retrieval and Data Cleaning

[1] https://developers.zomato.com/api#headline1
[2] Wikipedia

Unlike other popular APIs like Foursquare, etc. Zomato has a limit of retrieving just 20 set of data at a time in free version. Hence, the data retrieved in a loop to receive a list of 100 hotels beyond which premium it becomes premium.


```
{'location': {'entity_type': 'subzone',
  'entity_id': 6008,
  'title': 'Park Town',
  'latitude': '13.0798960000',
  'longitude': '80.2751110000',
  'city_id': 7,
  'city_name': 'Chennai',
  'country_id': 1,
  'country_name': 'India'},
 'popularity': {'popularity': '4.33',
  'nightlife_index': '0.98',
  'nearby_res': ['66330',
   '67408',
   '66334',
   '71679',
   '66342',
   '66312',
   '68329',
   '18428769',
```


The following data has a limit of 20 data. Hence, the data is written inside a loop and appended. The JSON formatted server side stored data has been accessed and the following columns are returned as the dataframe:

```
Index(['restaurant.R.res_id', 'restaurant.all_reviews_count',
       'restaurant.average_cost_for_two', 'restaurant.cuisines',
       'restaurant.establishment', 'restaurant.events_url',
       'restaurant.has_online_delivery', 'restaurant.highlights',
       'restaurant.is_table_reservation_supported',
       'restaurant.location.address', 'restaurant.location.city',
       'restaurant.location.latitude', 'restaurant.location.locality',
       'restaurant.location.longitude', 'restaurant.location.zipcode',
       'restaurant.name', 'restaurant.offers', 'restaurant.opentable_support',
       'restaurant.phone_numbers', 'restaurant.price_range',
       'restaurant.store_type', 'restaurant.timings',
       'restaurant.user_rating.aggregate_rating',
       'restaurant.user_rating.rating_color',
       'restaurant.user_rating.rating_text', 'restaurant.user_rating.votes'],
      dtype='object')
```

3. Methodology

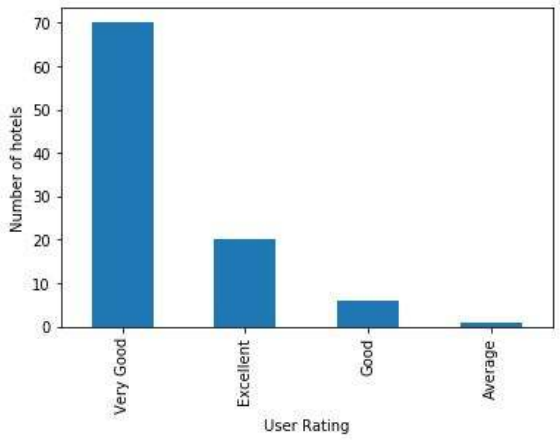
The feature set is identified to be user rating, latitude, longitude, average cost per two and the details of hotels which clusters the dataset on the basis of ratings. The best hotels have a strong relationship with the ratings. The ratings set is normalized and unnecessary columns were dropped in the process.

City	Latitude	Locality	Longitude	Price Range	Aggregate Rating	UserRating Color	UserRating Text	UserRating Votes
Chennai	13.000380	Alandur	80.200623	3	4.3	5BA829	Very Good	45755
Chennai	12.970815	Pallavaram	80.147182	2	4.5	3F7E00	Excellent	3700
Chennai	13.022305	Kotturpuram	80.242340	3	4.2	5BA829	Very Good	2731

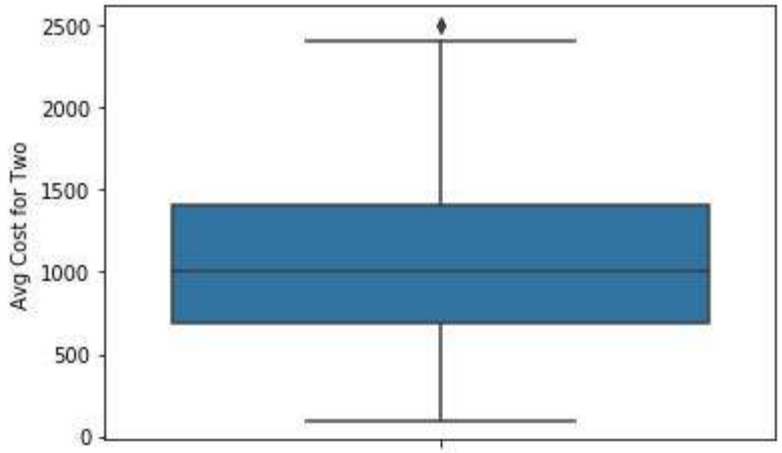
The description of the data is as follows:

	Restaurant ID	Reviews Count	Avg Cost for Two	Online Delivery Status	Is Table Reserved
count	1.000000e+02	100.00000	100.000000	100.000000	100.000000
mean	6.351519e+06	953.22000	989.500000	0.840000	0.520000
std	8.796904e+06	1111.55469	480.650601	0.368453	0.502117
min	6.505500e+04	252.00000	100.000000	0.000000	0.000000
25%	6.858775e+04	460.75000	687.500000	1.000000	0.000000
50%	7.167000e+04	625.50000	1000.000000	1.000000	1.000000
75%	1.844549e+07	1040.00000	1400.000000	1.000000	1.000000
max	1.895357e+07	7213.00000	2500.000000	1.000000	1.000000

The average cost for two people in best hotels across Chennai is 990. During, this Covid situation the online delivery status of the hotels were almost 84%. Table reservation is allowed only in 50% of the hotels.



The distribution has hotels that has very good rating as the highest, followed by excellent. There is an outlier when in average cost for two because the hotels with very high price cannot be compared to the hotels of similar range.

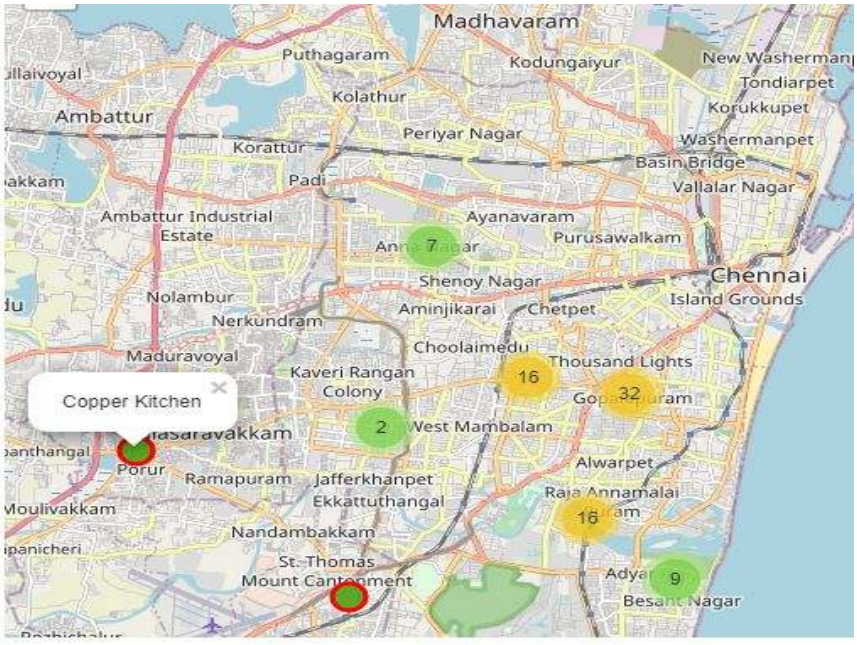


Hence, the outliers are removed and finally the analysis was carried out for 97 hotels.

As the data been extracted, retrieved and loaded after essential cleaning, the machine learning algorithms were run.

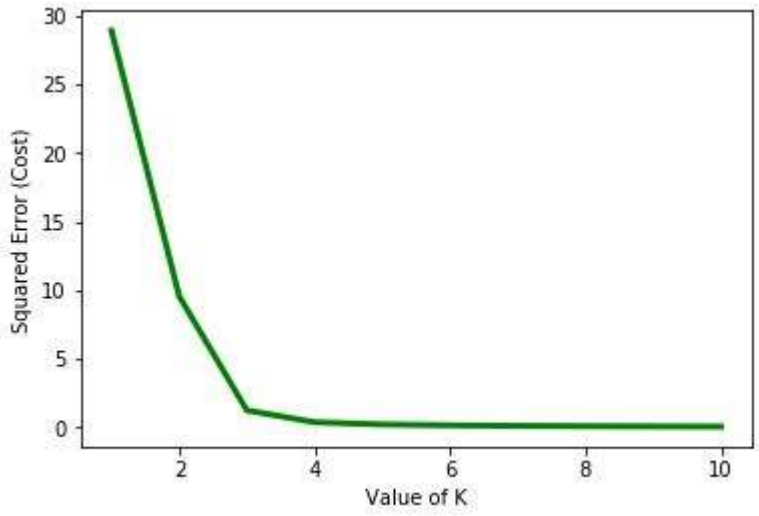
4. Results

The distribution of hotels across Chennai is mapped and the hotels were successfully placed.



The latitude and longitude data were separately collected and merged to the dataframe based on the area.

Then, in order to identify the right hotels with similar ratings, the unsupervised learning method is employed. K-Means Clustering is used to find the right clusters.

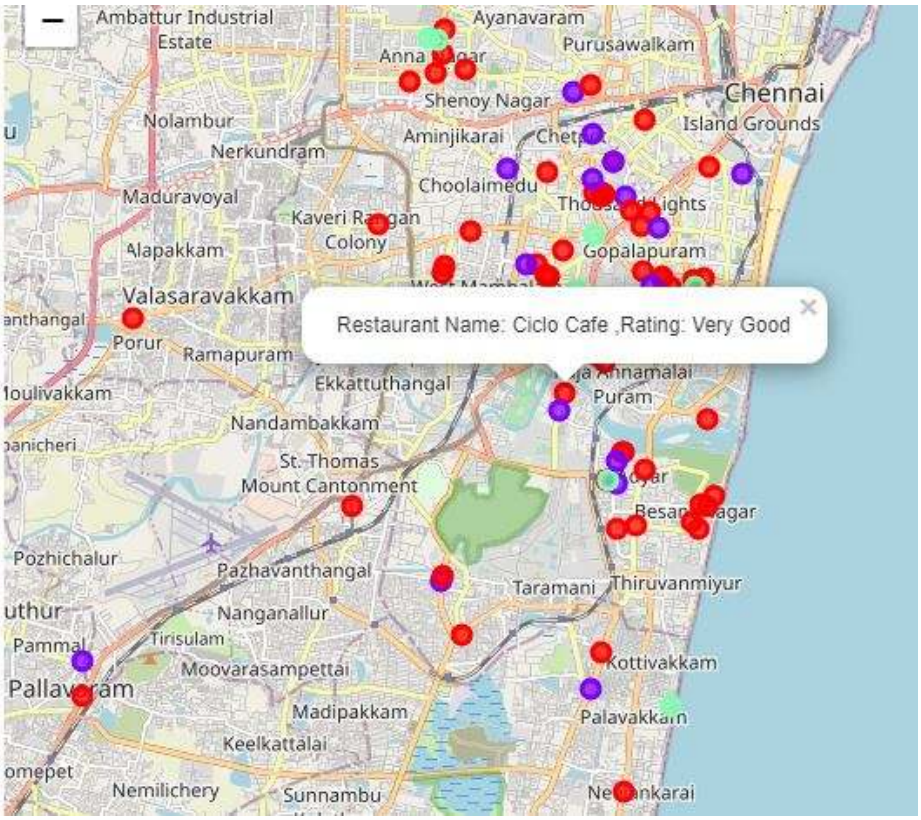


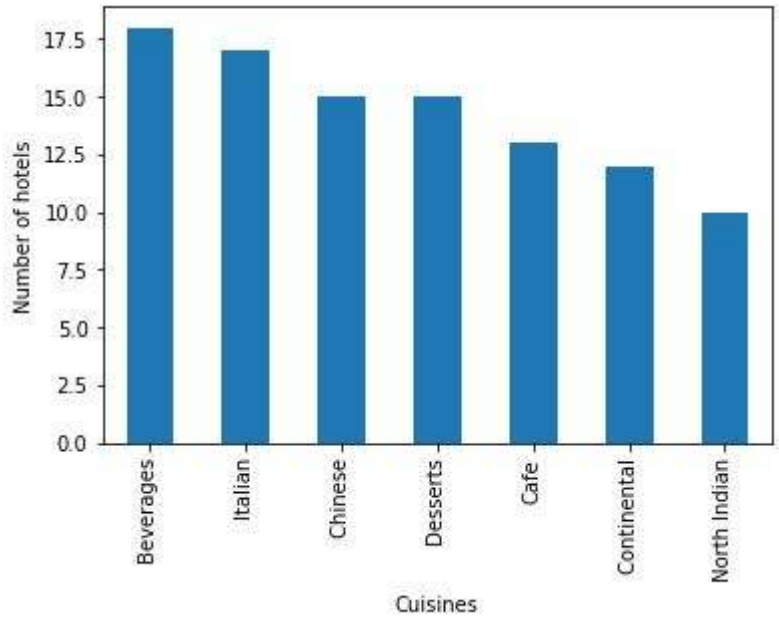
This specifies that at K = 3, the best clustering is possible and hence it is used as a clustering number.

```
df_data1['Classifier'].value_counts()
```

```
0    70
1    20
2     7
Name: Classifier, dtype: int64
```

The K-Means cluster classified the set of data based on the ratings, latitude and longitude and hence there exist three sets which maps excellent, very good and good set of hotels together.





4. Discussion

The red coloured hotels are the ones with Very good ratings and it is distributed across all parts of the city whereas the excellent rated hotels are more closer to the center of chennai. The

zomato API. The top cuisines across the above hotels are found using the above data by splitting the cuisines.

The cuisines and the region has a correlation which must be performed to get a regression values. However, the area is restricted to top 5 data based on the distribution.

5. Conclusion

In the following study, I had analyzed the data through a new API by directly invoking through calls. The top cuisines and top hotels of Chennai is found using machine learning algorithms. This helps data analyst to get the data of hotels across Chennai.