Feasibility Study for Opening a Restaurant in Pune, India

Anand Takale
13th July 2020

1. Introduction

1.1 Background

Restaurants, cafes, bars and eateries have always been a lucrative business worldwide. There has always been a demand for new and trendy places. Increase in migrants, cosmopolitan culture and increase in globe-trotting, has increased need to cater to a variety of cuisines. Many factors like location, type, cuisine, menu, ambience, pricing and services provided play an important part in determining the success of the business. The city of Pune, India has been rapidly expanding over the last few years and with the ever-increasing population the demand for restaurants is high.

1.2 Problem

Many factors like location, type, cuisine, menu, ambience, pricing and services provided play an important part in determining the success of restaurant business. This project aims at analyzing the current landscape of various restaurants in the city of Pune and attempts to determine various factors to maximize the returns. In particular this study attempts to determine the factors for opening a new restaurant: location, cuisine and features

- Location: Find an optimum location for opening a restaurant
- Predict the popularity: Based on existing restaurants and ratings, try to predict the rating of new restaurant based on price, cuisine and services provided
- Pricing: Find an optimal price point in terms of cost for persons per meal

1.3 Interest

New restaurant owners will benefit from this study as it will give them guidelines in starting their business. Restaurants also face the challenge of changing social trends, which are cyclic in nature . Apart from a few restaurants who have created a legacy in terms of cult following, the

majority of the restaurants need to revamp themselves over a period of time. This study will also assist existing restaurant owners by giving them insights to extend, modify and revamp their current offerings. Thirdly, this will also help allied businesses like food delivery services to determine popular areas and interesting restaurants to deploy optimal resources.

2. Data Acquisition and cleaning

2.1 Data sources

The geospatial coordinates of Pune city and various pincodes within the city are obtained using geopy. Using the geospatial coordinates, venues within 2 kms are fetched using Foursquare API (https://foursquare.com/). The following features were retrieved for each venue:

- Name: The name of the venue
- Type: Category of the venue (like restaurant, eatery, Cafe, Pub, Brewery etc)
- Location: The location of the venue including latitude, longitude and pincode

For the restaurants fetched from Foursquare API, details and reviews for restaurants in Pune using Zomato API (https://developers.zomato.com/api). The following features were retrieved for each restaurant:

- Name: The name of the restaurant
- Type: Category of the restaurant (like restaurant, eatery, Cafe, Pub, Brewery etc)
- Location: The location of the venue including latitude, longitude and pincode
- Cuisine: The type of cuisine the hotel serves
- User provided data: Votes, rating, photos, reviews
- Pricing: Average price for two people
- Features: Various features like table reservations, delivery, takeway etc

2.2 Data cleaning

- Initially data fetched from Foursquare API has some of the pincodes (zip postal codes) missing for few venues. Data imputation was based on prior knowledge as we had queried for individual pincodes, so missing data could be accurately populated
- Data obtained from Foursquare API was not comprehensive and did not include the local sentiments and ratings from local users. Zomato being quite actively used in Pune region has more users and ratings for each venue. Hence, data obtained from both sources viz. Foursquare and Zomato was combined into a single table using the hotel name.
- Duplicate records or overlapping records fetched were pruned first.

- All the data points with either the location of user ratings missing were pruned, to produce a usable data set of 950 unique restaurants.
- Furthermore the text data in highlights was processed to add column for different features and services provided by each individual entity

2.3 Feature selection

- Features were carefully selected taking any redundancy into consideration
- Additional features were created based on textual information from some of the fields. Namely: highlight which indicates the services provided by the restaurant and cuisine which lists the cuisines served by the restaurant

2.4 Links to fetched data

The collected data is available at:

Data fetched using Foursquare API:

https://github.com/atakale7/Coursera_Capstone/blob/master/Capstone/foursquare.csv

Data fetched using Zomato API:

https://github.com/atakale7/Coursera_Capstone/blob/master/Capstone/zomato.csv

2.5 Data Cleaning and Transformation:

- Few of the venues/restaurants were missing pincodes. Data imputation was done using prior knowledge as the query to fetch the restaurants contained the required pincode value
- Data fetched using the Zomato API has few missing values for cuisines. As they were very minute in number, these rows (data-points were dropped)
- “Cuisine” and “Highlights” were two fields which contained text data. The text data was processed to construct new columns/features and used for valuable inference
- Few of the restaurants also lacked any user rating. Such restaurants were also dropped for further analysis.
- Post this data transformation, the datasets fetched from both API (Foursquare and Zomato) were merged based on the hotel name. The resultant data used for processing contains 928 entries with 14 features (as latitude and longitude which were replicated in both datasets were discarded).

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 928 entries, 0 to 927
Data columns (total 15 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   name                  928 non-null    object
 1   lat                   928 non-null    float64
 2   lng                   928 non-null    float64
 3   pincode               928 non-null    object
 4   type                  928 non-null    object
 5   average_cost_for_two  928 non-null    int64
 6   price_range           928 non-null    int64
 7   aggregate_rating      928 non-null    float64
 8   votes                 928 non-null    int64
 9   cuisines              928 non-null    object
 10  highlights            928 non-null    object
 11  all_reviews_count     928 non-null    int64
 12  photo_count           928 non-null    int64
 13  has_table_booking     928 non-null    int64
 14  has_online_delivery   928 non-null    int64
dtypes: float64(3), int64(7), object(5)
memory usage: 116.0+ KB
```

- The resultant dataframe looks like below:

```
df.head()
```

	name	lat	lng	pincode	type	average_cost_for_two	price_range	aggregate_rating	votes	cuisines	highlights	all_reviews_count	p
0	Vohuman Cafe	18.532555	73.876834	411001	Breakfast	250	1	4.6	4603	Street Food	['Cash', 'Takeaway Available', 'Breakfast', 'L...	1996	
1	Gajalee	18.527478	73.878215	411001	Seafood	1000	3	4.2	647	Seafood, Mughlai, North Indian	['Dinner', 'Cash', 'Takeaway Available', 'Debi...	306	
2	La Bouchee d'Or	18.538931	73.876677	411001	Bakery	600	2	4.1	243	Cafe, Bakery	['Breakfast', 'Delivery', 'Credit Card', 'Lunc...	110	
3	La Bouchee d'Or	18.558395	73.803787	411007	Bakery	600	2	4.1	243	Cafe, Bakery	['Breakfast', 'Delivery', 'Credit Card', 'Lunc...	110	

3. Exploratory Data Analysis

As a first step we take a look at the basic statistical significance of all the features used for data analysis. From the basic statistical analysis we can see the following:

```
df.describe()
```

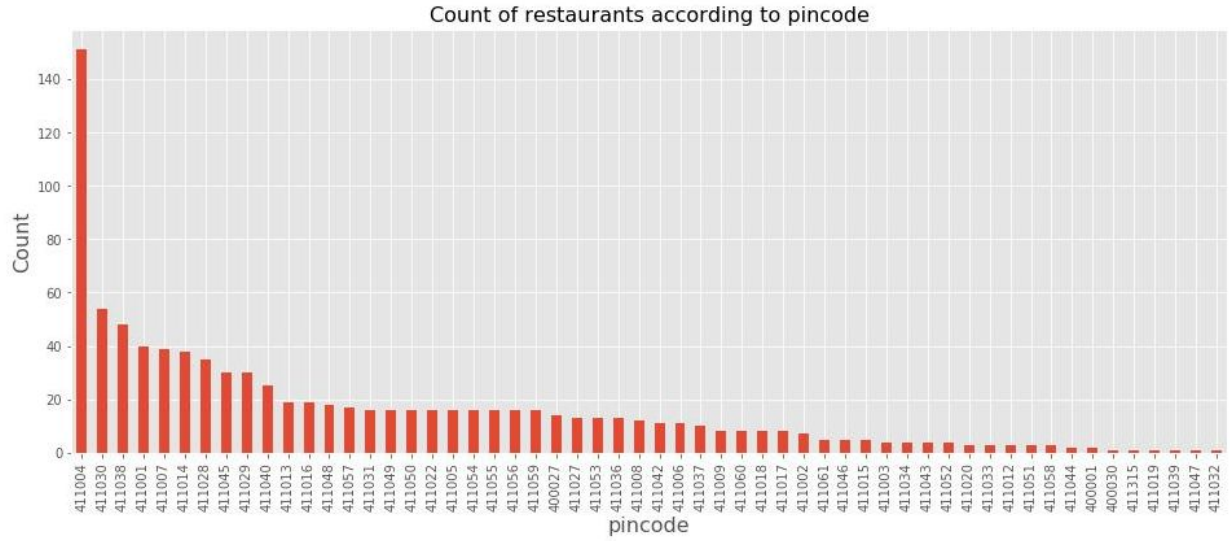
	lat	lng	average_cost_for_two	price_range	aggregate_rating	votes	all_reviews_count	photo_count	has_table_booking	has_o
count	928.000000	928.000000	928.000000	928.000000	928.000000	928.000000	928.000000	928.000000	928.000000	
mean	18.523253	73.851635	610.021552	1.743534	3.908297	1622.728448	362.492457	435.024784	0.082974	
std	0.028538	0.037079	420.742908	0.763912	0.616884	2172.220883	518.994746	981.523962	0.275992	
min	18.440685	73.766671	100.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	18.512350	73.833940	350.000000	1.000000	3.700000	262.000000	31.000000	46.000000	0.000000	
50%	18.518451	73.842713	500.000000	2.000000	4.000000	774.000000	198.000000	95.000000	0.000000	
75%	18.524296	73.872360	600.000000	2.000000	4.200000	2434.750000	356.750000	367.000000	0.000000	
max	18.671948	73.937133	2100.000000	4.000000	4.700000	30737.000000	2621.000000	7307.000000	1.000000	

- Pin code with maximum number of restaurants: 411004
- Most common type: Cafe
- Average Cost for 2 persons: Rs. 610
- Average Rating: 3.9

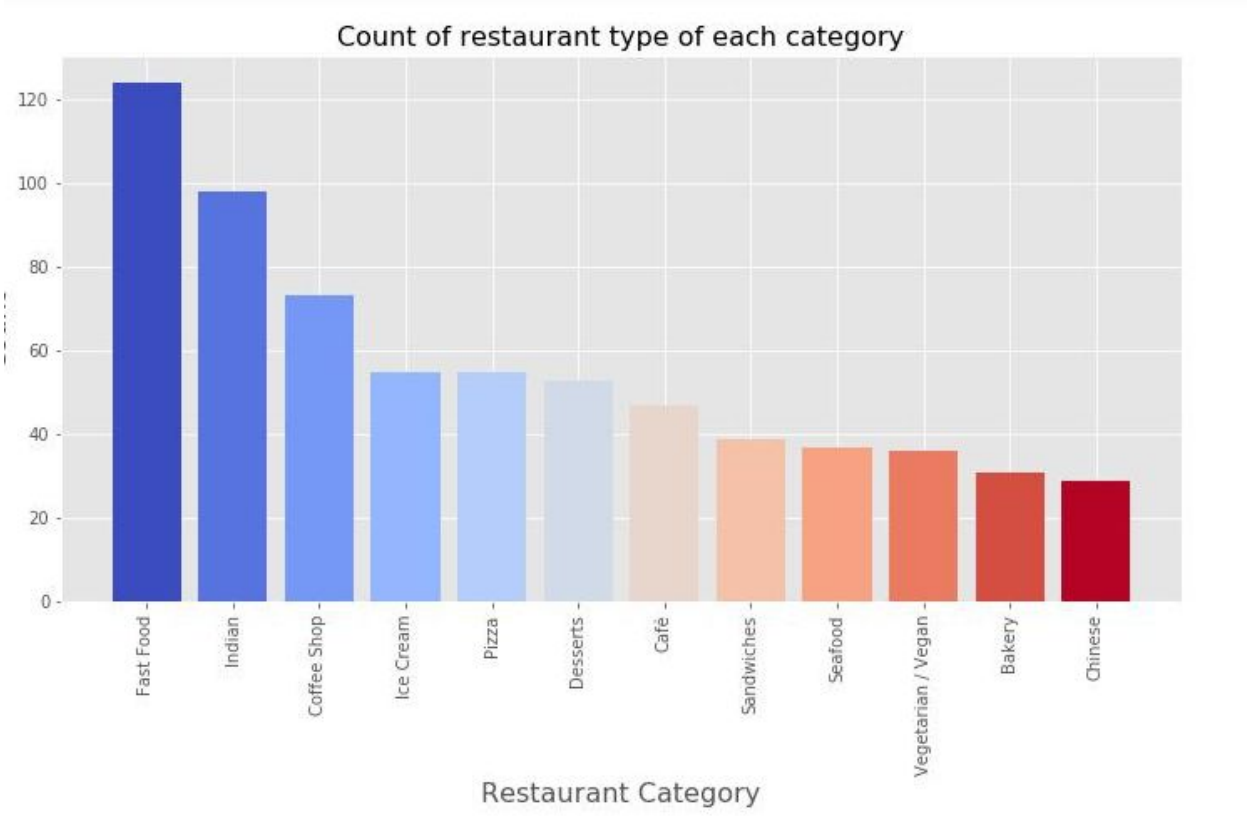
We then take a look at the distribution of restaurants based on various factors. As we are interested in determining optimal location, pricing, choice of cuisine and features/services to offer, we must first analyze the current distribution of restaurants/eateries. We look at the current restaurants using countplot, bar graph and histogram.

3.1 Distribution of restaurants based on pincodes

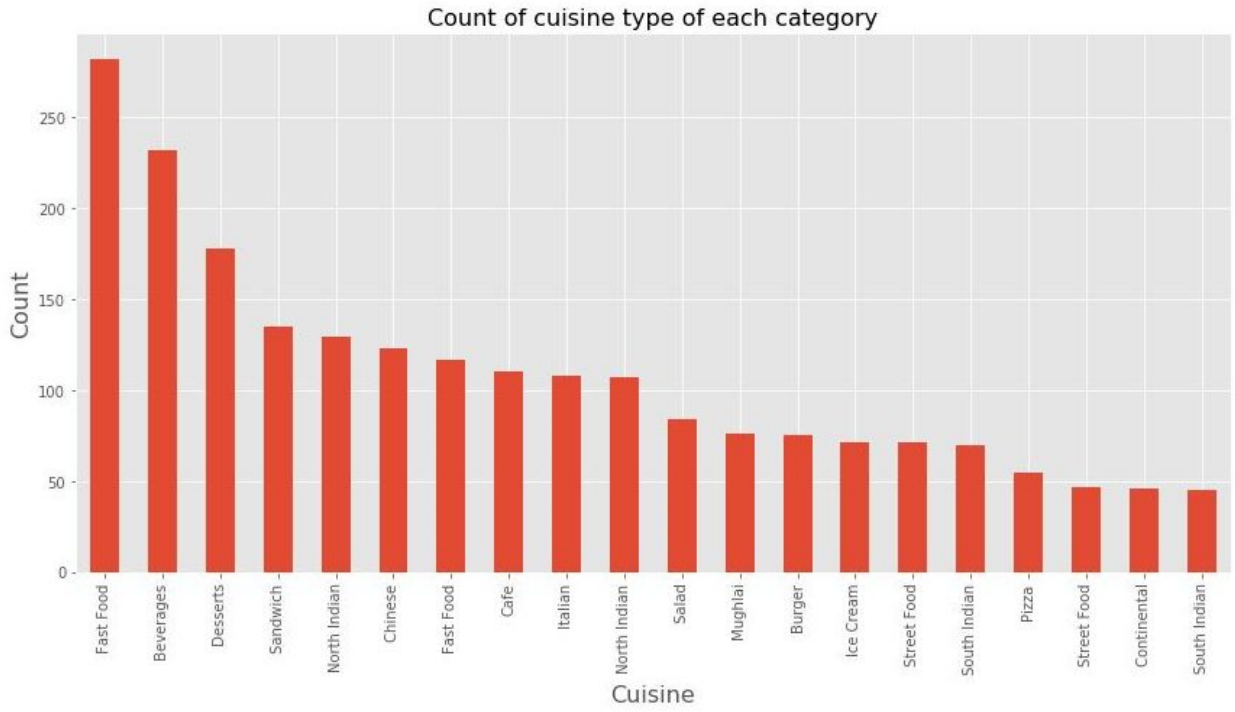
We can easily verify which locality based on pin-code is more popular for eating out and hence they are the areas to consider based on current scenario



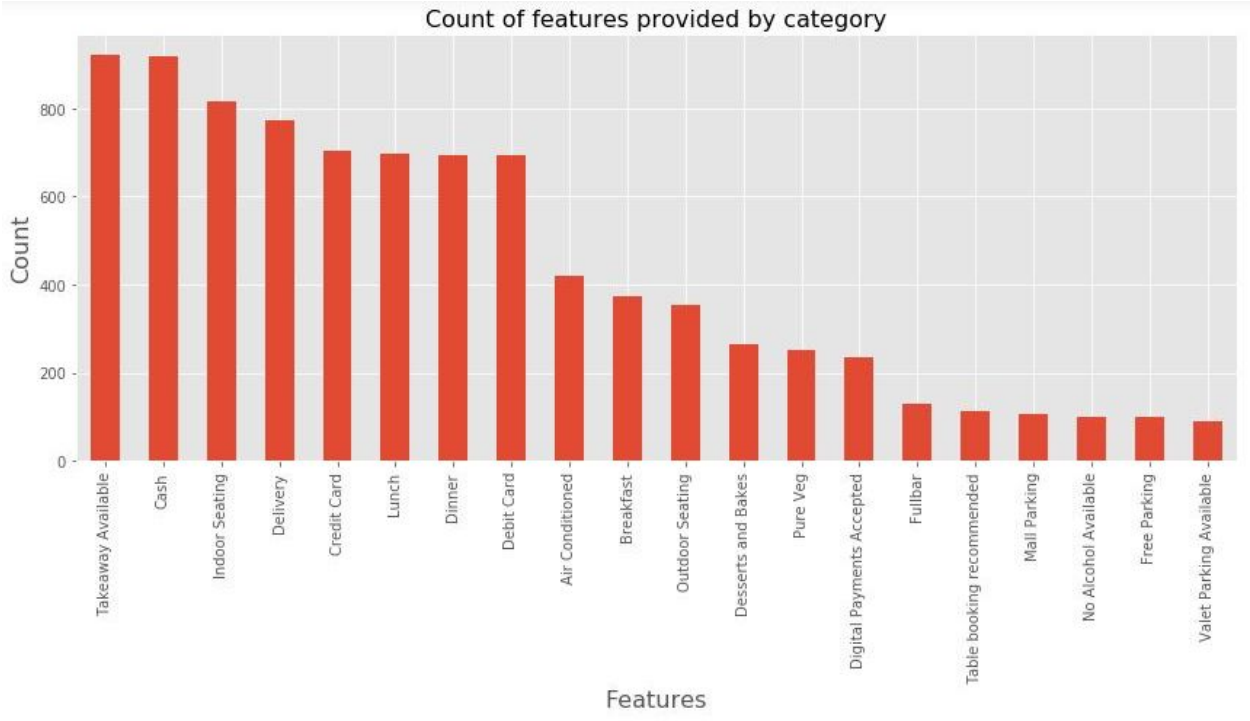
3.2 Distribution of restaurants based on type of restaurant:



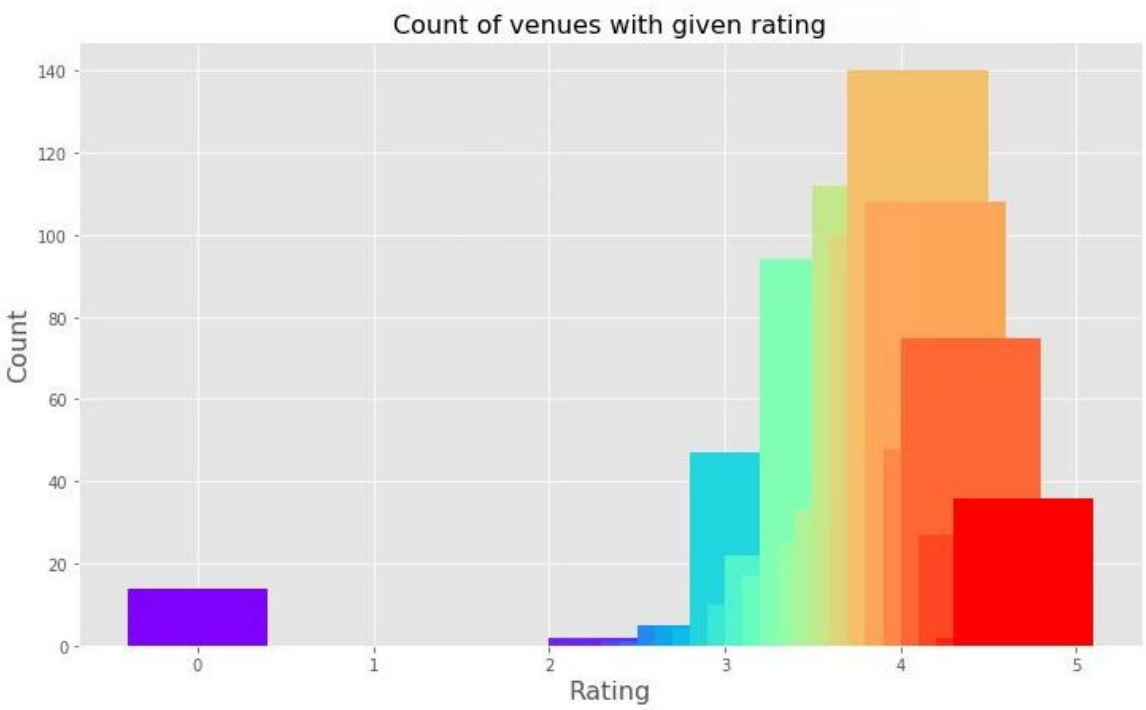
3.2 Distribution of restaurants based on cuisines served:



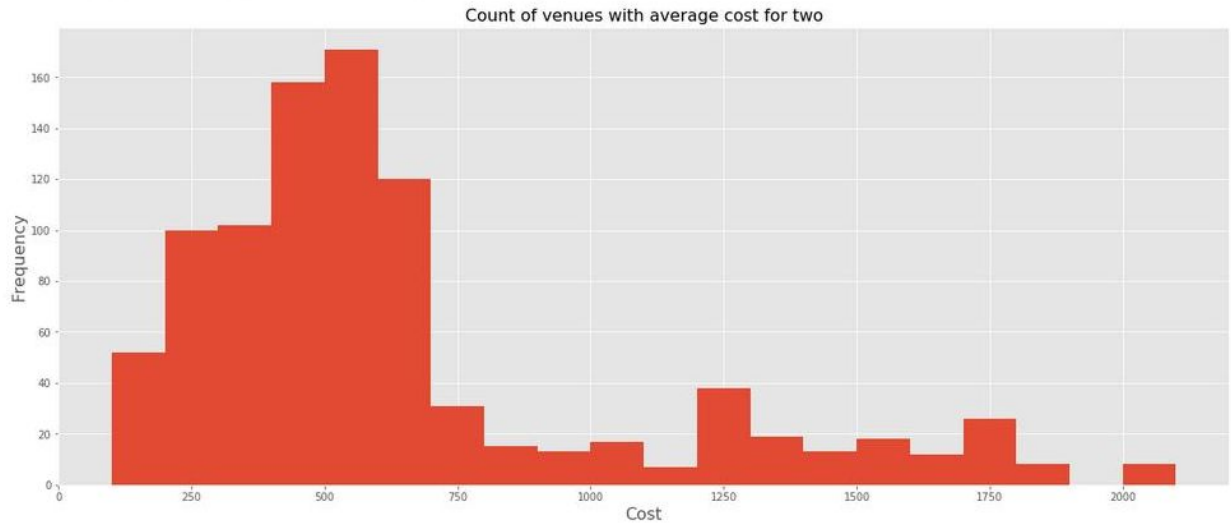
3.3 Distribution of restaurants based on features/services provided



3.4 Distribution of restaurants based on ratings

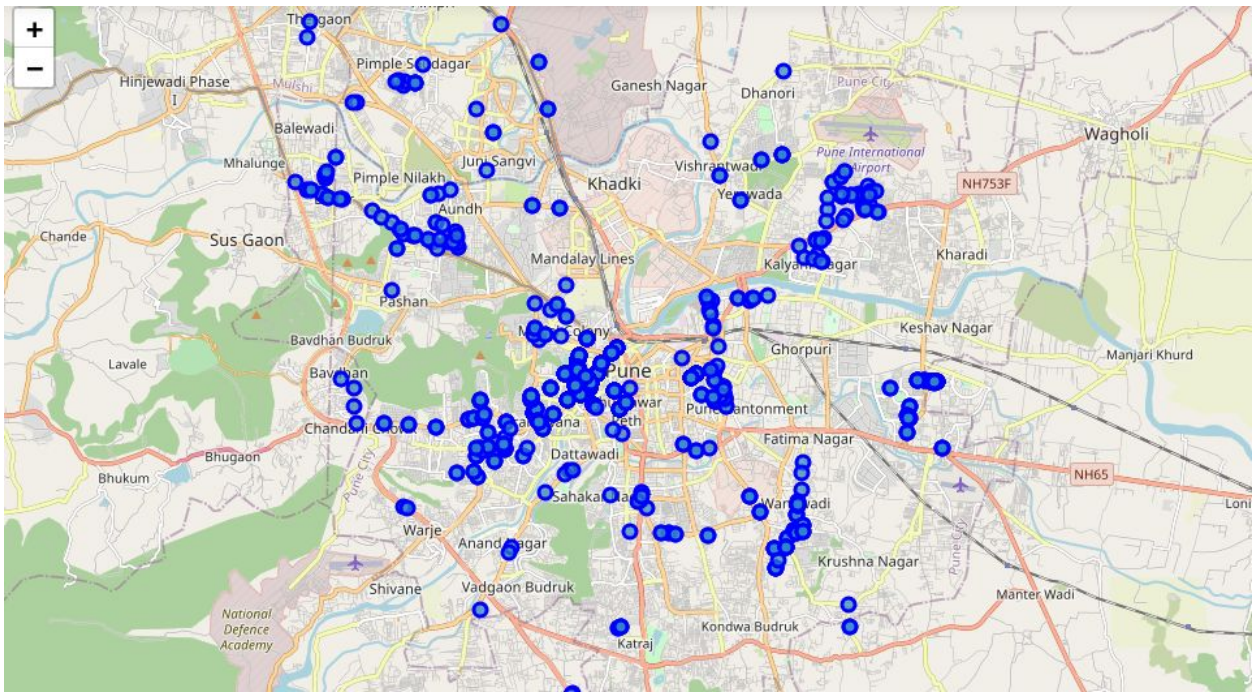


3.5 Distribution of restaurants based on average cost for two

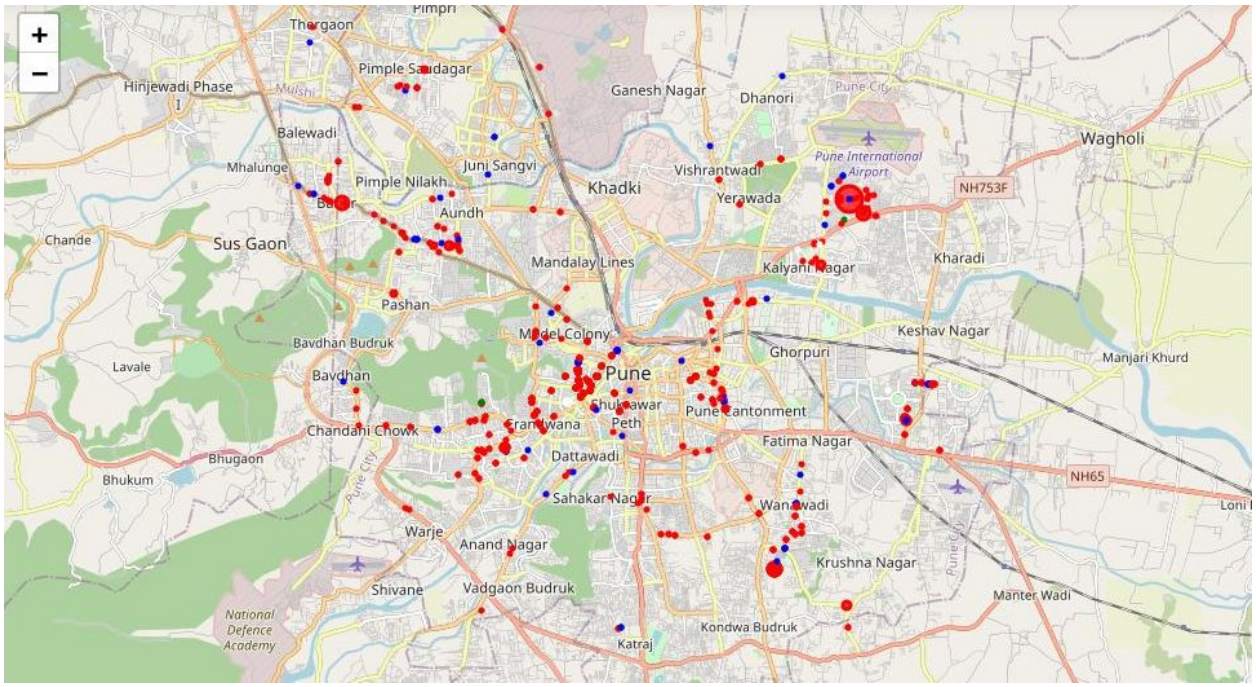


3.6 Geographical Distribution of restaurants:

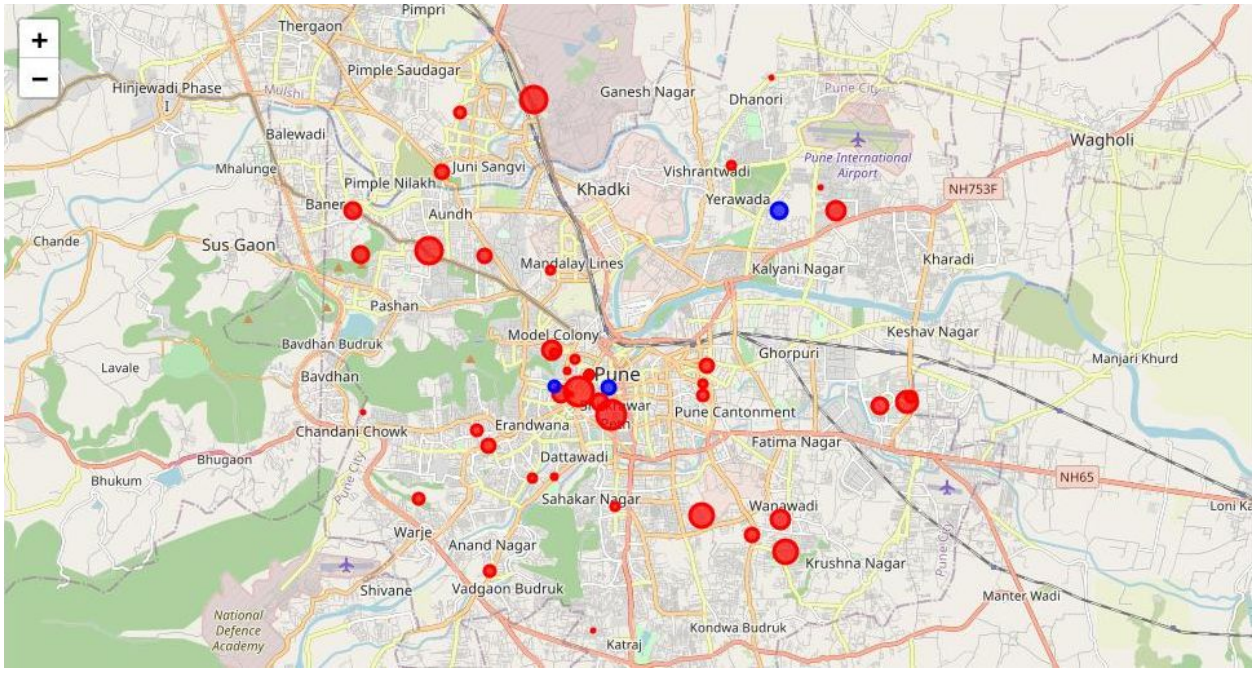
We also plot the chosen restaurants from the dataset on a map to showcase better visualization of spread of restaurants across the city:



1. Next, we plot the restaurants color-coded with ratings less than and above average. The red dots represent the restaurants which have higher ratings than average (>3.9) and the blue dots represent the restaurants which have lower ratings than average. The size of the circles also represent the popularity of the restaurants based on the number of votes received as that represents customers who have visited more times and have rated the restaurants online.



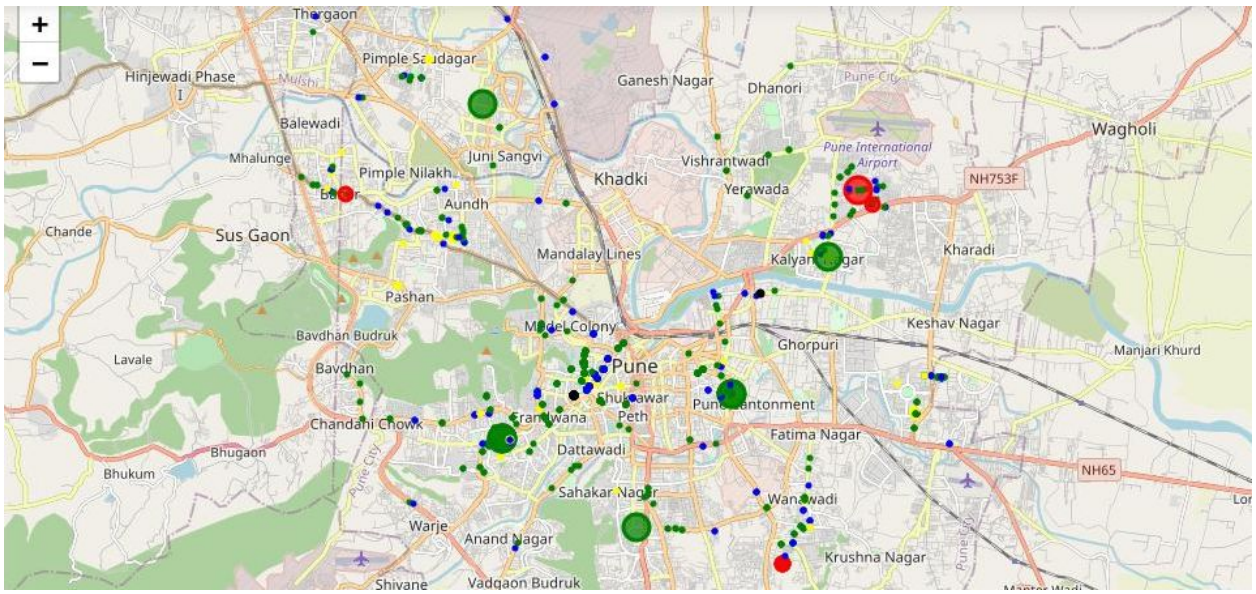
The following map represents the popularity of an area or a locality based on pincode. The more number of restaurants with more number of visits results in more number of reviews and ratings. There are only three localities which have below average rating, while rest of the localities have restaurants with more than average ratings. Choosing the localities with poor rating is one way to approach a business problem. However, we are choosing the aggressive method of finding a locality which is more popular, hence more number of customers and more need to serve varied as well as popular cuisines



4. Predictive Modeling

4.1 Clustering (Finding the correct characteristics of the restaurant)

First we use clustering which is an unsupervised learning algorithm to group the restaurants into sub-groups with similar characteristics. As clustering works only on numerical values, we convert the values in text based features such as highlights and cuisines into numerical/categorical values and then use KMeans to divide the current set of the restaurants into 5 distinct groups. The figure below showcases the various clusters:



5 categories of restaurants can be seen clearly:

Cluster label 0: (Blue) Lowly rated restaurants

Cluster label 1: (Yellow) Moderately rates but expensive restaurants

Cluster label 2: (Green) Moderately rated and affordable

Cluster label 3: (Red) Moderately rated, affordable and frequently visited and rated restaurants

Cluster label 4: (Black) Highly rated but expensive restaurants

As we are interested in opening a restaurant which has significant turnover of customers, it needs to have similar characteristics of popular restaurants. Also keeping initial investment into consideration, we need to look out for affordable restaurants. Keeping everything into consideration, common characteristics of restaurants in cluster groups 2 and 3 should be considered.

4.2 Predicting the popularity based on various features

In the second part, we use inferential statistics method called Linear Regression to predict the popularity using aggregate rating based on votes and ratings. We combine numerical features as input features and the aggregate rating is used as a dependent variable or the target feature. The dataset is first split into two subsets: training subset (consisting of 80% of the data points) and a testing subset (consisting of 20% of data points). The model is trained using multiple linear regression. Once the model is trained it is then used to predict the aggregate rating of the restaurants in the testing sub-set. Variance score of 0.88 is obtained on the testing sub-set.

```
y_hat = regr.predict(X_test)
x = np.asanyarray(X_test)
y = np.asanyarray(y_test)
print("Residual sum of squares: %.2f"
      % np.mean((y_hat - y) ** 2))

# Explained variance score: 1 is perfect prediction
print('Variance score: %.2f' % regr.score(x, y))
```

```
Residual sum of squares: 0.06
Variance score: 0.88
```

5. Results and Discussion

As seen above in sections 3 and 4, we first basically did an exploratory data analysis to zero in on localities which housed popular restaurants/eateries. We then used clustering to group together the restaurants with similar characteristics. As a result of both the steps, there are 3 localities which are suitable for opening a restaurant: 411004, 411045 and 411014. The results are in line with local knowledge of the city which claims that these areas are either localities with affluent households or corporate offices which indirectly results in increased consumption of higher prices goods and services. The study also showcased that the majority of the popular eating joints are either fast food joints or a Cafe. This also matches the fact that the majority of the population of Pune is young.

For the second part of the study, a multiple linear regression model was constructed using numerical data fields as input. The built model gave a fairly good accuracy with a variance score of 0.88. It can now be used as an effective tool while drawing up a business plan to take decisions on various features/services to be provided.

6. Conclusion

The study analyzes the distribution of restaurants across Pune city on basis of various factors including their geographical coordinates and popularity. It identified three core localities which are more suitable for opening a thriving restaurant business. By using predictive modelling methodology, the study was able to cluster restaurants with similar characteristics. With the help of exploratory data analysis and predictive modelling, the study was able to provide recommendations to new and existing business owners.

7. Future Direction

The text in the actual reviews can be processed further using sentimental analysis to draw up more features and refine current findings. Also a larger dataset and a dataset across the cities would be more beneficial to refine and better the current learning model