Capstone Project

Building a Recommender for a Restaurant in India

- Dipro Mondal

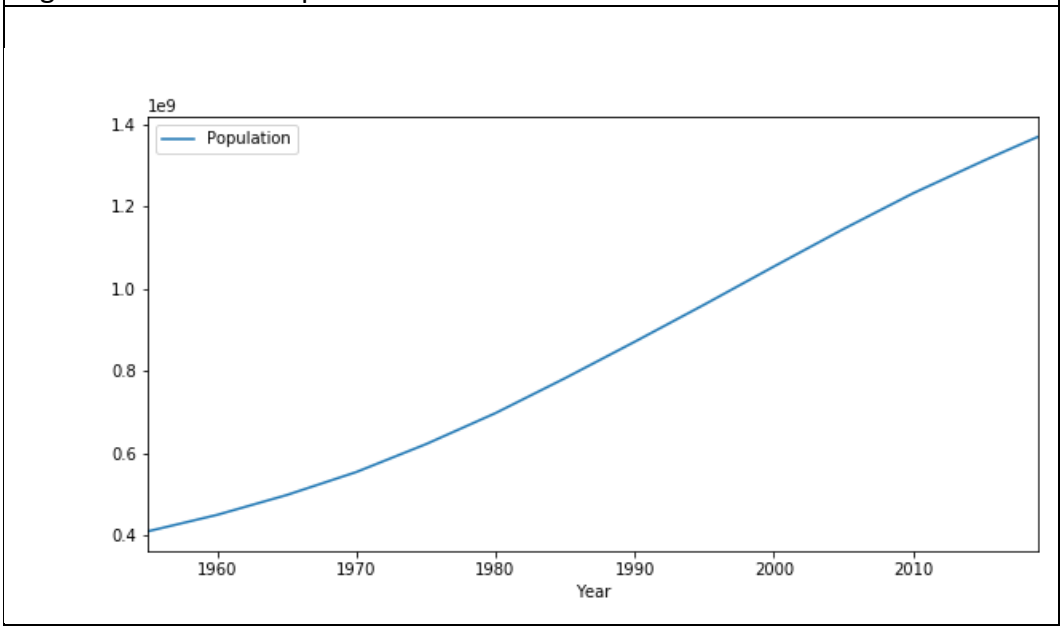
Table of Contents

1. Introduction

1.1 Background

India is a population of 1.4 billion and increasing (source: https://www.worldometers.info/world-population/india-population/)

Figure 1: Historical Population of India



Due to this increasing number of population there is a huge demand of food and certainly restaurants.

1.2 Data Science Question

The increasing demand of restaurants in India has been met with several types of restaurants established, encompassing tastes from various corners of the world- including and not limited to French, Italian, Mexican, Chinese and so on. However the question is if someone wants to open a new Restaurant in India **where (City) it should be?** Also another question is **what type of restaurant it should be?**

2. Methodology

In order to answer these questions I have acquired and used statistical data of population of the cities of India and compared it with the number of restaurants currently available in each city. Then I used the ratio of population vs restaurants available to recommend the city which has the highest ratio. To recommend the type of restaurant, I used k-means clustering to cluster the different cities based

on population and restaurants, and used the most frequent type of restaurant to recommend the type of restaurant for the recommended city.

2.1 Data Acquisition

For the analysis I required –

i) Data of population of India (historical) – scraped from https://www.worldometers.info/world-population/india-population/

ii) Data of population of India by cities – scraped from http://worldpopulationreview.com/countries/india-population/cities/

iii) Location data of all Indian cities – scraped from http://worldpopulationreview.com/countries/india-population/cities/

iv) List of all restaurants in different cities – obtained from FourSquare

2.2 Data Preparation

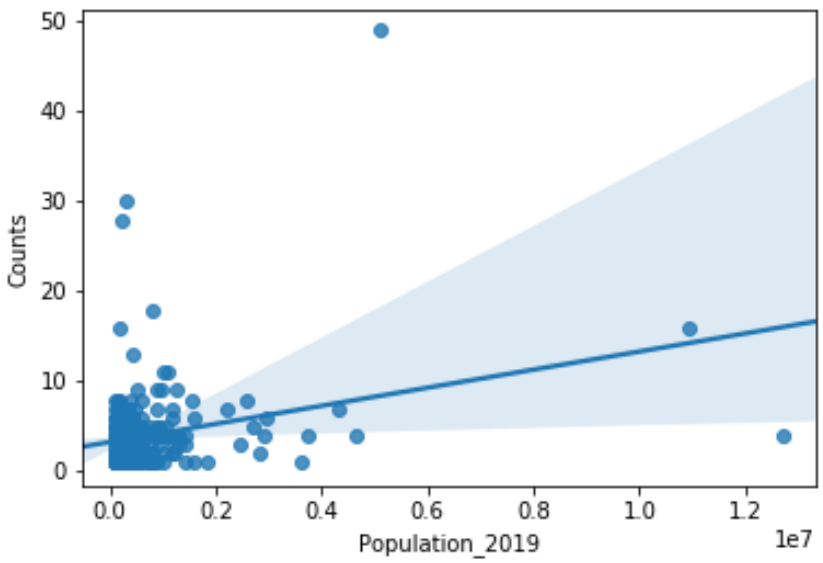
For scraping the data I have used BeautifulSoup. The population data (both the historical population data as well as the population of different cities) were readily available and had no missing values. For the location data of the cities I have used the google hyperlinks of the location from the site mentioned and used it as a starting parameter for the FourSquare Search. The Foursquare data containing all events of different cities contained some missing values. Some cities were registered as having no or '0' restaurants. Since there was only 3 cities in that category, I have neglected these data points for recommendation. These formed the final dataframe which I have used for my analysis.

3. Exploratory Data Analysis

Using the sources mentioned I formed a dataframe containing the cities of India, their population as of 2019 and the number of restaurants each city has. Now the question is whether there is any relation between the population and the number of restaurants.

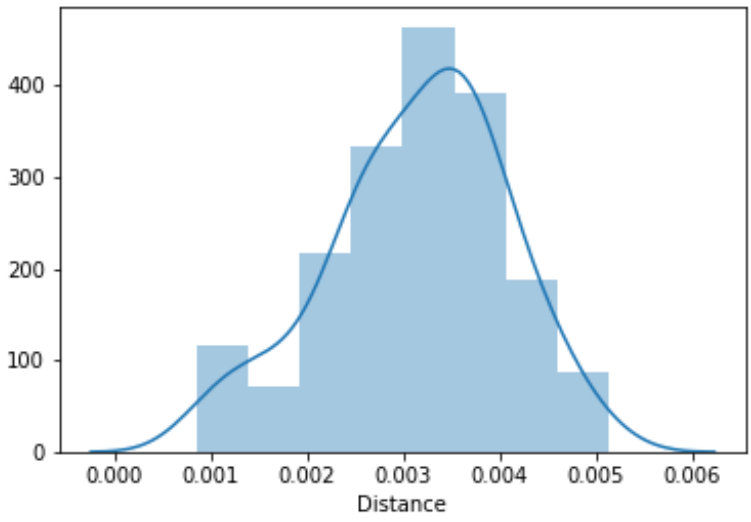
Figure 2 shows a linear regression plot of population of different cities with respect to the counts of the restaurants currently available which suggests there exist some linear correlation (though weak) between the population and number of restaurants.

Figure 2: Linear regression plot of population and number of restaurants (counts)



After calculating the ratio of the population with respect to the number of restaurants I have selected the city with the highest value to recommend building a restaurant there. Next I wanted to check how far from the city center the restaurant should be. The histogram plot in figure 3 shows that most restaurants are built with a distance of around 0.004 degrees from city center.

Figure 3: Histogram of the frequency of restaurants with distance from the city center



Now the question is what type of restaurant it should be. To recommend the type, I have clustered the cities using k-Means based on the restaurants and recommended the most frequent type of restaurant of the cluster containing the recommended city.

4. Results

My Data Analysis shows that the recommended city is Hyderabad and the type of restaurant is Fast Food. It should be built within a radius of 0.0035 – 0.0040 degrees from the city center. However this should be noted that this recommender system can only recommend based on previous statistical data on most frequent restaurants in India with respect to different cities.

5. Discussion

Though however the city recommended is Hyderabad, the type of restaurant changes with change of cluster type. Hence reproducing the recommendation of same type of restaurant multiple times is an issue.

Also the data obtained from FourSquare seems incomplete with respect to the number of events. I have used an iterative query search with foursquare with the location coordinates of each city. It's really a bit shocking to see several of the major cities in India having very few restaurants where there are positively much more restaurants. So more update from the Foursquare API is required for which I have automated my Data Analysis Script. If there are any updates from the API which thereby changes the dataset it will automatically be reflected in the results.

6. Conclusion

I have made an intelligent recommender system for recommending a restaurant in India with the knowledge of current restaurants and population of different cities in India. With availability of more data or more updated data from foursquare this recommender system can prove very successful.