Analysis of Indian demographics and suggestion of a locality with maximum number of Indian Restaurants near Bangalore

Geetha Prakash

July 13, 2019

This capstone project is a part of the Applied Data Science Capstone Course and has two parts.

PART-I :

In part one, a study of the demographics of India related to the total area, total population, male , female population, literacy rates and women entrepreneurs is done. Conclusions related to women entrepreneurs and the women literacy rates of the different states of India will be drawn. This analysis will give us information about states where women entrepreneurs can set up businesses. A study is done on the male literacy rate and the crimes on women and an analysis will be done to verify if the male literacy rate has an influence on the crimes on women. This will result in recommending a safe state for women with least number of crimes.

The data is obtained from the census data made available in the public domain. Different data sets from these sources will be scraped, cleaned, merged based on the requirements and necessary analysis completed

PART-II : Since there is a mandatory requirement to use Foursquare or a similar application and the demographic study of India does not really need such an application, I have decided to include a separate section in my project to use Foursquare to acquire data related to the eateries in and around Bangalore and a recommendation of the area with a good number of Indian Restaurants will be made.

Data from the wards of Bangalore and their latitude and longitude will be acquired using Foursquare and the suggestions will be made.

PART-1-DEMOGRAPHICS OF INDIA

1 Introduction

1.1 Background

India is the second most populated country in the world. India has 36 States and Union Territories combined.The study of demography is important as it allows us to study the nature in which our population changes over time, and this is important as it allows us to understand how changes to the population can influence the growth of the country. India has several states in which rural area dominates the urban area and the male population is many times larger than the female population. Literacy rates also vary from state to state and from rural to urban areas.Of late, entrepreneurship has been on the rise and there have been several women entrepreneurs who have made it to the forefront.

1.2 Problem

In this study we try to analyse some of the statistics such as male and female literacy rates, women entrepreneurs, crime rates and what factors may influence these numbers. Some analysis is done to predict states with higher women entrepreneurship numbers and also the influence of literacy rates on the number of crimes committed in a state.

2 Methodology

The execution methodology involves data collection, data cleaning, data sorting , feature selection and execution of machine learning techniques.

2.1 Data sources

To complete the analysis for the Indian demographics, data is obtained from different public domains such censusindia, districts and states of India, local government data etc. Data for some of the states are under dispute and so these states were not considered during analysis. Also some data pertaining to the recent states that were formed was not available. In some cases such data was retrieved from their individual portals and then appended to the main dataset.

2.2 Data Cleaning

Data from Wikipedia sources were extracted using BeautifulSoup and then appended with other data from the census sources. Data downloaded or scraped from multiple sources were combined into a single table. In some cases, columns not relevant to the analysis were dropped. Data sets combined from different sources had several problems. The numerics

were not in the same format and the data in all sets did not appear in the same order. Data were of different magnitudes and it was difficult to compare them without normalization. Standard normalization of all columns led to columns becoming zeros or NaN Format. In such cases, data type mismatch was addressed and columns were independently normalized.

3 Exploratory data analysis

Data analysis was related to women entrepreneurship and the factors which influence it. Population of women, Women literacy rate and the number of women entrepreneurs formed one data set where a comparison was made between this data of different states. Analysis was done on the influence of literacy rate on entrepreneurship. Histograms were plotted. The state with the highest number of women entrepreneurs was identified and a comparison was made with its literacy rate and that of the other states. Multivariate regression analysis was used to predict the possible number of women entrepreneurs when the literacy rate and the female population is known. With the output from the regression analysis, the name of the state which had the statistics closest to this result was extracted and was verified for correctness.
Similarly a study on the male population , male literacy rate and number of crimes in a particular state was studied. Number of crimes were predicted with male and female literacy rates considered as the factors influencing these numbers. Regression analysis was able to predict the number of crimes and the state very close to this number was predicted as output.

3.1 Normalization Methodoogy and interpretation of the graphs

Normalization was done for each column of the relevant data.
Interpretation of the bar graph: In the histogram depicting the comparison of women population, women literate population and the number of women entrepreneurs, all the similar coloured bars of the different states should be compared with one another. For example, in Figure 3 the cyan coloured bar represents women entrepreneurs. Tamilnadu has the maximum number and is therefore represented as 100 percent. For Kerala it is 84 percent and it implies that it is 16 percent less than that of Tamilnadu. The comparison is between the same parameter of different states.
Prediction is made as to which state would be the most cohesive state for a women entrepreneur . This is done by using Multivariate regression with the influencing parameters being the women population and the number of women literates. Prediction leads to the result that Maharashtra is the most conducive place for women entrepreneurs , though Tamilnadu has the maximum number of women entrepreneurs. Figure 6 shows the prediction analysis results.
Analysis is also done to check which state has the maximum number of crimes and prediction is made using Multivariate regression analysis with male and female literacy rates

being the factors of influence. With these it is predicted that Madhya Pradesh has the least predicted number of crimes based on the male and female literacy rates, though this is not directly evident from the histogram.

4 Results

[45]:

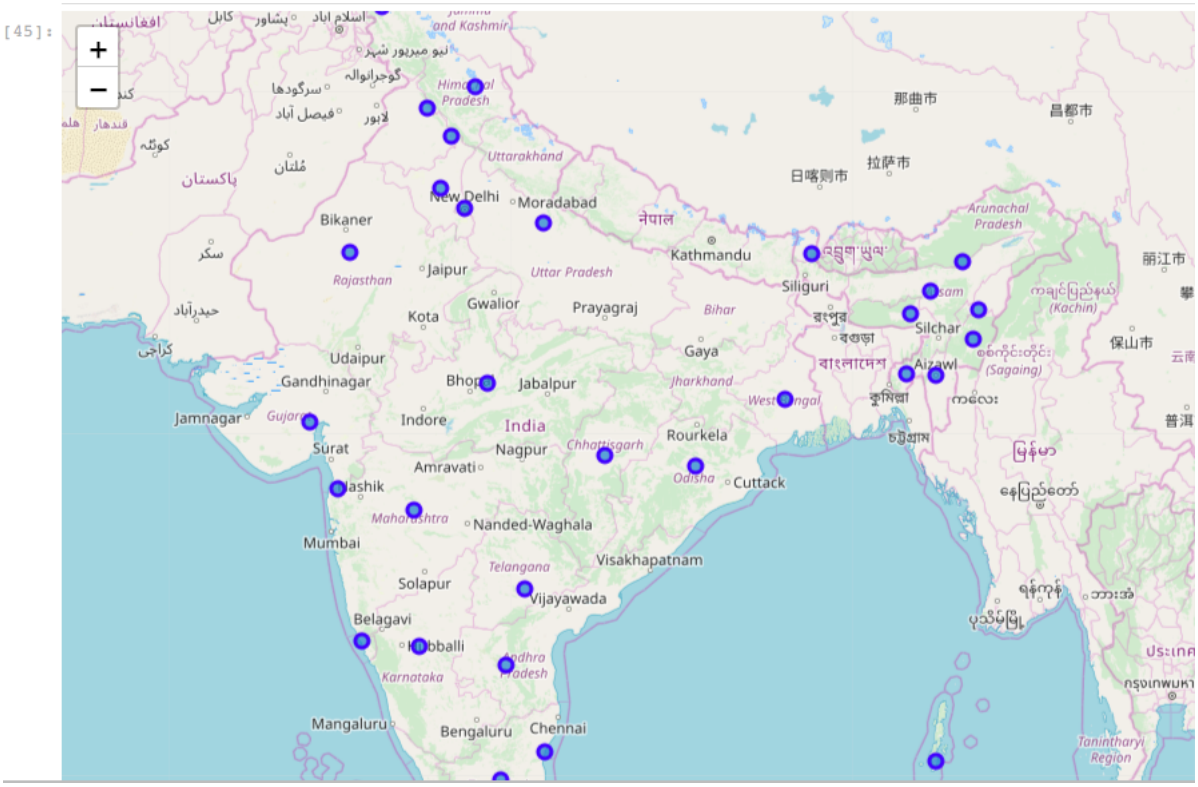

Figure 1: Different states on the Map of India represented by one important city-generated using Folium

[12]:

	States	total_population	Male_population	Female_Population	Women_Entrepreneurship	Female_Literates	Sex_Ratio
29	Tamil Nadu	72147030	36137975	36009055	1087609	24098521	996
17	Kerala	33406061	16027412	17378649	913917	14478339	1064
1	Andhra Pradesh	49386799	24738068	24648731	849912	22678728	996
34	West Bengal	91276115	46809027	44467088	831337	28106397	953
19	Maharashtra	112374333	58243056	54131277	664300	36218184	929

Figure 2: Sorted dataframe with relevant statistics

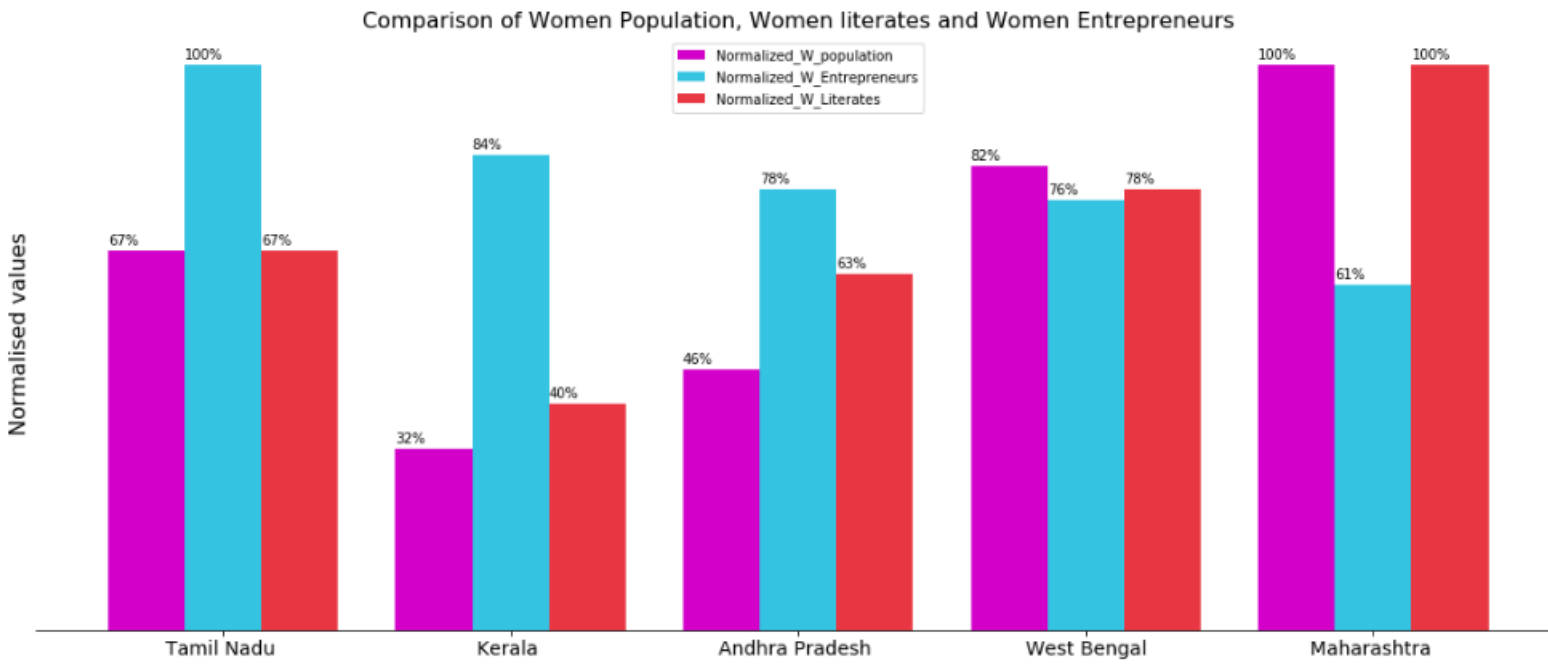

Figure 3: Comparison of Women literates, Women Entrepreneurs and Total women Population using Bar Graph

```
Intercept:
 1.195379976402363
Coefficients:
 [ 0.65894291 -1.19012735]
Normalized value of Predicted Women_Entrepreneurs:
 [0.66419553]
Predicted value of women entrepreneurs  [722385.03694365]
                            OLS Regression Results
==============================================================================
Dep. Variable:     Normalized_W_Entrepreneurs   R-squared:                       0.497
Model:                                    OLS   Adj. R-squared:                 -0.006
Method:                         Least Squares   F-statistic:                    0.9889
Date:                        Sat, 13 Jul 2019   Prob (F-statistic):              0.503
Time:                                09:06:19   Log-Likelihood:                 4.9722
No. Observations:                           5   AIC:                            -3.944
Df Residuals:                               2   BIC:                            -5.116
Df Model:                                   2
Covariance Type:                    nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                          1.1954      0.294      4.066      0.056      -0.070       2.460
Normalized_W_population        0.6589      0.989      0.666      0.574      -3.598       4.916
Normalized_W_Literates        -1.1901      1.230     -0.968      0.435      -6.481       4.101
==============================================================================
Omnibus:                          nan   Durbin-Watson:                   2.093
Prob(Omnibus):                    nan   Jarque-Bera (JB):                0.746
Skew:                           0.883   Prob(JB):                        0.689
Kurtosis:                       2.320   Cond. No.                         34.9
==============================================================================
```

Figure 4: Prediction of Women entrepreneurs based on the influence of women literates and total women population

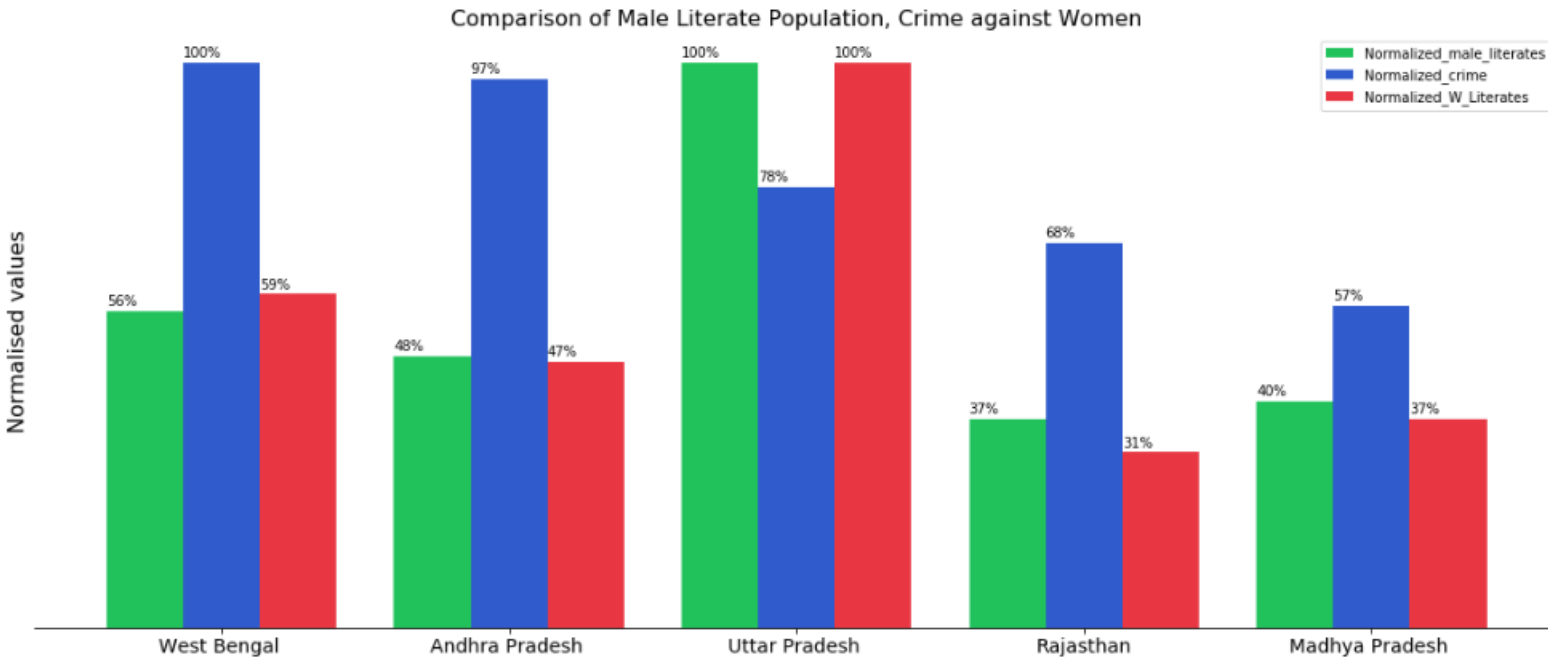

Figure 5: Comparison of number of crimes based on the number of male literates, women literates and total women population

```
Intercept:
 0.958857610791281
Coefficients:
 [ 4.69880564 -4.86441833]
Normalized value of Crime:
 [0.79324492]
Number of crimes predicted [23109.60427701]
                            OLS Regression Results
==============================================================================
Dep. Variable:       Normalized_crime   R-squared:                       0.588
Model:                            OLS   Adj. R-squared:                  0.176
Method:                 Least Squares   F-statistic:                     1.426
Date:                Sat, 13 Jul 2019   Prob (F-statistic):              0.412
Time:                        11:54:00   Log-Likelihood:                4.1214
No. Observations:                   5   AIC:                            -2.243
Df Residuals:                       2   BIC:                            -3.415
Df Model:                           2
Covariance Type:            nonrobust
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                          0.9589      0.247      3.882      0.060      -0.104       2.022
Normalized_W_Literates         4.6988      2.879      1.632      0.244      -7.687      17.084
Normalized_male_literates     -4.8644      3.085     -1.577      0.256     -18.137       8.408
==============================================================================
Omnibus:                        nan   Durbin-Watson:                   1.958
Prob(Omnibus):                  nan   Jarque-Bera (JB):                0.327
Skew:                        -0.523   Prob(JB):                        0.849
Kurtosis:                     2.310   Cond. No.                         72.4
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 6: Prediction of number of crimes based on the number of male literates, women literates and total women population

5 Conclusions

In this project we have studied the demographics and made an attempt to predict the Indian states with cohesive environment for women entretrepreneurs and also a state with lesser number of crimes, with factors like literacy influencing these.

DATA FOR PART-I :

Demographics of India: To complete the analysis for the Indian demographics, data is obtained from different public domains such censusindia, districts and states of India, local government data etc. The data will be converted to csv files where it is not available in that format and will then be merged depending on the requirements.

http://www.censusindia.gov.in/2011census/populationenumeration.html
http://censusindia.gov.in/CensusAndYou/literacyandlevelofeducation.aspx
http://districts.nic.in/districts.php
http://www.mospi.gov.in/statistical-year-book-india/2017/171
https://data.gov.in/catalogsv2format=json&offset=0&limit=9&sort%5Bcreated%5D=desc
http://www.indiaenvironmentportal.org.in/files/file/Crimes%20in%20India%20Statistics-2014.pdf
http://ncrb.gov.in/StatPublications/CII/CII2016/pdfs/NEWPDFs/Crime%20in%20India%20-%202016%20Complete%20PDF%20291117.pdf

PART-2-SUGGESTED LOCALITIES WITH INDIAN RESTAURANTS IN AND AROUND BANGALORE

6 Introduction

Bangalore has become a hub for varied culture and demographics. With this vibrant culture, there exist a variety of eateries which include Indian and Multicuisine restaurants. An attempt is made to use Foursquare to obtain the different eateries and the list is then sorted to suggest locations or wards with maximum number of Indian restaurants.

7 Methodology

The execution methodology involves data collection, data cleaning, data sorting. Data is extracted from Wiki pages using BeautifulSoup and then merged with other files when required. These include csv files with ward details and that with latitude and longitude of each of the wards.

7.1 Data sources

Data sources include wiki pages with Bangalore ward details and web pages with latitude and longitude of each of the wards.

7.2 Data Cleaning

Data is sorted and columns related to ward name, ward number is retained. Rest are dropped for purposes of sorting and processing

8 Exploratory data analysis

Folium is used to plot the different wards of Bangalore. Foursquare is used to obtain data relevant to eateries and this is then sorted. Once all the data has been retrieved, those with Indian Restaurants are then obtained and suggestion as to which location is the best is provided. In this case, Jayanagar and Basavanagudi are the locations where maximum number of Indian Restaurants are present.

9 Results

A Map of Bangalore with the wards is first obtained using Folium

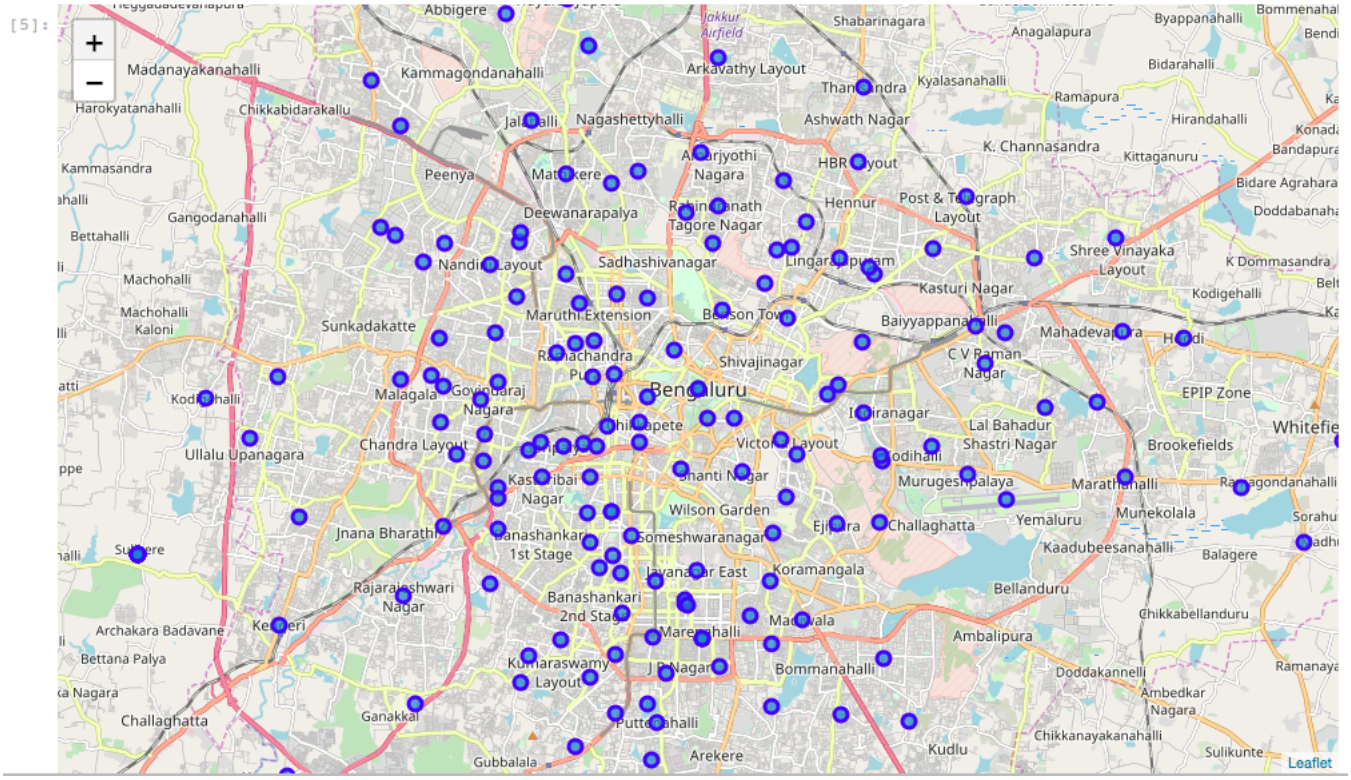

Figure 7: Wards in Bangalore

[4]:

	Ward_name	Ward_number_y	Latitude	Longitude
0	Kempegowda	1	13.109018	77.601900
1	Chowdeshwari	2	12.925190	77.588020
2	Attur	3	11.599586	78.596362
3	Yelahanka Satellite	4	13.095231	77.594296
4	Jakkur	5	13.078474	77.606894
5	Thanisandra	6	13.054713	77.633926
6	Byatarayanapura	7	13.062074	77.596392
7	Kodigehalli	8	12.976657	77.464564
8	Vidyaranyapura	9	13.076641	77.557731
9	Doddabommasandra	10	13.064967	77.562966
10	Kuvempu Nagar	11	13.073193	77.541713
11	Shettyhalli	12	12.884645	76.020329
12	Mallasandra	13	13.215966	78.159144
13	Bagalagunte	14	13.056476	77.507324
14	T-Dasarahalli	15	13.045141	77.514789

Figure 8: Wards in Bangalore and their co-ordinates

[16]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Kempegowda	13.109018	77.601900	Sri Raghavendra Food Line	13.111306	77.605188	Indian Restaurant
1	Chowdeshwari	12.925190	77.588020	1947	12.927642	77.586216	Indian Restaurant
2	Chowdeshwari	12.925190	77.588020	Meghana Foods	12.926237	77.584584	Indian Restaurant
3	Chowdeshwari	12.925190	77.588020	The Sofraah	12.923417	77.585262	Indian Restaurant
4	Chowdeshwari	12.925190	77.588020	Hot Chips	12.928670	77.585349	Indian Restaurant
5	Attur	11.599586	78.596362	Hotel Saravana Bhavan	11.599601	78.597274	Indian Restaurant
6	Byatarayanapura	13.062074	77.596392	Sanjay Dhaba	13.058612	77.593767	Indian Restaurant
7	Byatarayanapura	13.062074	77.596392	Swathi Gardenia	13.059108	77.593184	Indian Restaurant
8	Byatarayanapura	13.062074	77.596392	Bhagini Express	13.062840	77.592754	Indian Restaurant

Figure 9: Wards in Bangalore and the dfferent venue categories

```
[17]: Neighborhood
      Jayangar             9
      Basavangudi          9
      Maruthiseva Nagar    8
      Hosathippasandra     8
      Gandhi Nagar         8
      dtype: int64
```

Figure 10: Wards in Bangalore and wards sorted in accordance with the occurrence of Indian Restaurants

10 Conclusions

: In this section of the project we have made an attempt to suggest localities near Bangalore with maximum number of Indian restaurants.

DATA FOR PART-II : To make suggestions about Indian Restaurants in and around Bangalore, Foursquare is used to acquire all data pertaining to eateries and then sorting is done and areas with larger number of such restaurants is identified and suggestions made.

https://foursquare.com/explore?near=Bangalore,%20Karn%C4%81taka&cat=food
https://en.wikipedia.org/wiki/List of wards in Bangalore