IBM Professional Certificate Applied Data Science Capstone

Week – 5 (Final Report)

Using Data Science and ML to find the best location to open a restaurant in Maharashtra, India

- ATHARVA RAMGIRKAR

(nana chowk skywalk, Mumbai, Maharashtra, India)

(source : https://live.staticflickr.com/3919/15012790117_39732003d4_b.jpg)

INTRODUCTION

SOME BACKGROUND ON MAHARASHTRA, INDIA

With my basic understanding, India is divided into many states. Maharashtra is one of the states among those. It is located in the western part of the country. It is the second-most populous state and third-largest state by area. Spread over 307,713 km2 (118,809 sq mi). It is also the world's second-most populous subnational entity. It has over 112 million inhabitants and its capital, Mumbai, has a population around 18.4 million making it the most populous urban area in India. Nagpur hosts the winter session of the state legislature. Pune is known as the 'Oxford of the East' due to the presence of several well-known educational institutions. Nashik is known as the 'Wine Capital of India' as it has the largest number of wineries and vineyards in the country. Maharashtra is the most industrialised state in India while state capital Mumbai is India's financial and commercial capital. The state continues to be the single largest contributor to the national economy with a share of 15% in the country's gross domestic product (GDP). The economy of Maharashtra is the largest in India, with a gross state domestic product (GSDP) of ₹28.78 lakh crore and has the country's 13th-highest GSDP per capita of ₹208,000. Maharashtra has the 15th highest ranking among Indian states in human development index.

Looking at the above and considering the factors like popularity and population in the state of Maharashtra we can conclude that a restaurant business will be a decent idea. In this hypothetical scenario, an entrepreneur wants to open a restaurant in the state of Maharashtra. But is unable to decide the exact location of this new restaurant.

He appoints me to figure out some optimal locations in the state to find open the restaurant. This Capstone will use location data and via the Foursquare API will explore the venues around these locations. Finally based on K-Means clustering we will be able to cluster the neighbourhoods and inspecting them will allow us to come up with the optimal location for this new restaurant.

BUSINESS PROBLEM

The objective of this project will be to find the optimal location for an entrepreneur to open a new restaurant in the Maharashtra state in India. Using data science and Machine Learning Algorithms to figure out the optimal location.

We will do this and answer the question: - In Maharashtra, if an entrepreneur wants to open a restaurant, what should be the optimal location in the state?

TARGET AUDIENCE

This project is very useful to entrepreneurs who want to open a new restaurant in Maharashtra State, India. This project is also very useful now as Maharashtra is a growing state and with the large population there will be large demands to restaurants and even small eateries in the state.

As stated earlier, the state of Maharashtra has 'Oxford of the East' which will surely drive in a lot of youth population which likes to eat outside rather than eat at home.

DATA

(MAHARASHTRA, INDIA)

(Source : https://upload.wikimedia.org/wikipedia/commons/thumb/1/16/IN-MH.svg/250px-IN-MH.svg.png)

1. FROM THE INTERNET

To solve this capstone, we need lots of data. But to keep the project simple and a bit basic, I have not included complicated data like average income of the residents, population density, poverty rate, etc.

First of all we need the list of the subdivisions of Maharashtra state into various Districts. I got this data from the following website: -

https://www.census2011.co.in/census/state/districtlist/maharashtra.html

Next, I needed the division of the Districts into Talukas. This data was found on this website: -

https://en.wikipedia.org/wiki/List_of_talukas_of_Maharashtra

The information used to find facts on Maharashtra, India was found on Wiki: -

https://en.wikipedia.org/wiki/Maharashtra

2. INDIRECT DATA

- Once we get the data of the Districts and Talukas in Maharashtra State, we use Geocoder to get the exact Longitude and Latitude coordinates of the place.
- Once we have the location coordinates of the Talukas we can input the Talukas in the Foursquare API to get the information on the nearby venues.

METHODOLOGY

Step 1: -

We need to look into the structure of the region.

- How is the state of Maharashtra divided?
- What are the Subdivisions?
- To what level should we divide the state for the exploratory analysis?

So, after looking up on the internet, we find that Maharashtra state is divided into several Districts. Each district is further divided into several Talukas. For this project we mainly focus on the Taluka region. From some analysis we find that there are around 356 Talukas in Maharashtra state. If we plan to divided the Talukas further, the data might be excessive and might take lot of time to load. It might also exceed the Foursquare API calls limit.

We find the list of Talukas and Districts from the above-mentioned sources.

Not logged in Talk Contributions Create account Log in

Article Talk | Read Edit View history | Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Download as PDF
Printable version

Languages
मराठी
नेपाल भाषा
Edit links

List of talukas of Maharashtra

From Wikipedia, the free encyclopedia

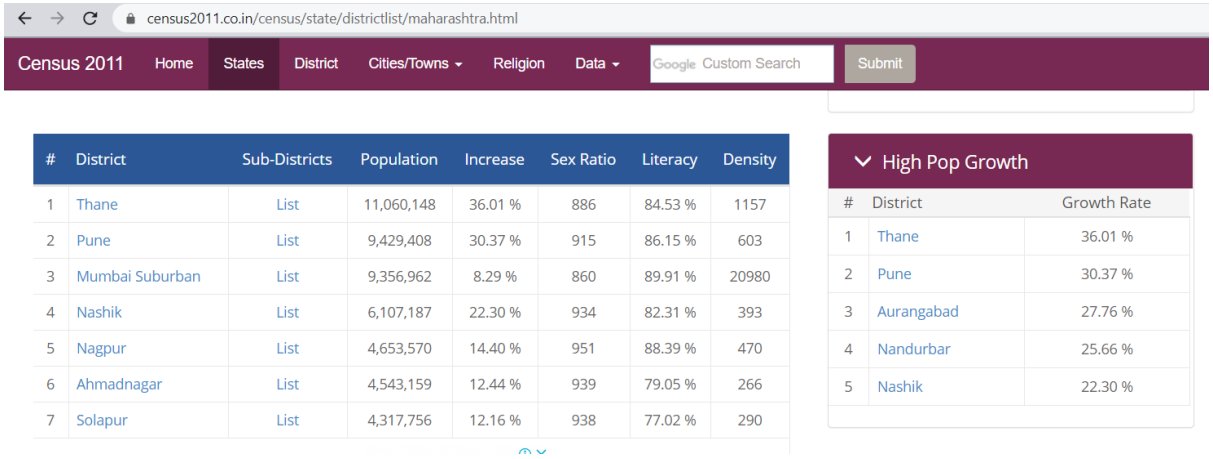
The Table below list all the talukas (tahsils/tehsils) of all the thirty-six districts in the Indian state of Maharashtra, along with district-subdivision and urban status information of headquarters villages/towns, as all talukas are intermediate level panchayat between the zilla panchayat (district councils) at the district level and gram panchayat (village councils) at the lower level.[1]

Table [edit]

District	District Subdivision	Taluka	
Sindhudurg district	Kankavli 416602	Kankavli	
		Vaibhavwadi 416810	
		Devgad 416613	
		Malwan 416606	
	Sawantwadi 416510	Sawantwadi	
		Kudal 416520	
		Vengurla 416512	
		Dodamarg 416512 (Kasal)	
Ratnagiri district	Ratnagiri	Ratnagiri	
		Sangameshwar 415611 (Deorukh)	i
		Lanja 416701	
		Rajapur 416702	
	Chiplun	Chiplun 415605	
		Guhagar 415703	
		Dapoli 415712	
		Mandangad 415203	
		Khed 415718	
Raigad district	Alibaug	Pen 402107	
		Alibaug	
		Murud 413510	

(Here is a screenshot of the table found on Wikipedia)

Additional information of the Districts can be found from one of the above mentioned links.



#	District	Sub-Districts	Population	Increase	Sex Ratio	Literacy	Density
1	Thane	List	11,060,148	36.01 %	886	84.53 %	1157
2	Pune	List	9,429,408	30.37 %	915	86.15 %	603
3	Mumbai Suburban	List	9,356,962	8.29 %	860	89.91 %	20980
4	Nashik	List	6,107,187	22.30 %	934	82.31 %	393
5	Nagpur	List	4,653,570	14.40 %	951	88.39 %	470
6	Ahmadnagar	List	4,543,159	12.44 %	939	79.05 %	266
7	Solapur	List	4,317,756	12.16 %	938	77.02 %	290

#	District	Growth Rate
1	Thane	36.01 %
2	Pune	30.37 %
3	Aurangabad	27.76 %
4	Nandurbar	25.66 %
5	Nashik	22.30 %

(Here is another table with more detailed information about each District)

We import these tables into the PyNotebooks and perform some exploratory Data Analysis and Data Cleaning on it. We create few data frames to hold the necessary columns from the above tables.

```
In [23]: MH_Dist_Tal = pd.read_html("https://en.wikipedia.org/wiki/List_of_talukas_of_Mah
```

```
In [24]: MH_Dist_Tal_df = MH_Dist_Tal[0]
         MH_Dist_Tal_df.head()
```

Out[24]:

	District	District Subdivision	Taluka	Unnamed: 3
0	Sindhudurg district	Kankavli 416602	Kankavli	NaN
1	NaN	NaN	Vaibhavwadi 416810	NaN
2	NaN	NaN	Devgad 416613	NaN
3	NaN	NaN	Malwan 416606	NaN
4	NaN	Sawantwadi 416510	Sawantwadi	NaN

```
In [25]: MH_Dist_Tal_df.drop(['District Subdivision','Unnamed: 3'],axis=1,inplace=True)
         MH_Dist_Tal_df.head(10)
```

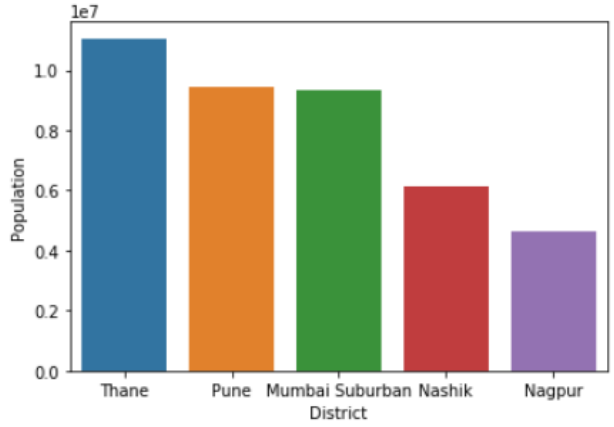
Out[25]:

	District	Taluka
0	Sindhudurg district	Kankavli
1	NaN	Vaibhavwadi 416810
2	NaN	Devgad 416613
3	NaN	Malwan 416606
4	NaN	Sawantwadi
5	NaN	Kudal 416520
6	NaN	Vengurla 416512
7	NaN	Dodamarg 416512 (Kasal)
8	Ratnagiri district	Ratnagiri
9	NaN	Sangameshwar 415611 (Deorukh)

```
In [146]: # Visualizing the Districts on a barplot using the seaborn library

          sns.barplot(y=MH_top5['Population'],x=MH_top5['District'])

Out[146]: <matplotlib.axes._subplots.AxesSubplot at 0x1d1922ba550>
```



(Data Visualization to see the District with maximum population)

Step 2: -

After we have successfully imported the data from the internet, we will find the Longitude and Latitude of each Taluka. This will allow us to plot these Talukas on the map of Maharashtra, India using the **Folium library.** We get the coordinates from the **Geocoder** library.

Once the Coordinates are retrieved from the Geocoder library, we save them and add them to the table so the coordinates correspond to the particular Taluka.

We then make a new Data Frame with the Talukas and their coordinates.

```
In [31]: MH_latlng_df.head(10)
```

Out[31]:

	District	Taluka	Latitude	Longitude
0	Sindhudurg district	Kankavli	16.569380	73.668790
1	Sindhudurg district	Vaibhavwadi 416810	16.505630	73.768000
2	Sindhudurg district	Devgad 416613	16.381198	73.394170
3	Sindhudurg district	Malwan 416606	16.104241	73.502825
4	Sindhudurg district	Sawantwadi	15.903190	73.823350
5	Sindhudurg district	Kudal 416520	16.016940	73.678220
6	Sindhudurg district	Vengurla 416512	15.852100	73.632210
7	Sindhudurg district	Dodamarg 416512 (Kasal)	15.748480	74.042510
8	Ratnagiri district	Ratnagiri	17.079650	73.400640
9	Ratnagiri district	Sangameshwar 415611 (Deorukh)	17.188715	73.512550

```
In [32]: MH_latlng_df.to_csv("maharashtra.csv", index=False)

In [34]: # get the coordinates of Maharashtra
```

Use folium to visualize these places.

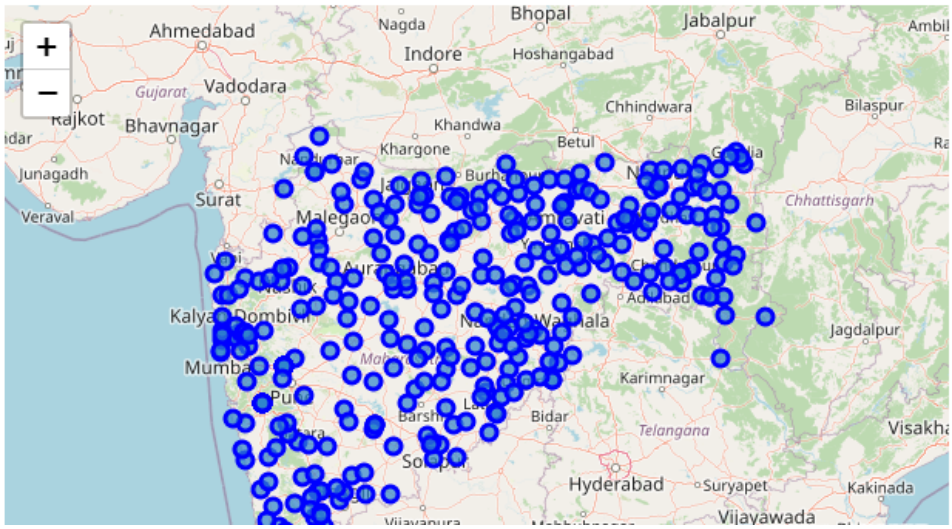
Step 4: Creating a map of Maharashtra, India with Talkuas superimposed on it

In [36]:
```
# create map of Toronto using latitude and longitude values
map_mh = folium.Map(location=[latitude, longitude], zoom_start=6)

# add markers to map
for lat, lng, Taluka, District in zip(MH_df['Latitude'],
                                      MH_df['Longitude'],
                                      MH_df['Taluka'],
                                      MH_df['District']):
    label = '{},{}'.format(Taluka,District)
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7).add_to(map_mh)

map_mh
```

Out[36]:



Step 3: -

Next, we use the **Foursquare API** to get the data of the nearby venues of the Talukas. We use the Client_ID and the Client_Secrets to access the venues. The Foursquare API returns the data in a JSON file. We extract the required information from this file for example, ‘venue category’, ’venue latitude’, ‘venue longitude’, etc.

Now we examine each venue using the provided data. Ur project focuses on restaurants. So we group each District and Taluka and check for ‘Restaurants’ in their ‘Venue names’. We then take the mean of the frequency of occurrence of each venue category and find the ‘Restaurants’ as venue category.

Step 4: -

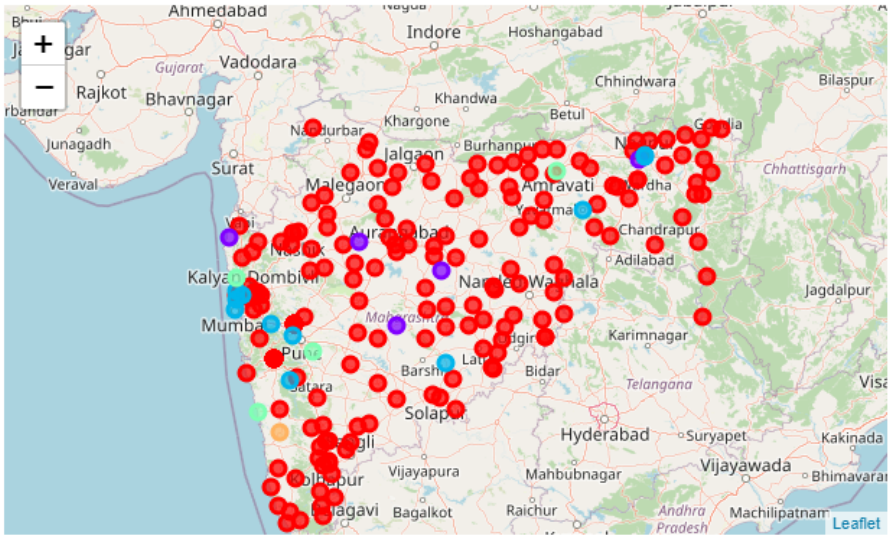
Now we use the **K-Means clustering** to perform the clustering of data. In this project, due to the vast area of the state, we plan to cluster the data into 5 clusters. The method generates 5 centroids on the data and allocates each data point(Taluka) to one of the five clusters.

After segmentation we visualise the clusters using **Folium.**

```
                                    mh_merged['Longitude'],
                                    mh_merged['Taluka'],
                                    mh_merged['Cluster Labels']):
    label = folium.Popup(str(poi) + ' - Cluster ' + str(cluster), parse_html=Tru
    folium.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.7).add_to(map_clusters)

map_clusters
```

Out[69]:




```
In [70]: # save the map as HTML file
```

Step 5: -

Finally, we analyse each cluster to find the optimum cluster and the optimum location in Maharashtra state to start a new Restaurant.

66	Hadgaon 431717	0.0	0	Nanded district	18.552840	77.572378
67	Hatkanangale 416109	0.0	0	Kolhapur district	16.752310	74.277330
61	Georai 431127	0.0	0	Beed district	19.267490	75.745510
84	Kalamb 445401	0.0	0	Yavatmal district	20.468415	78.365271

187 rows × 6 columns

Cluster Label : 1

```
In [72]: mh_merged.loc[mh_merged['Cluster Labels'] == 1]
```

Out[72]:

	Taluka	Restaurant	Cluster Labels	District	Latitude	Longitude
196	Vaijapur 423701	0.250000	1	Aurangabad district	19.92509	74.72936
76	Jamkhed 413201	0.250000	1	Ahmednagar district	18.73425	75.31078
71	Hingna 441110	0.250000	1	Nagpur district	21.10039	78.98157
45	Dahanu 401602	0.285714	1	Palghar district	19.98848	72.74471
62	Ghansawangi 431209	0.250000	1	Jalna district	19.52122	75.98533

Cluster Label : 2

```
In [73]: mh_merged.loc[mh_merged['Cluster Labels'] == 2]
```

Out[73]:

	Taluka	Restaurant	Cluster Labels	District	Latitude	Longitude
15	Andheri 400069	0.030000	2	Mumbai Suburban District	19.10393	72.86698
59	Gaganbawada	0.047619	2	Kolhapur district	21.17497	79.07479
101	Khandala 410301	0.037037	2	Satara district	18.76027	73.38759

RESULT

The examination of the cluster brings forward these results: -

1. Cluster Label 0 has the least (pretty much zero) number of restaurants
2. Cluster 1 and Cluster 4 have very high competition.
3. Cluster 2 and Cluster 3 have fair number of restaurants.
4. Top 5 populated Districts: Thane, Pune, Mumbai Suburban, Nashik, Nagpur

DISCUSSION

Cluster 0 has very less competition and almost zero restaurants this is a great opportunity for the business investors who want to open a restaurant in the Maharashtra State as in Cluster Label 0 there is no Competition. Cluster 4 has a single competitor but the mean is quite high suggesting that the competition will be hard. Cluster 1 also has high competition. So, opening a restaurant in any of Cluster 1 or 4 would be a bad idea due to existing competitors and already set up numerous restaurants. Cluster 2 and Cluster 3 have a moderate number of restaurants but the mean is not as high. So, there is competition but the competition is not much resistant. Add on further analysis of the Districts we find the top 5 populated Districts. Thus the project recommends that the entrepreneur opens the restaurant in Cluster 0 and in one of the top 5 populated Districts namely Thane, Pune, Mumbai Suburban, Nashik, Nagpur.

LIMITATION IN THE PROJECT

The project was kept simple and not much data was analysed. The literacy rate, poverty, average income of the residents per Taluka, population density in each region and many such factors are ignored in this project. The project mainly focused on the population and the number of restaurants set up in each Taluka and tells the entrepreneur to set up his restaurant in the area with less competition and more population and then used k-Means clustering to find the optimal location.

The project also uses Foursquare Sandbox account which is basic and limited to use. So much information of the venues could not be retrieved from the API.

CONCLUSION

Even thought the project is simple and very basic, it goes through some of the most important segments in the firld of Data Science and Machine Learning including things like Data Analysis, Visualization, Data Exploration, K-Means Clustering and also using API's

Finally, the business question is also answered.

Question: -

I am an entrepreneur and want to set up a new restaurant in the state of Maharashtra, where is the optimum location to set it up?

Answer: -

Any of the Talukas in Cluster 0 and among the following Districts - Thane, Pune, Mumbai Suburban, Nashik, Nagpur.

The findings in the project provides relevant information the stakeholders and helps them make some crucial financial decisions based on the location of high potential to open a new restaurant in Maharashtra State, India.

THANK YOU