IBM Developer
SKILLS NETWORK

Best Place to setup Business

Sagar Ratnaparkhi
27-09-2021

Outline

Contents

1. Introduction

1.1. Background Information

1.2. Problem Statement

2. Data

3. Methodology

3.1. Pune Regions, Postal Codes, and Geographical Coordinates

3.2. Exploring Venues in the Subdistricts of Pune

3.3. Most Common Venues Overall

3.4. One-Hot Encoding

3.5. Clustering Subdistricts Based on Venues Similarity

3.6. Examining Each Cluster

3.7. Visualizing Distribution of Coffee Shop Locations in Pune City

4. Results and Discussion

5. Conclusion

Introduction

Pune, is the second largest city in the state of Maharashtra and the seventh most populous city in India, with an estimated population of 7.4 million as of 2020. It has been ranked as "the most livable city in India" several times. Along with the municipal corporation limits of PCMC and the three cantonment towns of Camp, Khadki and Dehu Road, Pune forms the urban core of the eponymous Pune Metropolitan Region (PMR).

According to the 2011 census the urban area had a combined population of 5.05 million whilst the population of the metropolitan region was estimated at 7.4 million. Situated 560 metres (1,837 feet) above sea level on the Deccan plateau on the right bank of the Mutha river,[26] Pune is also the administrative headquarters of its namesake district.The largest city of Maharashtra, Pune contributes a GDP(PPP) of $78 billion.

Because of this people from across the archipelago to move to the city in search of opportunities and a potentially better standard of living.

Business opportunities abound in Pune City, but the food-and-beverage (F&B) sector has long been an attractive target for investors. It has recorded the largest investment realization among secondary sectors in Indonesia over the last five years, totaling IDR 293 trillion2. According to a research by Toffin3, the coffee shop has been a booming F&B business in Indonesia, reflected on the significant rise in number of outlets and domestic coffee consumptions in the recent years4. The market value of coffee shops is also estimated to reach over IDR 4 trillion per year.

1.2. Problem Statement

With the aforementioned prospect, various stakeholders (entrepreneurs, investors) may be interested to explore coffee shop business opportunities in Pune City. This data science project is thus carried out to help them answer the following question:

“Which of the Pune City regions are strategic for opening a coffee shop business?”

Apart from business stakeholders, the project may also be of interest to fellow coffee enthusiasts.

2. Data

In order to explore potential answer to the problems, the following data are required:

1. The names of administrative regions in Pune City and their corresponding postal codes. The regions include three levels of subdivision: city, district, subdistrict. The region names are useful to perform analysis across different sub-regions. The postal codes are needed to obtain coordinates of each subdistricts.

2. Geographical coordinates of Pune City and its subdistricts, which will in turn be needed to utilize Foursquare API in the subsequent step.

3. Information about venues in Pune City regions: the names, category, venue latitudes, venue longitudes. These are obtained using Foursquare API. The subdistricts of Pune City will be clustered based on their surrounding venues to find the best location candidates for opening a coffee shop.

Methodology

Methodology

- Web API call made to fetch pune regions and postal codes as well as retrieval of geographical coordinates. Leveraging Foursquare API, these coordinates data were given as inputs to explore venues within the Pune subdistricts. Two dataframes were then created for use in the analysis:
- 1. ***df***: contains postal codes and geographical coordinates of all Jakarta regions (city, district, subdistrict).
- 2. ***pune_venues***: contains at most 100 venues and venues details (name, category, latitude, longitude) for every subdistrict in Jakarta.

- One-hot encoding was performed to analyze and to narrow down the most common venues in each of the subdistricts. Given all the venues surrounding them, subdistricts were clustered using *K*-means algorithm. The number of optimal clusters was decided using the elbow method and silhouette score. Each cluster was separately analyzed in order to examine one discriminating venue that characterizes them. Analysis of the clusters and visualization of coffee shop distribution across Jakarta would provide insights as to where the strategic regions to set up the business are.
- The following Python libraries and dependencies were used: pandas, NumPy, string, Requests, time.sleep, BeautifulSoup, GeoPy (Nominatim geocoder), JSON, Folium, Matplotlib, and scikit-learn.

3.1. Pune City Regions, Postal Codes, and Geographical Coordinates

The data to scrape are the names of all Pune City regions and their corresponding postal codes. For reference, Pune City is a province that consists of 5 cities (mainland Pune City) and 1 regency:

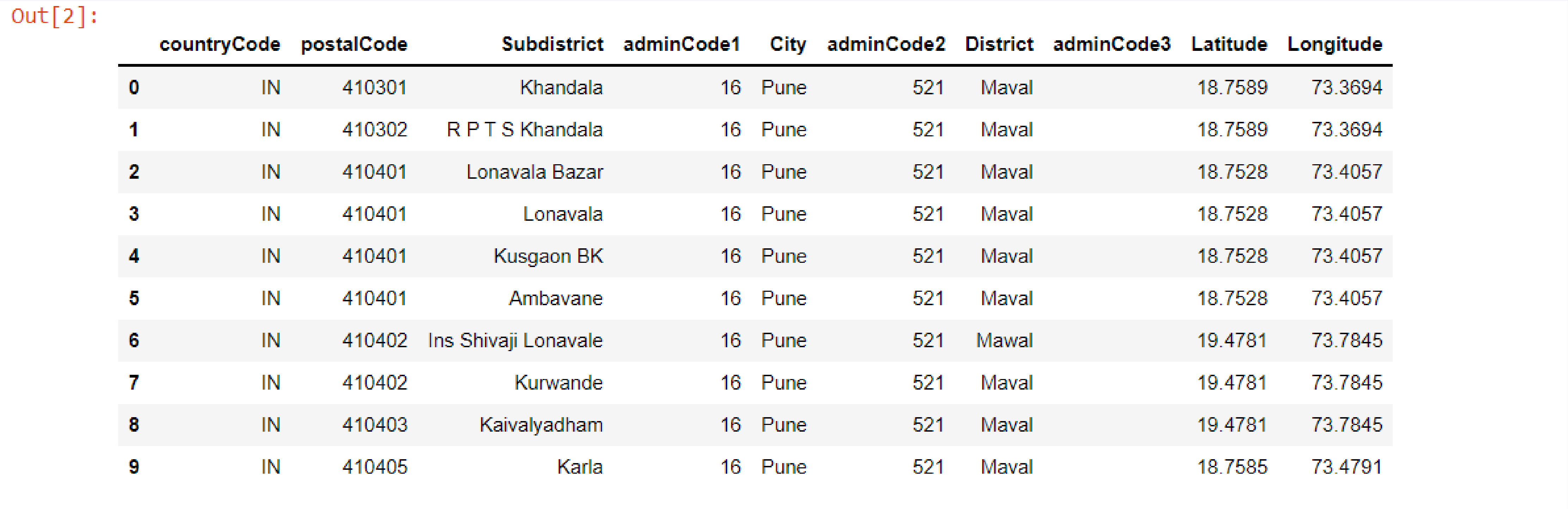
1. Pune City Pusat (Central Pune City)

2. Maval(North Pune City)

3. Mulshi (West Pune City)

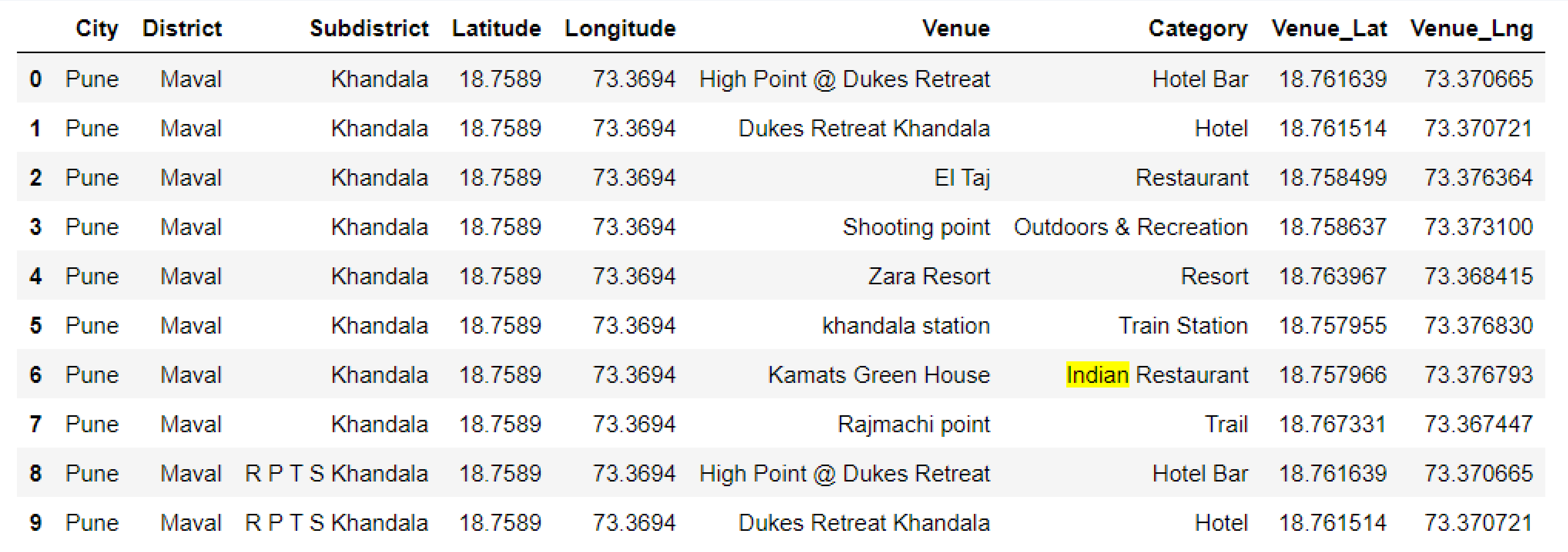
Each of these cities is further subdivided into districts and then subdistricts/ small colonies. In total, there are 44 districts and 267 subdistricts across Pune City. Pune City Selatan and Pune City Timur are tied as the cities with the highest number of sub-regions, each having 65 subdistricts.

Used API to retrieve latitudes and longitudes of every subdistrict. Figure 1 displays the first 10 rows of the resulting dataframe: post.



Out[2]:

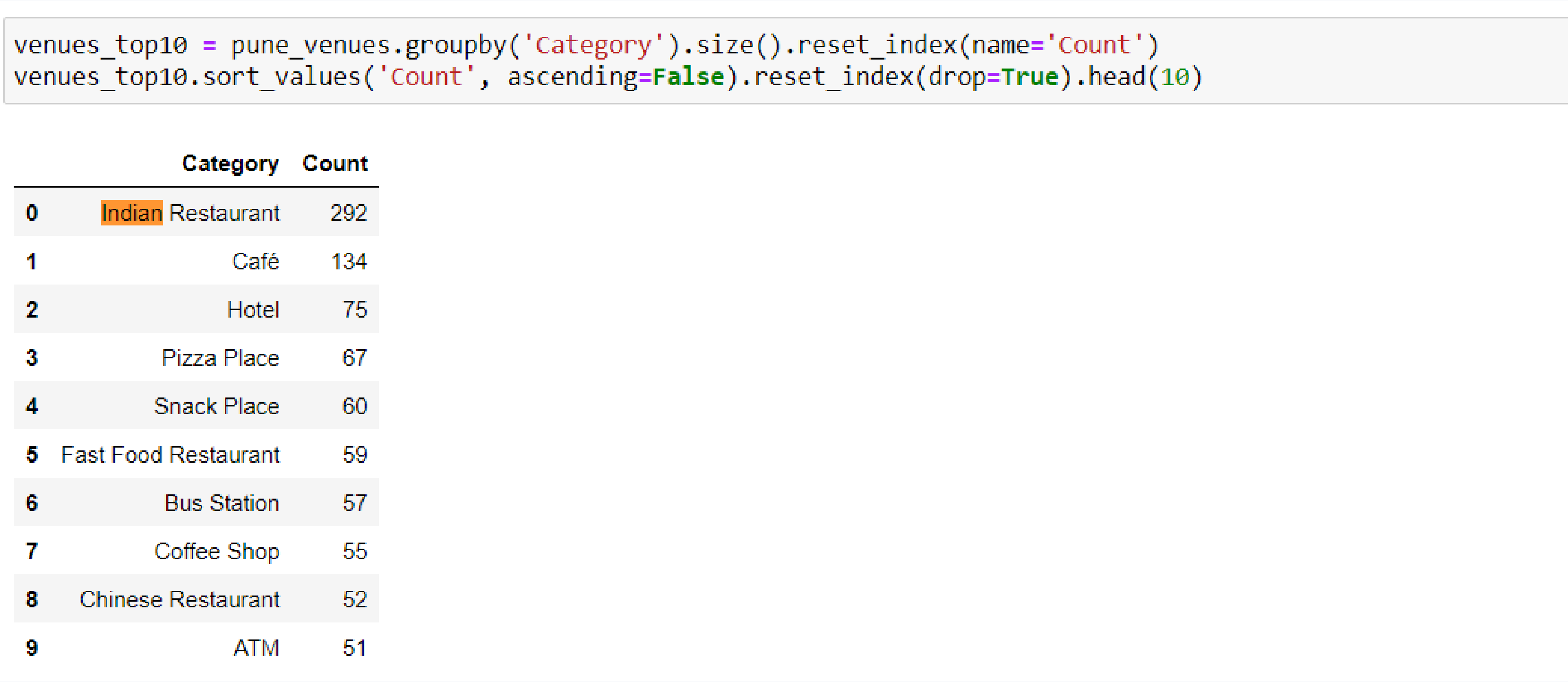
	countryCode	postalCode	Subdistrict	adminCode1	City	adminCode2	District	adminCode3	Latitude	Longitude
0	IN	410301	Khandala	16	Pune	521	Maval		18.7589	73.3694
1	IN	410302	R P T S Khandala	16	Pune	521	Maval		18.7589	73.3694
2	IN	410401	Lonavala Bazar	16	Pune	521	Maval		18.7528	73.4057
3	IN	410401	Lonavala	16	Pune	521	Maval		18.7528	73.4057
4	IN	410401	Kusgaon BK	16	Pune	521	Maval		18.7528	73.4057
5	IN	410401	Ambavane	16	Pune	521	Maval		18.7528	73.4057
6	IN	410402	Ins Shivaji Lonavale	16	Pune	521	Mawal		19.4781	73.7845
7	IN	410402	Kurwande	16	Pune	521	Maval		19.4781	73.7845
8	IN	410403	Kaivalyadham	16	Pune	521	Maval		19.4781	73.7845
9	IN	410405	Karla	16	Pune	521	Maval		18.7585	73.4791



	City	District	Subdistrict	Latitude	Longitude	Venue	Category	Venue_Lat	Venue_Lng
0	Pune	Maval	Khandala	18.7589	73.3694	High Point @ Dukes Retreat	Hotel Bar	18.761639	73.370665
1	Pune	Maval	Khandala	18.7589	73.3694	Dukes Retreat Khandala	Hotel	18.761514	73.370721
2	Pune	Maval	Khandala	18.7589	73.3694	El Taj	Restaurant	18.758499	73.376364
3	Pune	Maval	Khandala	18.7589	73.3694	Shooting point	Outdoors & Recreation	18.758637	73.373100
4	Pune	Maval	Khandala	18.7589	73.3694	Zara Resort	Resort	18.763967	73.368415
5	Pune	Maval	Khandala	18.7589	73.3694	khandala station	Train Station	18.757955	73.376830
6	Pune	Maval	Khandala	18.7589	73.3694	Kamats Green House	Indian Restaurant	18.757966	73.376793
7	Pune	Maval	Khandala	18.7589	73.3694	Rajmachi point	Trail	18.767331	73.367447
8	Pune	Maval	R P T S Khandala	18.7589	73.3694	High Point @ Dukes Retreat	Hotel Bar	18.761639	73.370665
9	Pune	Maval	R P T S Khandala	18.7589	73.3694	Dukes Retreat Khandala	Hotel	18.761514	73.370721

3.3. Most Common Venues Overall

As shown in Figure 4, various kinds of restaurant top the list of most common venues in Pune City. Coffee shop, which is our venue of interest, comes in second. With almost 1000 coffee shops in Pune City, it sure is a quite competitive business.

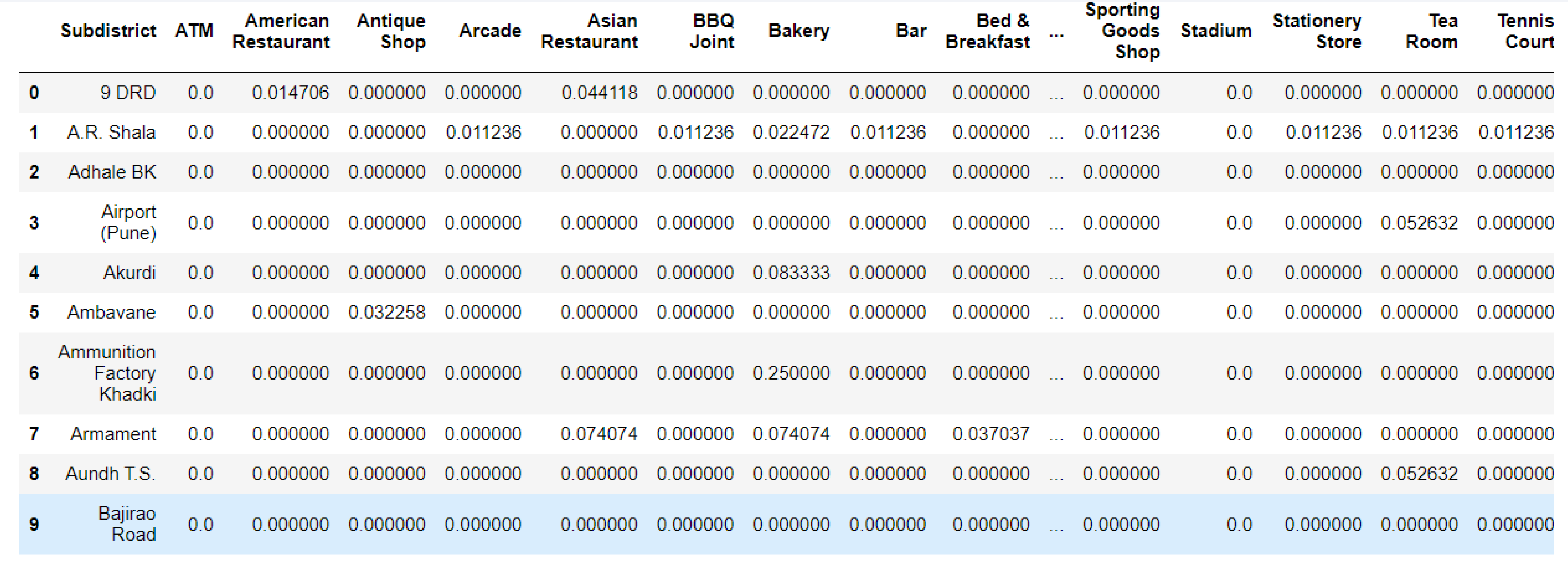


```
venues_top10 = pune_venues.groupby('Category').size().reset_index(name='Count')
venues_top10.sort_values('Count', ascending=False).reset_index(drop=True).head(10)
```

	Category	Count
0	Indian Restaurant	292
1	Café	134
2	Hotel	75
3	Pizza Place	67
4	Snack Place	60
5	Fast Food Restaurant	59
6	Bus Station	57
7	Coffee Shop	55
8	Chinese Restaurant	52
9	ATM	51

3.4. One-Hot Encoding

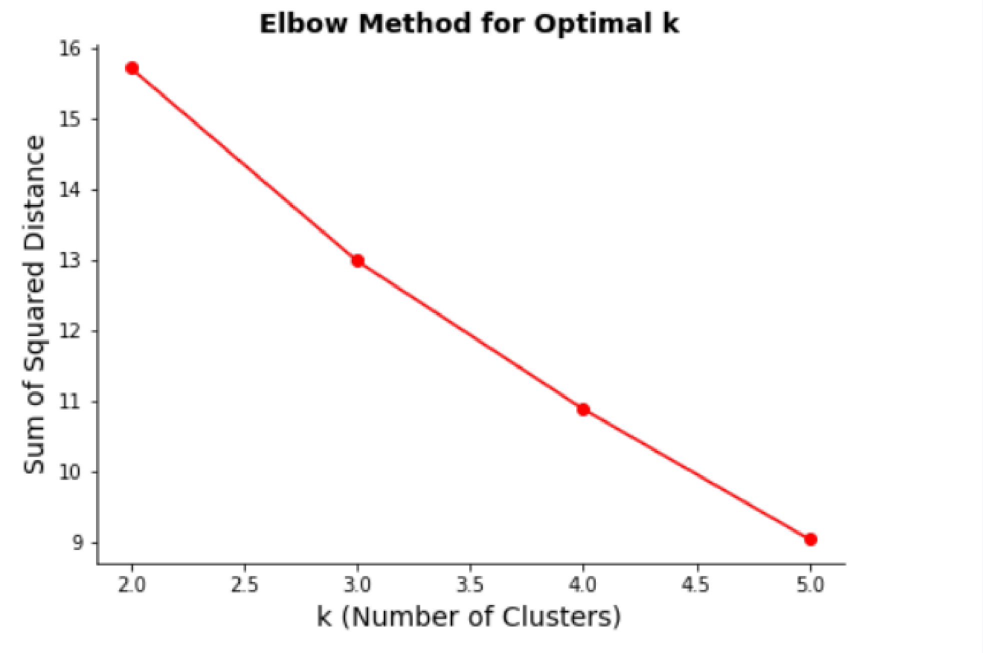
One-hot encoding converts categorical variables (i.e., venues) into numeric variables. In this case, a dummy of all the venues was made and the mean of the frequency of venue occurrence were calculated. The dataframe is then grouped by subdistrict, as

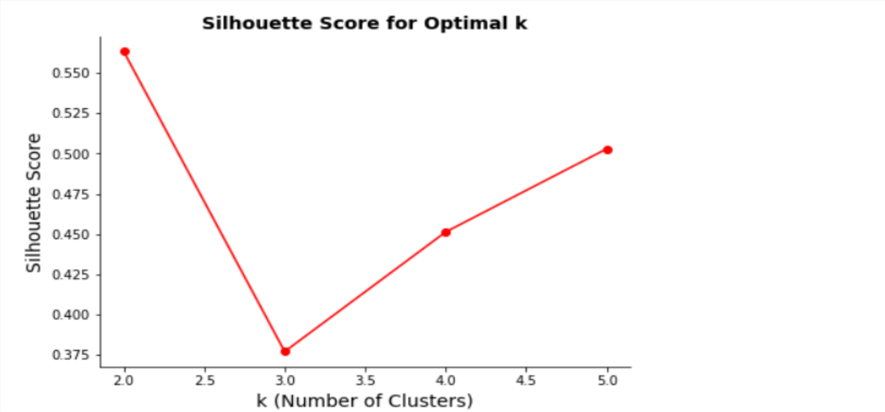


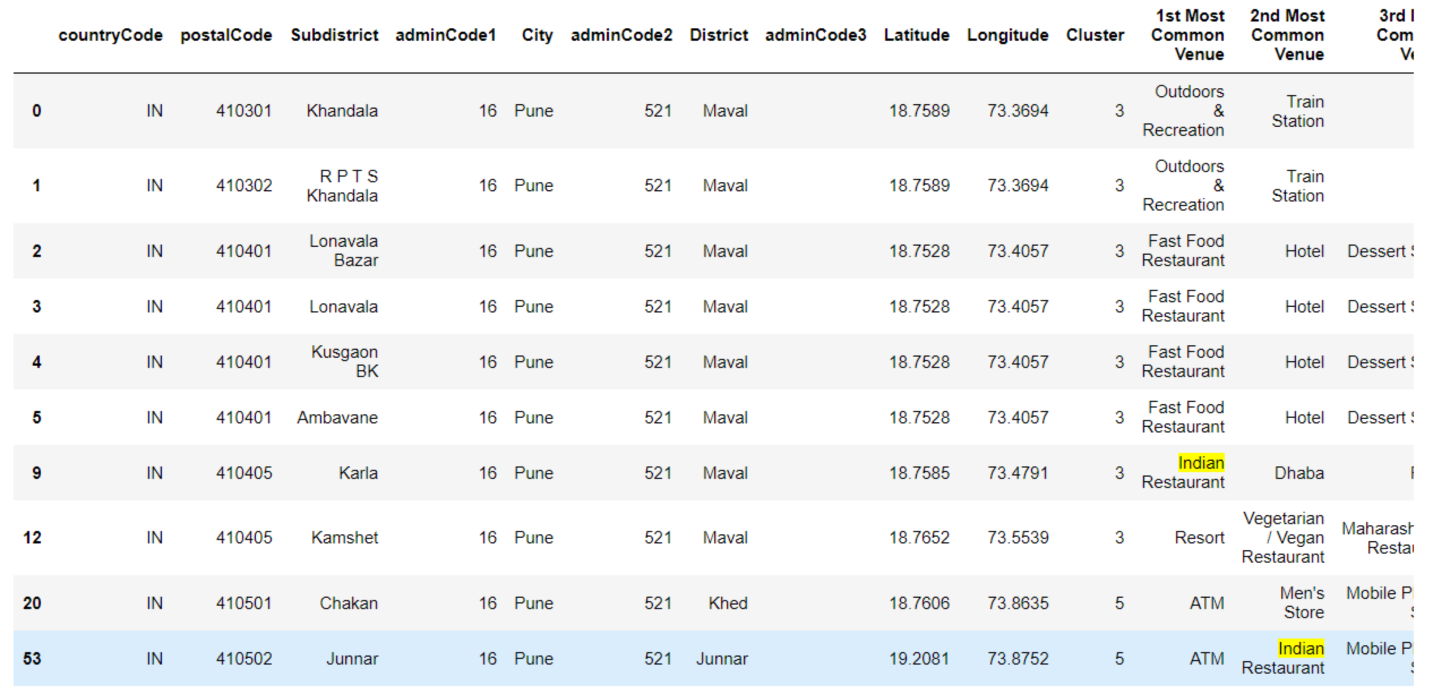
	Subdistrict	ATM	American Restaurant	Antique Shop	Arcade	Asian Restaurant	BBQ Joint	Bakery	Bar	Bed & Breakfast	...	Sporting Goods Shop	Stadium	Stationery Store	Tea Room	Tennis Court
0	9 DRD	0.0	0.014706	0.000000	0.000000	0.044118	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.0	0.000000	0.000000	0.000000
1	A.R. Shala	0.0	0.000000	0.000000	0.011236	0.000000	0.011236	0.022472	0.011236	0.000000	...	0.011236	0.0	0.011236	0.011236	0.011236
2	Adhale BK	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.0	0.000000	0.000000	0.000000
3	Airport (Pune)	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.0	0.000000	0.052632	0.000000
4	Akurdi	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.083333	0.000000	0.000000	...	0.000000	0.0	0.000000	0.000000	0.000000
5	Ambavane	0.0	0.000000	0.032258	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.0	0.000000	0.000000	0.000000
6	Ammunition Factory Khadki	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.250000	0.000000	0.000000	...	0.000000	0.0	0.000000	0.000000	0.000000
7	Armament	0.0	0.000000	0.000000	0.000000	0.074074	0.000000	0.074074	0.000000	0.037037	...	0.000000	0.0	0.000000	0.000000	0.000000
8	Aundh T.S.	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.0	0.000000	0.052632	0.000000
9	Bajirao Road	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.0	0.000000	0.000000	0.000000

3.5. Clustering Subdistricts Based on Venues Similarity

The subdistricts were clustered based on a set of similar characteristics or features, i.e., their surrounding venues. K-Means clustering, which was used in this part of the analysis, is a machine learning algorithm that creates homogeneous subgroups/clusters from unlabeled data such that data points in each cluster are as similar as possible to each other according to a similarity measure (e.g., Euclidian distance).







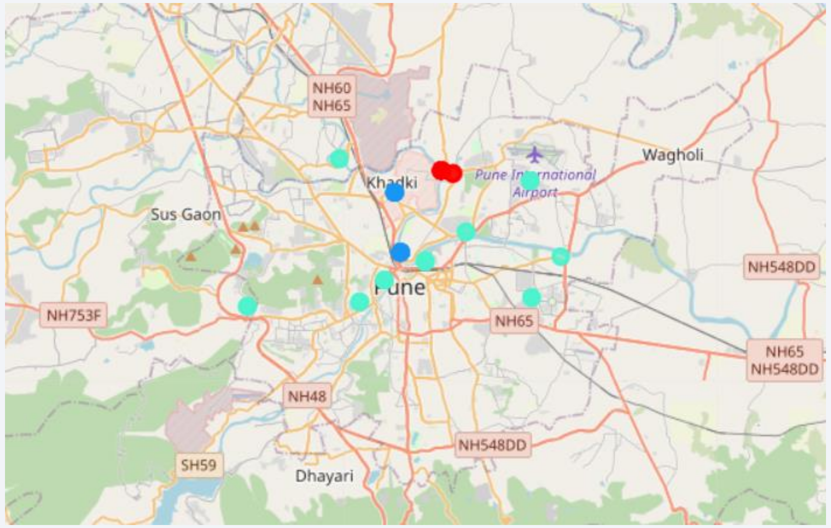
	countryCode	postalCode	Subdistrict	adminCode1	City	adminCode2	District	adminCode3	Latitude	Longitude	Cluster	1st Most Common Venue	2nd Most Common Venue	3rd I Com Ve
0	IN	410301	Khandala	16	Pune	521	Maval		18.7589	73.3694	3	Outdoors & Recreation	Train Station	
1	IN	410302	R P T S Khandala	16	Pune	521	Maval		18.7589	73.3694	3	Outdoors & Recreation	Train Station	
2	IN	410401	Lonavala Bazar	16	Pune	521	Maval		18.7528	73.4057	3	Fast Food Restaurant	Hotel	Dessert S
3	IN	410401	Lonavala	16	Pune	521	Maval		18.7528	73.4057	3	Fast Food Restaurant	Hotel	Dessert S
4	IN	410401	Kusgaon BK	16	Pune	521	Maval		18.7528	73.4057	3	Fast Food Restaurant	Hotel	Dessert S
5	IN	410401	Ambavane	16	Pune	521	Maval		18.7528	73.4057	3	Fast Food Restaurant	Hotel	Dessert S
9	IN	410405	Karla	16	Pune	521	Maval		18.7585	73.4791	3	Indian Restaurant	Dhaba	I
12	IN	410405	Kamshet	16	Pune	521	Maval		18.7652	73.5539	3	Resort	Vegetarian / Vegan Restaurant	Maharash Restau
20	IN	410501	Chakan	16	Pune	521	Khed		18.7606	73.8635	5	ATM	Men's Store	Mobile P S
53	IN	410502	Junnar	16	Pune	521	Junnar		19.2081	73.8752	5	ATM	Indian Restaurant	Mobile P S

A value of k (number of clusters) needs to be defined before proceeding with the clustering. The “Elbow Method” was used, which calculates the sum of squared distances of data points to their closest centroid (cluster center) for different values of k.

The optimal value of k is the one after which there is a plateau (no significant decrease in sum of squared distances).

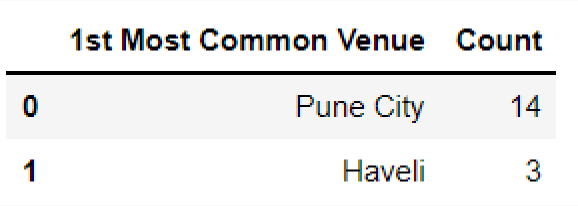
However, because there is no discernible "elbow" from the plot (Figure 7), another measure was used: “Silhouette Score”. Silhouette score varies from -1 to 1. A score value of 1 means the cluster is dense and well-separated from other clusters.

A value nearing 0 represents overlapping clusters, data points are close to the decision boundary of neighboring clusters. A negative score indicates that the samples might have been assigned into the wrong clusters. Given that there is a peak at k = 6 , the K-Means clustering was proceeded with that value.



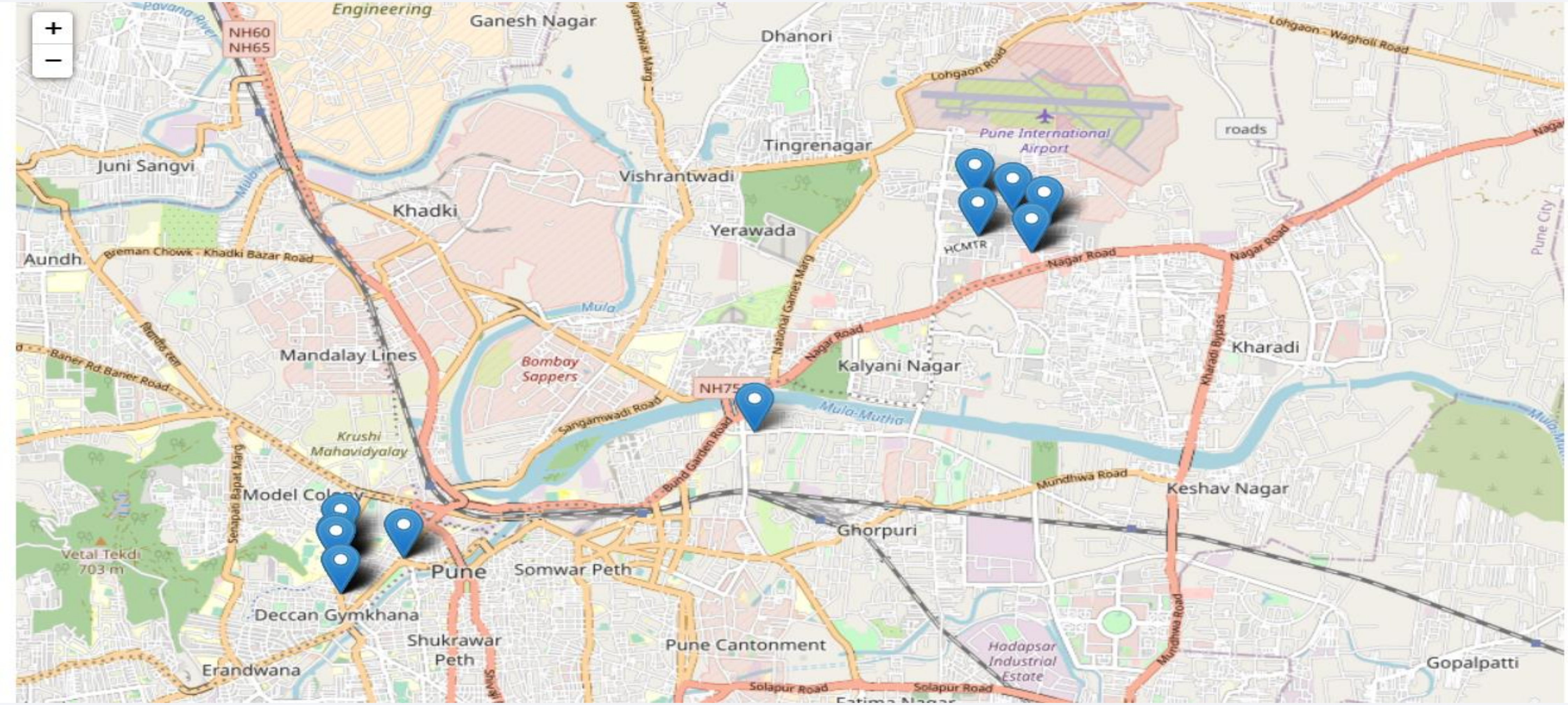
3.6. Examining Each Cluster

Each cluster was filtered from the dataframe previously created in the clustering stage. The clusters were separately analyzed in order to gain an understanding of a discriminating venue that characterize each of them. The number one most common venue categories from each cluster, as well as the regions (cities) in which a particular cluster is highly concentrated were singled out.



	1st Most Common Venue	Count
0	Pune City	14
1	Haveli	3

Visualizing Distribution of Coffee Shop Locations in Pune City



Results

Exploratory data analysis as well as machine learning and visualization techniques have provided us with some insights into the problem at hand.

A total of 1915 venues from all Pune regions (249 subdistricts) were returned at the time the API call was made. There are on average 55 venues within a kilometer of a subdistrict center, where two of the most common categories overall are Indian Restaurant and Coffee Shops.

After deciding on an optimal *k* value of 6, *K*-Means algorithm was run to cluster the subdistricts based on their most common surrounding venues. Each of the six clusters, labeled 0-5, is characterized by a dominant venue as follows:

A considerable number of coffee shops can be found within Cluster 5 (41 shops out of 151 venues). In fact, it is the second-most common venues in that cluster. Choropleth map of coffee shop locations across mainland Pune shows that Pune City has a very high concentration of the business, i.e., 426 shops while the rest are below 200. The districts in Pune City, therefore, are not viable options for opening up a coffee shop business because they are already way too saturated.

It is recommended that stakeholders look into opportunities in Maval and Haveli, as these two cities have the least concentration of coffee shops and would significantly minimize competition.

Conclusion

Stakeholders searching for opportunities to open a coffee shop in Pune may want to consider setting up their business someplace where competitions are not severe. Pune regions were explored and then clustered based on the similarity of their surrounding venues using *K*-Means algorithm. Analysis results show that districts in Jakarta Utara and Jakarta Timur are among the best candidates for a new coffee shop location.

Reference

United Nations, Population Division (2018). World Urbanization Prospects: The 2018 Revision, Online Edition.

Thank You