THE BATTLE OF THE NEIGHBORHOODS

Finding the best location to open a hotel in Province 2, Nepal

MAY 3, 2020

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

Coursera

Table of Contents

1. Introduction 2
2. Data Description 2
3. Data Preparation 3
4. Methodology 4
5. Results 10
6. Discussion 10
7. Conclusion 10

1. Introduction

For this project, I selected to work with selecting a suitable location for a fictional Hotel in Province Number 2 of Nepal. Below is the business problem description:

The hotel chain "XYZ" is interested in opening a new hotel in Province Number 2. This would be the chain's second hotel location, after having successfully opened a venue in Kathmandu, a very lively neighborhood from Province Number 3. Considering they had very good results with their Kathmandu location, they have requested our data science team to find a neighborhood with similar characteristics. The problem question would be: What neighborhood from Province 2 has the most similar characteristics compared to Kathmandu in Province 3?

Since there are already a number of hotels in Province 2 we will try to detect locations that are not already crowded with hotels. We would also prefer locations with greater population and Per Capita Income, assuming we have found neighborhoods similar to Kathmandu in Province 2. We will use our data science powers to generate a few most promising neighborhoods based on this criterion.

2. Data Description

The data to be used for this project comes from four different locations:

i. Foursquare: It is a local search-and-discovery service which provides information on different types of entertainment, drinking, living and dining venues. Foursquare has an API that can be used to query their database and find information related to allocation, such as the venues, overall category, reviews and tips.

ii. Province 2 and Province 3 Neighborhood Names and geographic coordinates. Available on https://opendatanepal.com/dataset/nepal-municipalities-wise-geographic-data-shp-geojson-topojson-kml. It is a shape file containing the boundary information for all the Neighborhoods of Nepal.

iii. Province 2 and Province 3 Neighborhood Population Information. Data available on https://opendatanepal.com/dataset/total-population-by-sex-country-province-district-and-local-level. It is a csv file with population information for all the neighborhoods of Nepal, divided into male and female population for each neighborhood. We will determine the total population for each neighborhood from it.

iv. Province 2 and Province 3 District Income Information. Data available on https://opendatanepal.com/dataset/life-expectancy-income-of-nepal-by-district. It is a csv file with Per Capita Income (USD) information for all the districts of Nepal. I could not find income information at Neighborhood level. We will assume each districts income represents income of neighborhoods located in the district as well.

*Understand that I am interpreting neighborhoods as local levels in context of Nepal.

3. Data Preparation

On this section, we will read data from the later three of the sources, and will consolidate the data from our sources into new datasets.

Here are the first 10 rows from province three, new consolidated dataframe:

	PROVINCE	DISTRICT	NEIGHBORHOOD_TYPE	NEIGHBORHOOD	LONGITUDE	LATITUDE	POPULATION	PER CAPITA INCOME (USD)
0	3	SINDHULI	Nagarpalika	Dudhouli	86.241117	27.032294	65644	822.0
1	3	SINDHULI	Gaunpalika	Ghanglekh	85.788581	27.343092	13761	822.0
2	3	SINDHULI	Gaunpalika	Golanjor	86.069936	27.251908	19490	822.0
3	3	SINDHULI	Gaunpalika	Hariharpurgadhi	85.543347	27.251372	27727	822.0
4	3	SINDHULI	Nagarpalika	Kamalamai	85.932950	27.187530	66391	822.0
5	3	SINDHULI	Gaunpalika	Marin	85.699617	27.255092	41106	822.0
6	3	SINDHULI	Gaunpalika	Phikkal	86.288506	27.186659	16968	822.0
7	3	SINDHULI	Gaunpalika	Sunkoshi_sindhuli	85.883817	27.379378	21969	822.0
8	3	SINDHULI	Gaunpalika	Tinpatan	86.134532	27.133561	36420	822.0
9	3	RAMECHHAP	Gaunpalika	Doramba	85.923530	27.549137	22773	951.0

And here are the first 10 rows from province two, new consolidated dataframe:

	PROVINCE	DISTRICT	NEIGHBORHOOD_TYPE	NEIGHBORHOOD	LONGITUDE	LATITUDE	POPULATION	PER CAPITA INCOME (USD)
0	2	SAPTARI	Gaunpalika	Agnisair Krishna Savaran	86.796975	26.646790	27129	801.0
1	2	SAPTARI	Gaunpalika	Balan Bihul	86.522056	26.571204	21842	801.0
2	2	SAPTARI	Gaunpalika	Bishnupur_saptari	86.720324	26.518306	23166	801.0
3	2	SAPTARI	Nagarpalika	Bode Barsain	86.573847	26.564128	43293	801.0
4	2	SAPTARI	Gaunpalika	Chhinnamasta	86.734078	26.461687	28437	801.0
5	2	SAPTARI	Nagarpalika	Dakneshwori	86.632882	26.529857	42833	801.0
6	2	SAPTARI	Nagarpalika	Hanumannagar Kankalini	86.878736	26.512161	45840	801.0
7	2	SAPTARI	Nagarpalika	Kanchanrup	86.900719	26.635802	53366	801.0
8	2	SAPTARI	Nagarpalika	Khadak	86.607061	26.640448	45428	801.0
9	2	SAPTARI	Gaunpalika	Mahadeva	86.818177	26.550320	28542	801.0

We need to concat the location information of Province 2 and Province 3 as we will be using unsupervised machine learning algorithm KMeans Clustering, to find neighborhoods from Province 2 which fall in the same cluster as Kathmandu. We won't be needing Population and Per Capita Income for this. Here’s the final dataframe, containing location information on neighborhoods of both province 2 and 3:

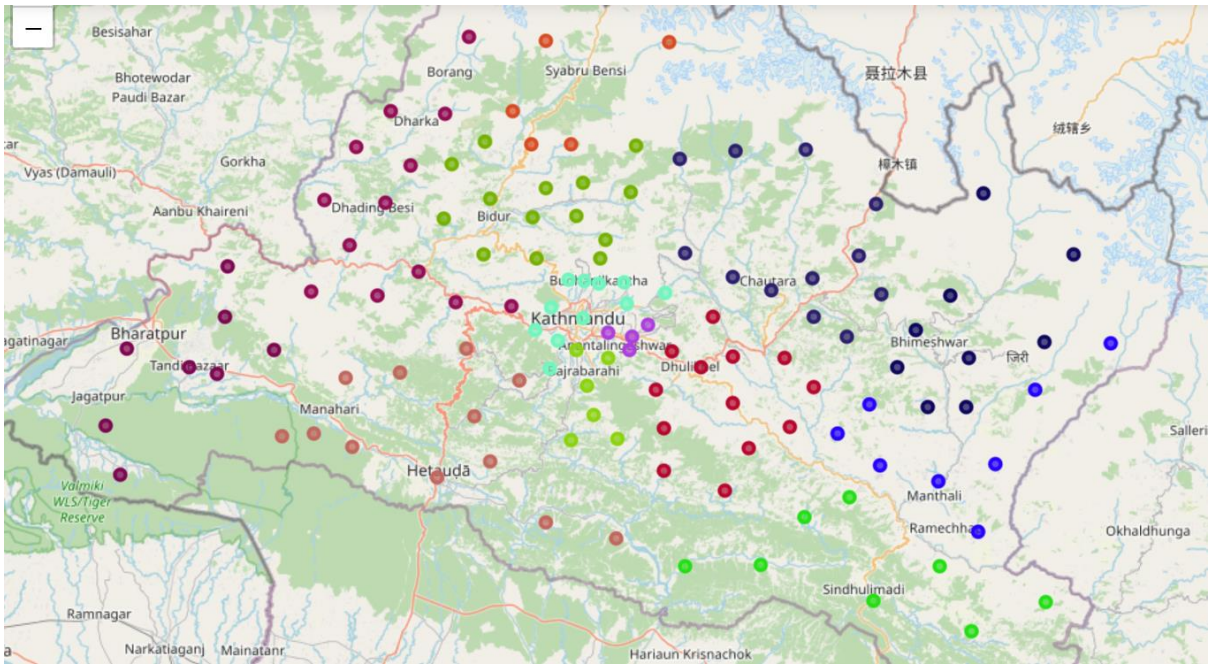
	PROVINCE	DISTRICT	NEIGHBORHOOD_TYPE	NEIGHBORHOOD	LONGITUDE	LATITUDE
0	3	SINDHULI	Nagarpalika	Dudhouli	86.241117	27.032294
1	3	SINDHULI	Gaunpalika	Ghanglekh	85.788581	27.343092
2	3	SINDHULI	Gaunpalika	Golanjor	86.069936	27.251908
3	3	SINDHULI	Gaunpalika	Hariharpurgadhi	85.543347	27.251372
4	3	SINDHULI	Nagarpalika	Kamalamai	85.932950	27.187530
...	...	...	...	...	...	...
257	2	PARSA	Gaunpalika	Paterwa Sugauli	84.792380	27.192946
258	2	PARSA	Nagarpalika	Pokhariya	84.765851	27.076862
259	2	PARSA	Gaunpalika	Sakhuwa Prasauni	84.827764	27.166429
260	2	PARSA	Gaunpalika	Thori	84.636899	27.320525
261	2	PARSA	National Park	Chitawan National Park_parsa	84.782004	27.342900

262 rows × 6 columns

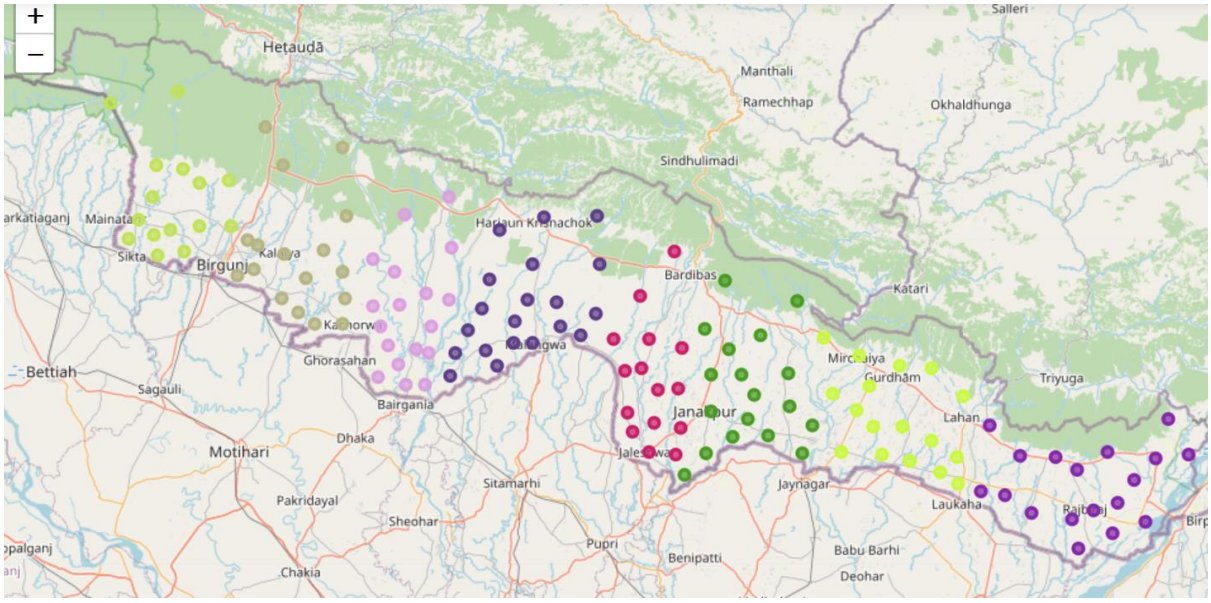
4. Methodology

We will start with visualizing the neighborhoods on respective map of province 2 and 3 with the help of folium maps. Understand that markers of same color, represents neighborhoods in same district.

Here's the map with neighborhoods of province 3:



and here is the map with neighborhoods of province 2:

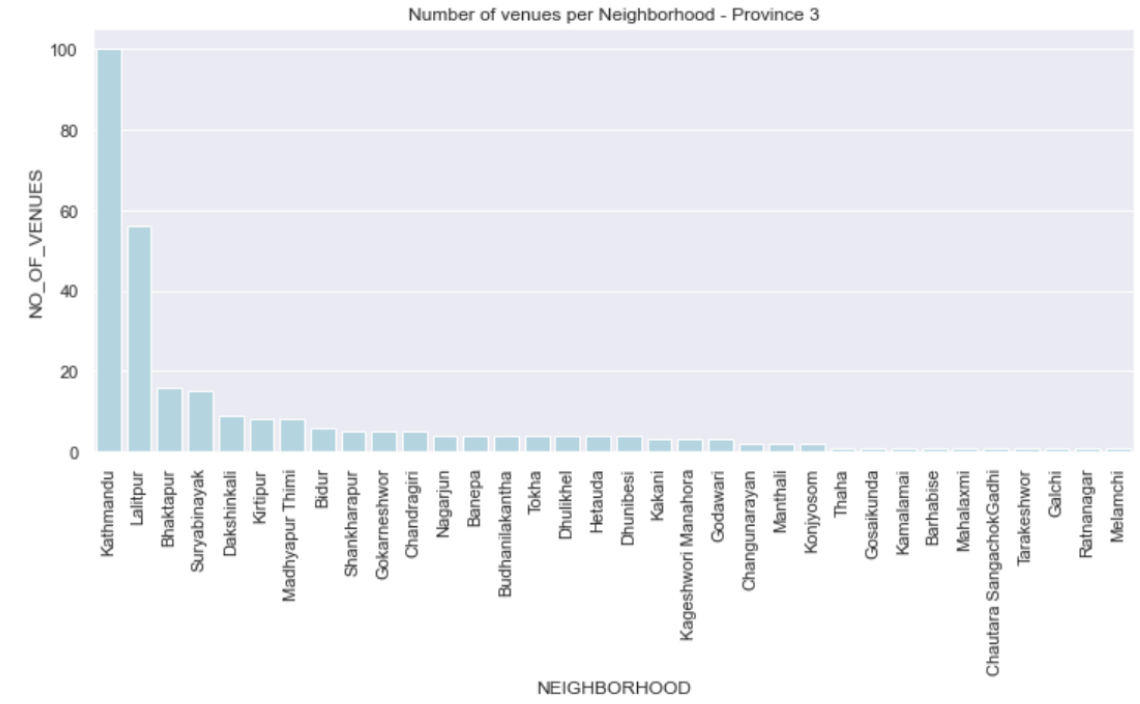


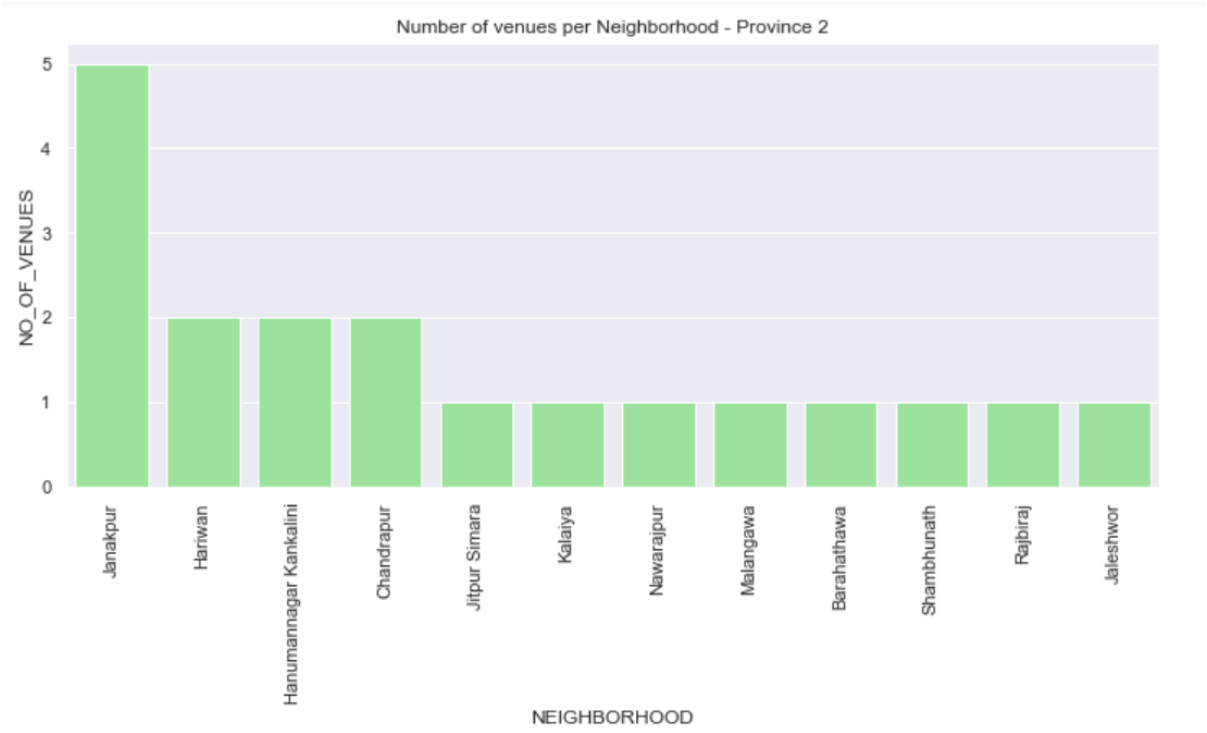
We will now use Foursquare API to explore venues in neighborhoods of Province 2 and 3. The output from the API call will be a json file, from where we will read required information and process it to get output in form of a dataframe, shown below. Understand that each row here, represents a venue. So, if there are 2 venues in the neighborhood we will have 2 rows below.

	PROVINCE	NEIGHBORHOOD	DISTRICT	VENUE	VENUE_LATITUDE	VENUE_LONGITUDE	VENUE_CATEGORY
0	3	Kamalamai	SINDHULI	Sindhulimadi	27.202838	85.911664	Historic Site
1	3	Manthali	RAMECHHAP	Trekkers Transit Café	27.393574	86.061830	Café
2	3	Manthali	RAMECHHAP	Manthali Airport	27.383976	86.059990	Airport
3	3	Barhabise	SINDHUPALCHOK	friends adventure canyoning spot	27.818253	85.877411	River
4	3	Chautara SangachokGadhi	SINDHUPALCHOK	Tudikhel	27.775686	85.712906	Soccer Field
...	...	...	...	...	...	...	...
300	2	Malangawa	SARLAHI	Janaki Family Lodge	26.861152	85.563609	Motel
301	2	Chandrapur	RAUTAHAT	Chandranigapur	27.121268	85.358287	Bus Station
302	2	Chandrapur	RAUTAHAT	Alpine Hotel	27.125422	85.351013	Motel
303	2	Jitpur Simara	BARA	Anmol Travels	27.198462	84.982239	Moving Target
304	2	Kalaiya	BARA	Gadi	27.033601	85.003298	Breakfast Spot

305 rows × 7 columns

Now, Foursquare database has more information on venues from province three than province two. This will lead to biasness in the result. However, for the purpose of this example we will disregard this biasness. We can plot a bar-graph showing how many venues were returned for each neighborhood from API call.





Machine learning models don't understand words and so we will use pandas one-hot encoding to transform the venues in each neighborhood in binary terms with help of their venue categories, where 1 represent the venue category and 0 represents not the venue category. We will then add neighborhood names to it. Understand that each row in dataframe below is a venue and the value 1 points towards its category.

Bar	Basketball Court	Bed & Breakfast	Beer Garden	Bistro	...	Steakhouse	Theme Park Ride / Attraction	Trail	Train Station	Tree	Vacation Rental	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Village	NEIGHBORHOOD
0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Kamalamai
0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Manthali
0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Manthali
0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Barhabise
0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Chautara SangachokGadhi
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Malangawa
0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Chandrapur
0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Chandrapur
0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Jitpur Simara
0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Kalaiya

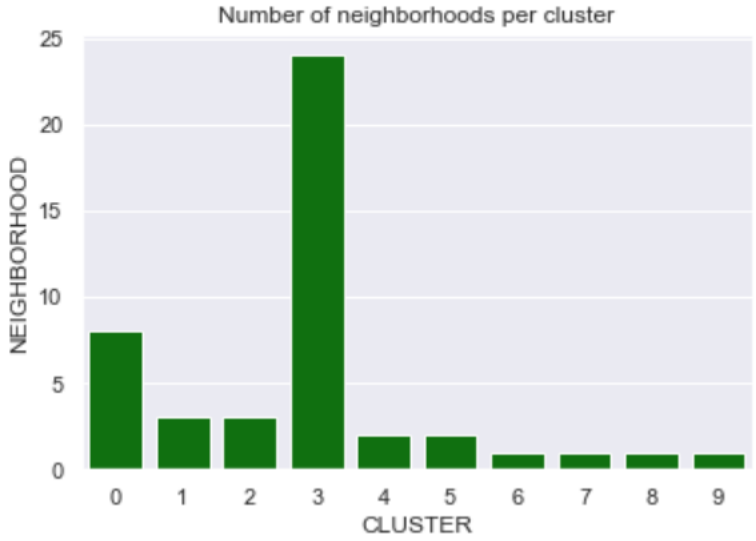
We will use the dataframe above for machine learning specifically KMeans clustering, but before that let us look at most common venue categories in some of the neighborhoods. In order to do that, we will first group above values by neighborhood and take a mean of groupby object, this will give the probability of finding the venue category in that neighborhood, in comparison to other venue categories. If it is difficult to understand let me explain with an example. Suppose there are 2 Asian Restaurants named 'A' and 'B' in a neighborhood and 1 Bakery named C. Our above data frame will have three rows representing each venue, as each row represents a venue, and not a venue category. Now, if we groupby neighborhood and calculate the mean of groupby object, it will calculate mean of Asian Restaurants as (1+1+0)/3 = 2/3 and Bakery as (0+1+0)/3=1/3. So the comparative probability of finding an Asian restaurant in that neighborhood is greater than finding a Bakery. If you still didn't understand it, no worries. You can skip the two tables below. It is just a side-note, not the main idea. The first 10 rows of above computation are shown below.

NEIGHBORHOOD	Airport	Asian Restaurant	Athletics & Sports	BBQ Joint	Bakery	Bar	Basketball Court	Bed & Breakfast	Beer Garden	...	Spanish Restaurant	Steakhouse	Theme Park Ride / Attraction	Trail	Train Station
Banepa	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
Barahathawa	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
Barhabise	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
Bhaktapur	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
Bidur	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
Budhanilakantha	0.0	0.0	0.25	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
Chandragiri	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
Chandrapur	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
Changunarayan	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
Chautara SangachokGadhi	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0

Now, as we have comparative probabilities, we can sort venue categories based on it, to determine most common ones for our neighborhoods, as shown below.

	NEIGHBORHOOD	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Banepa	Café	Fast Food Restaurant	Bus Station	Village	Diner	Eastern European Restaurant	Farm	Fish & Chips Shop	Food	Forest
1	Barahathawa	Bus Stop	Village	Department Store	Dumpling Restaurant	Eastern European Restaurant	Farm	Fast Food Restaurant	Fish & Chips Shop	Food	Forest
2	Barhabise	River	Village	Department Store	Dumpling Restaurant	Eastern European Restaurant	Farm	Fast Food Restaurant	Fish & Chips Shop	Food	Forest
3	Bhaktapur	Hostel	Bus Station	Café	Plaza	Historic Site	Tree	Lake	Restaurant	Himalayan Restaurant	Indian Restaurant
4	Bidur	Indian Restaurant	Scenic Lookout	Vacation Rental	Restaurant	River	Village	Diner	Dumpling Restaurant	Eastern European Restaurant	Farm
5	Budhanilakantha	Resort	Athletics & Sports	Historic Site	Outdoors & Recreation	Village	French Restaurant	Dumpling Restaurant	Eastern European Restaurant	Farm	Fast Food Restaurant
6	Chandragiri	Coffee Shop	Record Shop	Historic Site	Park	French Restaurant	Dumpling Restaurant	Eastern European Restaurant	Farm	Fast Food Restaurant	Fish & Chips Shop
7	Chandrapur	Motel	Bus Station	Diner	Eastern European Restaurant	Farm	Fast Food Restaurant	Fish & Chips Shop	Food	Forest	French Restaurant

Now, we will use machine learning algorithm KMeans clustering. It basically involves four steps, importing the model, initializing the model, fitting the model with data and finally predicting with the help of the model. We can import the model from sklearn.cluster, we defined/initialized KMeans to cluster our neighborhoods into 10 clusters resulting from 300 iterations, which are practical numbers, we will fit the output from one-hot encoding to the model and the model will return 10 cluster values, with similar neighborhoods in one cluster. After the process, most of the neighborhoods clustered into one cluster i.e. cluster 3, as shown below. Cluster 3 is also the cluster where Kathmandu belongs.



We will now see which neighborhoods from province 2 belong to cluster 3. Turn out for this particular run of program, we have 5 neighborhoods from province 2 in cluster 3, as shown below.

	DISTRICT	NEIGHBORHOOD
0	SAPTARI	Hanumannagar Kankalini
1	SARLAHI	Hariwan
2	DHANUSHA	Janakpur
3	BARA	Kalaiya
4	SIRAHA	Nawarajpur

We will further analyze our results to obtain top three based on more filtering criteria i.e. our ranking system. Understand that a number of ranking systems can be developed for same task. This one is just an arbitrary one.

We will rank each neighborhood based on a composite ranking using the following items:

a. Total Population. Weight: 50%

b. Average income per household within each neighborhood. Weight: 35%

c. Amount of already existing Hotel. Weight: 15%

To create this ranking, lets first normalize each of the three metrics. The three metrics are measurement of non-compatible items. Normalizing them, we can add them together to calculate our ranking system with weights provided above.

Also, from here we will be focusing only on province number 2 datasets as our competing neighborhoods are from province number 2 only.

We normalized these parameters and final dataframe, with normalized values, is shown below. Understand that Number of hotels in the data frame is actually in normalized form, and bar of number of hotels is its composite. We will be using bar of number of hotels for our ranking system.

NEIGHBORHOOD_TYPE	LONGITUDE	LATITUDE	POPULATION	PER CAPITA INCOME (USD)	POPULATION_NORMALIZED	INCOME_NORMALIZED	NUMBER OF HOTELS	BAR NUMBER OF HOTELS
Nagarpalika	86.878736	26.512161	45840	801.0	0.187803	0.541216	0.0	1.0
Nagarpalika	85.576516	27.100324	43924	809.0	0.179953	0.546622	0.0	1.0
Upamahanagarpalika	85.936686	26.725416	162842	938.0	0.667150	0.633784	0.4	0.6
Upamahanagarpalika	85.011567	27.028639	123363	1480.0	0.505408	1.000000	0.0	1.0
Gaunpalika	86.435139	26.607590	19056	689.0	0.078071	0.465541	0.0	1.0

5. Results

Applying our ranking system to above dataframe we can obtain our top three neighborhoods, as shown below. The maximum values is 1 and minimum is 0, for ranking.

	DISTRICT	NEIGHBORHOOD	RANKING	LONGITUDE	LATITUDE
0	BARA	Kalaiya	0.752704	85.011567	27.028639
1	DHANUSHA	Janakpur	0.645399	85.936686	26.725416
2	SAPTARI	Hanumannagar Kankalini	0.433327	86.878736	26.512161

6. Discussion

After performing a clustering analysis, a group of possible neighborhoods, with similar characteristics to the target neighborhood from Province 3, were identified. We decided to filter further our results based on the potential customer base (Population), spending power of the population (income), and the number of competitors. We came up with the three most suitable locations based on available data.

It is necessary to address the difference in data available between province two and province 3, and the bias generated because of it. There may be a perfect reason for the small number of hotels in any of those areas, logic which would make them unsuitable for a new restaurant regardless of lack of competition in the area. It is also necessary to point out that, hyper parameters tuning and comparison between alternative models (KMeans, DBSCAN) we not part of this analysis. KMeans, with its defects like having no concept of outliers, and other models with their own shortcomings may not provide full-proof results.

7. Conclusion

The purpose of this analysis was only to provide info on areas similar to Kathmandu in province 2 with the available information, not crowded with existing hotels, and amidst many people with higher spending power. Recommended zones should, therefore, be considered only as a starting point for more detailed analysis, which could eventually result in a location that has not only no nearby competition but also other factors taken into account and all other relevant conditions met. Domain knowledge and field evaluation are inseparable from data analysis.

Final decision on optimal hotel location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood, etc.