Optimizing Where to Establish a New Hotel on the Most Tourist-Visited Philippine Provinces as Analyzed Through Foursquare API Location Data Using Unsupervised Machine Learning Algorithm

Sandy C. Lauguico

July 2020

Coursera Applied Data Science Capstone Project

Image Source: https://www.thecrazytourist.com/25-best-islands-visit-philippines/

Introduction

Having the Department of Tourism's (DOT) slogan, "It's more fun in the Philippines," the country's tourism is one of the growing sectors that significantly contribute in the economy. According to several articles (Erika Hueneke, 2019; Amy Thomas, 2019; Karen Hastings, 2020), the Philippines has some of the most beautiful islands in the world, which attracted majority of the 8,211,535 foreign tourists in different provinces last 2018 as reported by DOT. The Philippines as an archipelago has a lot of geographical and natural settings that can offer attractions, fun activities, leisure, and relaxation to accommodate those who are wandering.

However, having a good vacation is not just about the experiences from those beautiful sceneries. One of the major factors that needs to be considered are the hotels where the guests are going to stay. Hotel experiences can make or break the whole vacation. It is very important that hotels also offer the services which could best accommodate the tourists to enhance and to promote tourism further. Without considering this, a relaxing getaway can turn into a stressful experience which could rate down the country's tourism.

Business Problem

The current construction industry that is working together with the DOT is possibly relying on different inferences on where to best place a new hotel. This may result to potential investment losses and inefficient use of resources. A data-driven analysis would significantly provide a decision that can maximize profit while promoting tourism. The objective of this project is to be able to effectively locate the best provinces where to establish new hotels based on tourism data, existing hotels, and hotel customer-based clustered experiences. This optimal approach may help in decreasing the risks of losses and increasing the possibilities of win-win arrangement for businesses, communities, and customers. The system will also be able to assist the

construction industry in determining what type of hotel (from 1- to 7-star hotels) can be established to offer variety in prices and services.

Data

In achieving the objectives, the following data are necessary to accomplish the analysis:

- List of the Philippine provinces
- The provinces' corresponding coordinates in terms of latitude and longitude
- The total number of tourists visiting a province
- Venues within the provinces, specifically the hotel category.

The list of provinces was obtained from a Wikipedia page through scraping using Beautiful Soup (https://en.wikipedia.org/wiki/List_of_Philippine_provinces_by_population). This dataset lists originally each of the provinces and the population residing in the area. It contains all the 81 provinces and the key cities of each province.

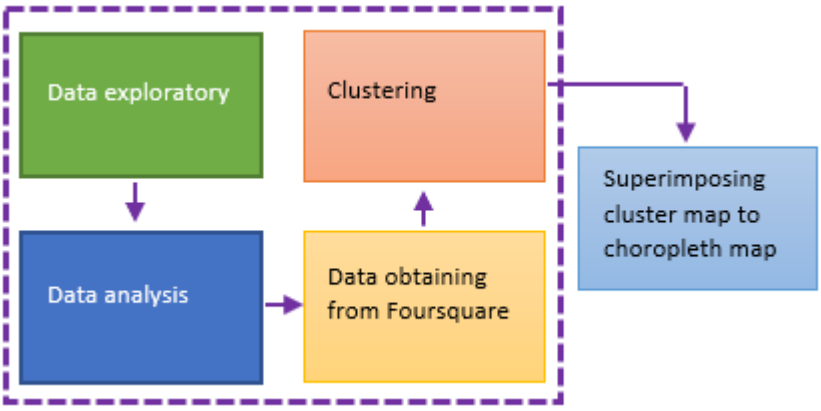
To get the coordinates of the provinces in terms of latitude and longitude, the geocoder library was used. The coordinates were listed and merge into a dataframe with the names of the provinces.

A separate dataset was used to get the total number of tourists from domestic, overseas, and international guests visiting a certain province in 2018. The data was from the Philippines' Department of Tourism and it is recorded and presented in a pdf format. This results to manual integrating of the 'number of tourists' column to the dataframe which consists the provinces and coordinates.

Foursquare API was then used to get the venues in all of the provinces to be extracted in the 'Hotel' venue category.

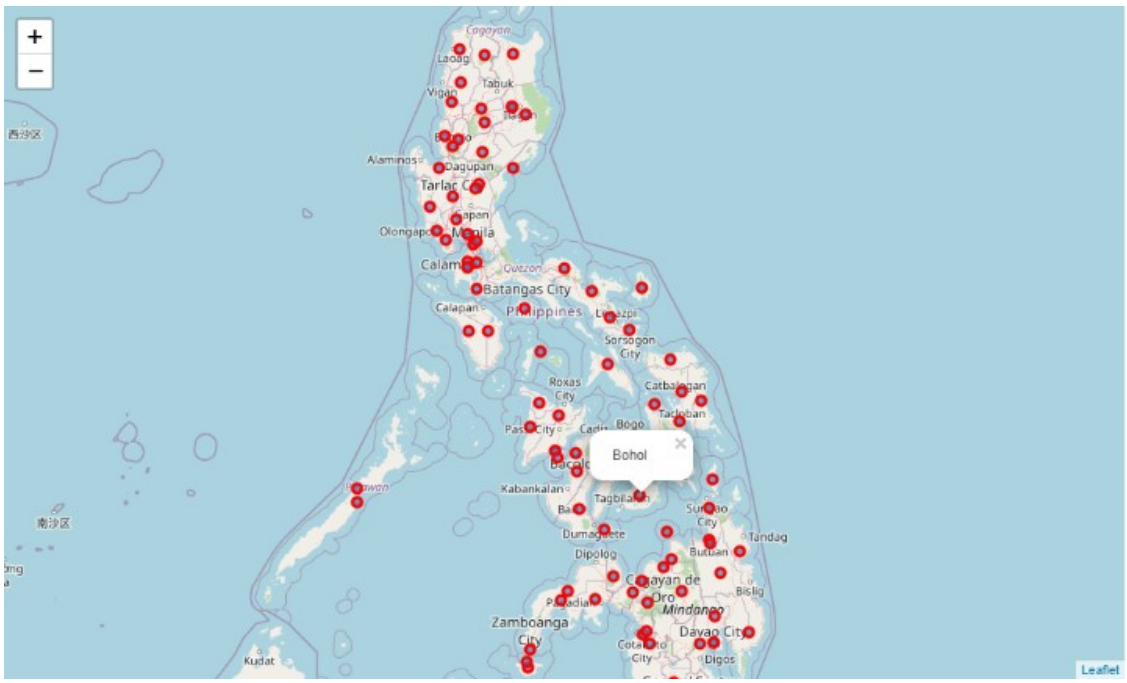
Methodology

The figure shows the methodology followed in conducting the project. Mainly, the steps were divided into 5 and will be thoroughly discussed in the following sections. These steps are data exploratory, where data were preprocessed. Data were also analyzed to determine some pre notions regarding the data to be used and integrated to the data to be obtained from Foursquare. After that, clustering was made, and the cluster map was superimposed to a choropleth map.



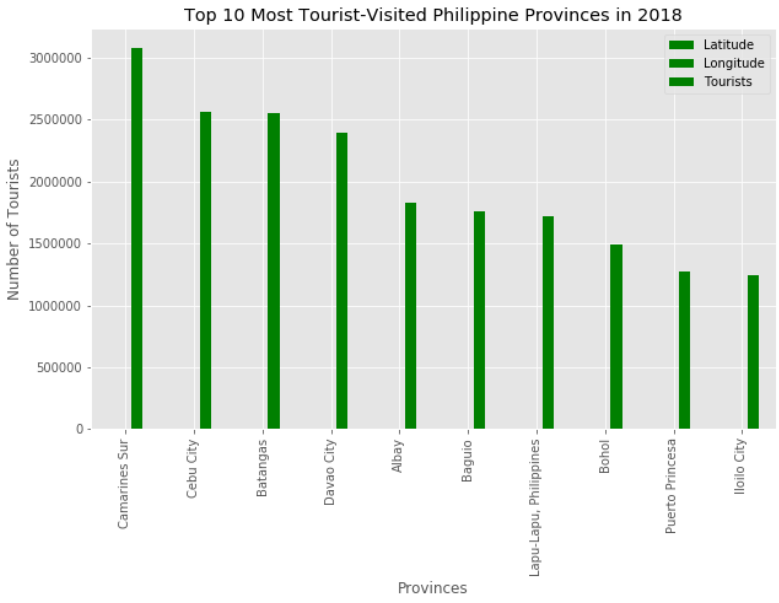
A. Data Exploratory

The list of provincial data from Wikipedia were scraped through the use Beautiful Soup. The class wikitable sortable was stored in a table to find the links that contained the string a. This string contains the values of the Provinces in the column of the table. By using a for loop, a list was created for each of the Philippine province. This was then turned into a dataframe. Geocoder was used to get the coordinates of each province and added to the dataframe which contains the provinces. Rows with NaN values were also dropped in the dataframe for further cleaning. Shown in the figure is the Philippine map with Province markings.

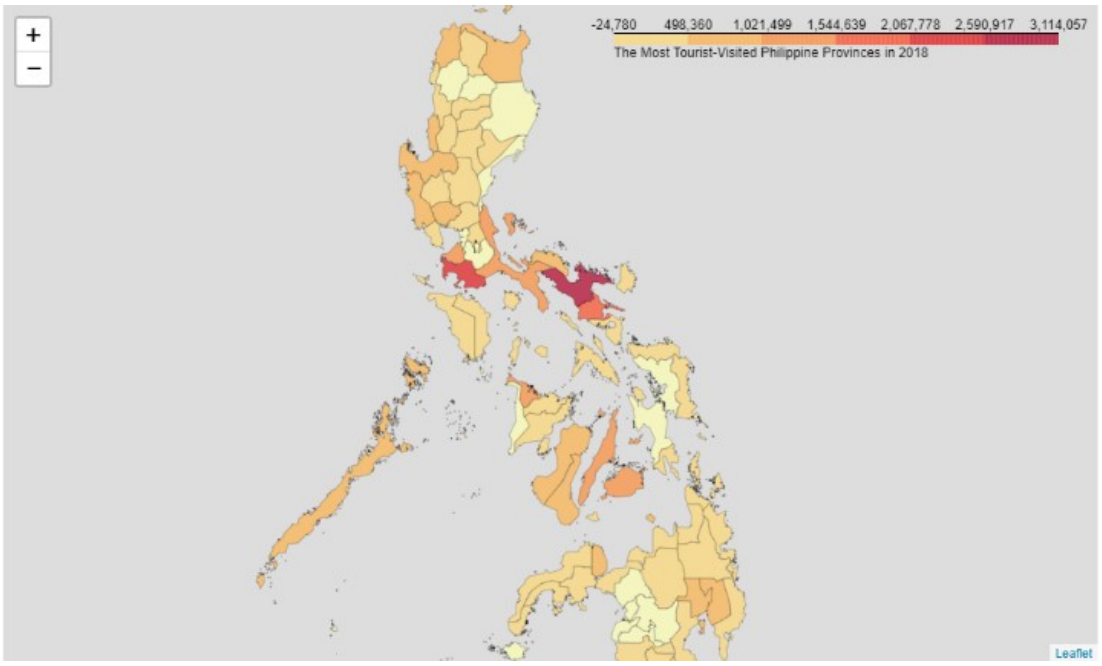


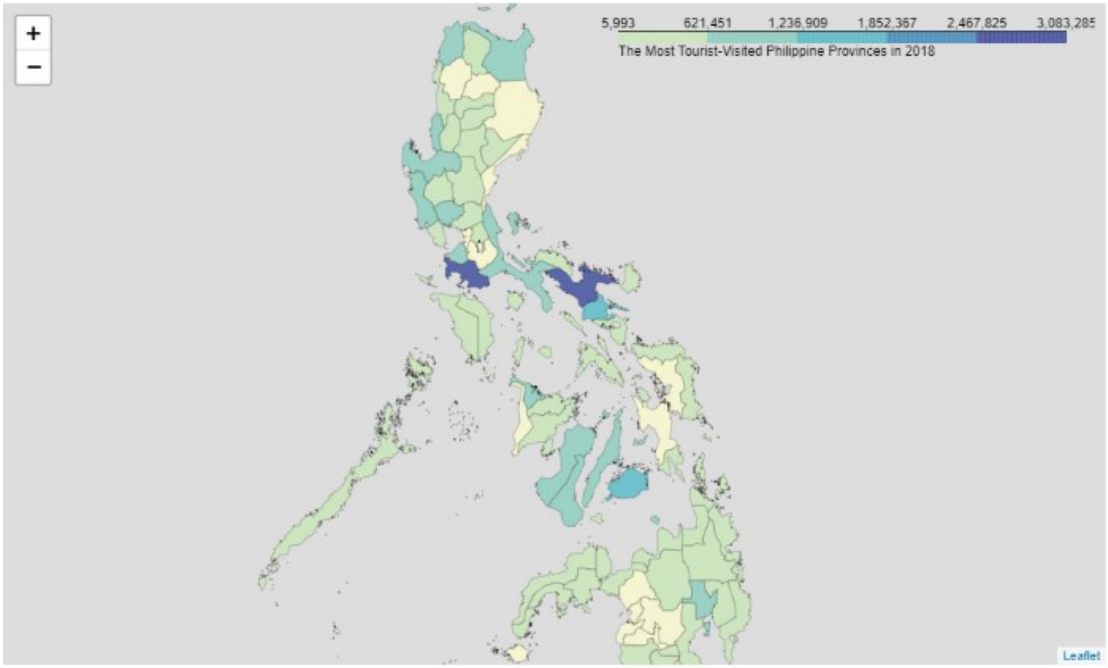
B. Data Analysis

Another set of data was manually obtained and integrated to the dataframe. This new set of data consists the total number of tourists visiting a province. The figure below shows a bar graph for the top 10 most visited province in the Philippines in 2010 as this was extracted from the new data being merged to the original dataframe.



A choropleth map was also generated to represent relatively the number of tourists visiting a province in terms of color intensity. Shown in YlOrRd fill color is the choropleth map. Since there are still negative values in the scale, a numpy array was used to scale the threshold as shown in the YlGnBu color fill.



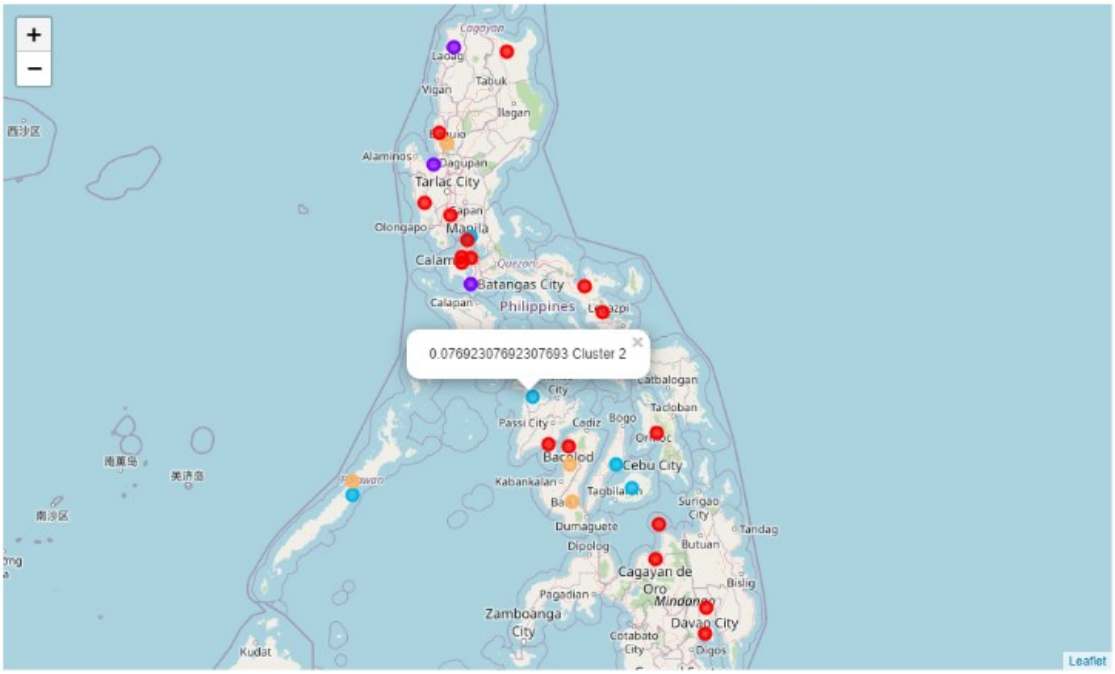


C. Data Obtaining from Foursquare

Foursquare credentials were entered for accessing the data on venues existing in each of the province. A function was defined to to obtain the 100 nearby venues with a 500 m radius. A new dataframe that contains the venues and the venue categories was formed. One hot encoding was performed to represent data in terms of unique venue. The provinces by the mean frequency of the venue category was then stored in a variable to extract the dataframe which contains the provinces and hotel columns.

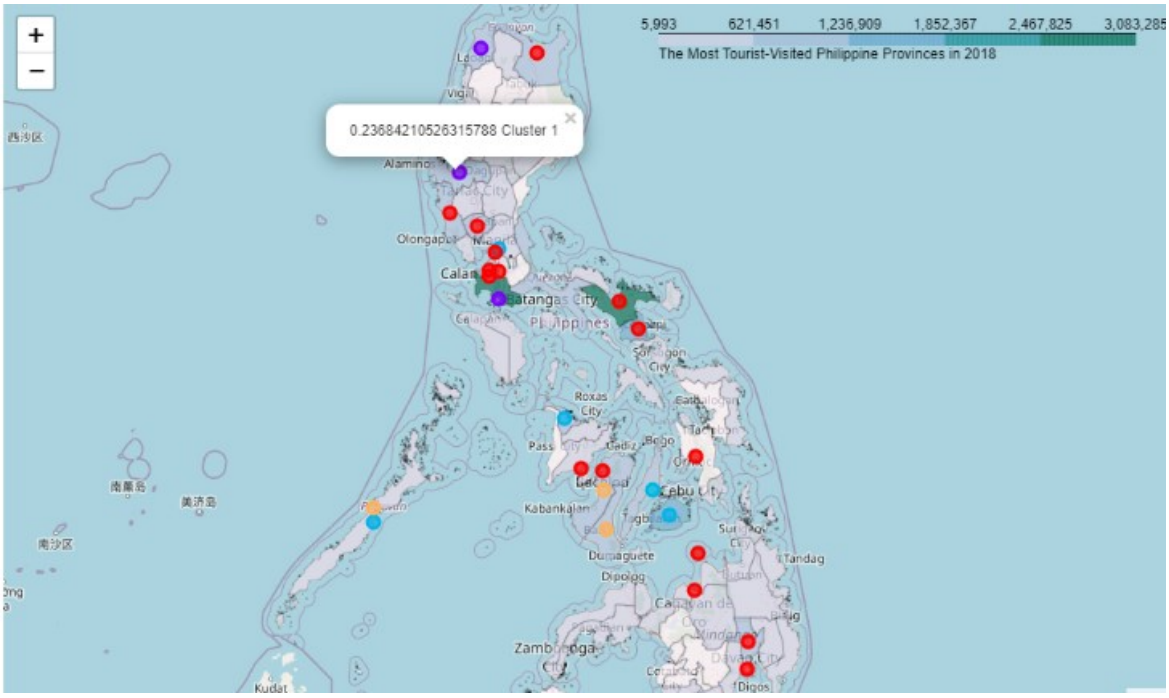
D. Clustering

A total of five clusters were done through K-means Clustering to segment the hotel categories. The cluster labels produced were then merged into the original dataset which contains the provinces, their corresponding coordinates, and the total visiting tourists. The results were then visualized through a cluster map.



E. Superimposing Cluster Map to Choropleth Map

The resulting cluster map was then superimposed to the choropleth map which represents the total number of tourists visiting the province. Thus, it integrates the type of hotels that ware already existing in the provinces.



Results

Evaluating the five clusters provided, the following are the observation for each cluster:

- Cluster 0 (red) – The cluster representing the province with the least number of hotels within a 500 m radius.
- Cluster 1 (violet) – The cluster representing the province with the second to the highest number of hotels within a 500 m radius.
- Cluster 2 (blue) – The cluster representing the province with the third to the highest number of hotels within a 500 m radius.

- Cluster 3 (yellow) – The cluster representing the province with the highest number of hotels within a 500 m radius.
- Cluster 4 (orange) – The cluster representing the province with the second to the least number of hotels within a 500 m radius.

Shown in the following evaluation are the provinces with the hotels belonging to a cluster.

```
df_tourist_merged.loc[df_tourist_merged['Cluster Labels'] == 0]
```

	Provinces	Hotel	Cluster Labels	Tourists	Latitude	Longitude
0	Angeles, Philippines	0.0	0	3083284	13.700560	123.267780
3	Batangas	0.0	0	2393395	7.065740	125.610800
4	Butuan	0.0	0	1829768	13.212181	123.616992
6	Cavite	0.0	0	1716938	14.164863	120.861630
9	Cotabato	0.0	0	1242087	10.705070	122.567850
10	Cotabato City	0.0	0	1116996	14.251965	121.055045
12	Dinagat Islands	0.0	0	1105886	14.254445	120.871101
17	Laguna (province)	0.0	0	876536	8.485850	124.647960
18	Lapu-Lapu, Philippines	0.0	0	876536	18.097458	121.758702
19	Mandaue	0.0	0	835453	10.667960	122.949700
20	Metro Manila	0.0	0	815692	7.568651	125.629228
22	Quezon	0.0	0	767105	9.173147	124.719241
23	Rizal	0.0	0	674359	15.290143	120.141984
25	Siquijor	0.0	0	657267	16.584761	120.424453
26	Sorsogon	0.0	0	647157	6.118870	125.174150
27	Southern Leyte	0.0	0	630899	15.058517	120.644345
28	Tacloban	0.0	0	626168	14.588640	120.984540
29	Tarlac	0.0	0	587659	10.916670	124.666670

```
df_tourist_merged.loc[df_tourist_merged['Cluster Labels'] == 1]
```

	Provinces	Hotel	Cluster Labels	Tourists	Latitude	Longitude
2	Baguio	0.203390	1	2552149	13.762350	121.057070
21	Puerto Princesa	0.236842	1	793890	15.995279	120.313643
24	Romblon	0.250000	1	659242	18.197215	120.728426

```
df_tourist_merged.loc[df_tourist_merged['Cluster Labels'] == 2]
```

	Provinces	Hotel	Cluster Labels	Tourists	Latitude	Longitude
7	Cebu	0.100000	2	1496129	9.851989	124.197340
8	Cebu City	0.100000	2	1278318	9.740010	118.744210
11	Davao City	0.081081	2	1108235	14.647660	121.051500
13	General Santos	0.076923	2	1103334	11.605807	122.251303
14	Iligan	0.066667	2	1093421	10.311210	123.892340

```
df_tourist_merged.loc[df_tourist_merged['Cluster Labels'] == 3]
```

	Provinces	Hotel	Cluster Labels	Tourists	Latitude	Longitude
1	Bacolod	0.384615	3	2559742	10.31121	123.89234

```
df_tourist_merged.loc[df_tourist_merged['Cluster Labels'] == 4]
```

	Provinces	Hotel	Cluster Labels	Tourists	Latitude	Longitude
5	Cagayan de Oro	0.036364	4	1760729	16.413020	120.590760
15	Iloilo	0.043478	4	920242	10.309462	122.986406
16	Iloilo City	0.043478	4	883295	9.604146	123.033766
30	Zamboanga City	0.030303	4	584444	9.981076	118.748465

Discussion

From the results obtained, it can be deduced that the province with the highest number of hotels for the tourists to stay within every 500-meter radius is Bacolod, which is the only province with hotels that belong in Cluster 3. However, looking at the choropleth map, the province has low color intensity. This means that the number of tourists visiting the province is not that high. From the data, it has exactly 835,453 total tourists that visited the province in 2018. This relatively low in comparison to the maximum of 3,083,284 in Camarines Sur, and yet Bacolod has the highest number of hotels. The provinces with least number of hotels within a 500 meter radius are Angeles, Batangas, Butuan, Cavite, Cotabato, Cotabato City, Dinagat Islands, Laguna, Lapu-Lapu, Mandaue, Metro Manila, Quezon, Rizal, Siquijor, Sorsogon, Souther Leyte,

Tacloban, and Tarlac. Two of these listed provinces belonging in Cluster 0 however, belongs to the top 10 most tourist-visited provinces. Thus, this deduces to the need to establish more hotels in the two provinces, specifically: Batangas and Lapu-Lapu, with 2,552,149 and 1,716,938 total tourists respectively in 2018.

To further determine where to establish new hotels, the top 10 most visited provinces will be evaluated with their corresponding hotel clustering.

Ranking	Provinces	Number of Tourists	Hotel Cluster
1	Camarines Sur	3,083,284	Cluster 0
2	Cebu City	2,559,742	Cluster 2
3	Batangas	2,552,149	Cluster 0
4	Davao City	2,393,395	Cluster 2
5	Albay	1,829,768	Cluster 0
6	Baguio	1,760,729	Cluster 1
7	Lapu-Lapu	1,716,938	Cluster 0
8	Bohol	1,496,129	Cluster 2
9	Puerto Princesa	1,278,318	Cluster 1
10	Iloilo City	1,242,087	Cluster 4

Analyzing the table, none of the top 10 most tourist-visited province has a hotel cluster of 3 on which it is the cluster representing the highest number of hotels. Cluster 1, being the second to the highest number of hotels within every 500 meter-radius is only evident in Baguio and Puerto Princesa. This means that somehow these provinces can somehow accommodate the tourists visiting their areas. Cluster 2 is labeled on the top 2, 4, and 8 provinces which were Cebu City, Davao City, and Bohol. This cluster is the third highest frequency of existing hotel which means these provinces on the average can accommodate an acceptable number of tourists. However, Cebu City and Davao City which ranked 2nd and 4th on Cluster 2 needs at least Cluster 1, or better yet Cluster 3 to provide the best accommodation services to the tourists looking for hotels

to stay in. The second to the least number of hotels represented by Cluster 4 is labeled on Iloilo City. This barely can accommodate 1.2 million tourists in the province. At least Cluster 2 or 1 is needed in this province. Lastly, the least number of hotels with Cluster 0 can be found on four of the top 10 most tourist-visited province, which are: Camarines Sur, Batangas, Albay, and Lapu-Lapu which ranked 1, 3, 5, and 7 respectively. With these rankings, these provinces deserve a clustering of at least Cluster 1 with the exception of Camarines Sur which needs Cluster 3, the highest number of hotels possible to exist in within every 500 meter-radius in the province.

Conclusion

From all the evaluation and discussion of results, it can be recommended that 8 out of the top 10 provinces should be added with hotels. These 8 provinces are all except Baguio and Puerto Princesa. To further optimize the analysis, Camarines Sur should be the focus of the construction industry to establish so much more hotels as it has the highest number of tourists having their vacation back in 2018, yet grouped in the cluster with the least number of hotels available, in order to better accommodate all the tourists. The other seven provinces should also be added with more hotels in proportion to their tourism data and the hotel clustering labels they were labeled.