Final Report for Applied Data Science Capstone - Using Machine Learning & Data Science to Open Up a Hotel in The Philippines

Paul Ycay
IBM Data Science Professional Certificate | Coursera

August 16, 2020

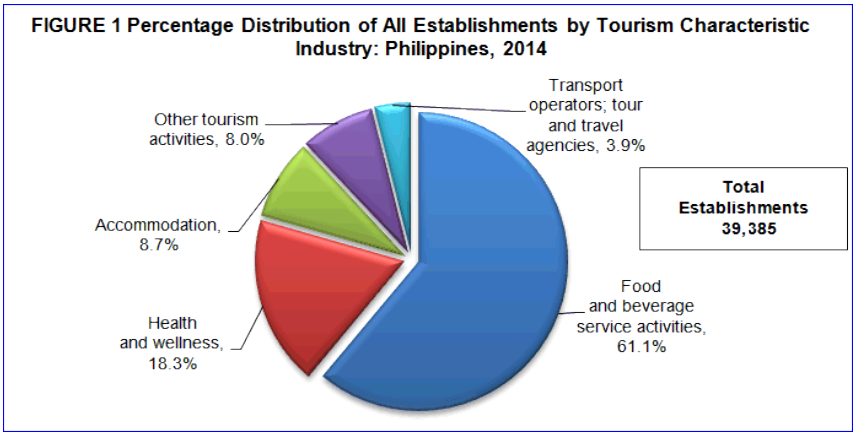
1 Introduction

1.1 Background

The Philippines is made up of 7,641 islands inhabited by a population of 106.7 million, last reported as of 2018. Jane, a Torontonian who is about to complete her degree in hospitality & hotel management, would like to travel back to her home country and open up an. Apart from planning her expenses and sorting out the logistics to operate an inn, Jane must figure out the optimal location to start her business.

For this project, we will use several tools in Data Science & Machine Learning to find the best location that will suit Jane's business needs. Such tools & methods include implementing Exploratory Data Analysis, *Folium* (an interactive data visualization map tool), and the use of the K-Means Clustering algorithm to create a conclusion from location data.

Below is a pie chart representing the tourism activities distributed in the country in 2014. Source: http://www.psa.gov.ph/content/2014-survey-tourism-establishments-philippines-step-economy-wide-preliminary-results



1.2 The Data

The spreadsheet obtained from https://simplemaps.com/data/ph-cities consists of several cities in the Philippines. This dataset will provide necessary information on their coordinate systems. These coordinates will then be used on the Foursquare Places API to gather information on venues, tourist attractions, restaurants, and dense population areas that will compliment an inn.

	city	lat	lng	iso2	admin	population	population_proper
0	Barilan	14.633300	121.000000	PH	Quezon	11100000.0	3077575.0
1	Quezon City	14.648800	121.050900	PH	Quezon	2761720.0	2761720.0
2	Davao	7.073056	125.612778	PH	Davao	1402000.0	1212504.0
3	Cagayan de Oro	8.481111	124.643056	PH	Cagayan de Oro	1121561.0	602088.0
4	General Santos	6.112778	125.171667	PH	General Santos	950530.0	538086.0
5	Bacolod	10.667700	122.953300	PH	Bacolod	949354.0	511820.0
6	Cebu City	10.311111	123.891667	PH	Cebu	815000.0	798634.0
7	Zamboanga City	6.910255	122.071715	PH	Zamboanga	773000.0	457623.0
8	Dinaga	13.650000	123.166667	PH	Camarines Sur	741635.0	174931.0
9	Iligan	8.230833	124.236111	PH	Iligan	464599.0	464599.0
10	Baguio City	16.417155	120.590998	PH	Baguio	447824.0	272714.0

This data set contains geographical information on the various cities in The Philippines, represented by columns 'lat' & 'lng'. The 'admin' column represents the province of that city.

2 Methodology & Results

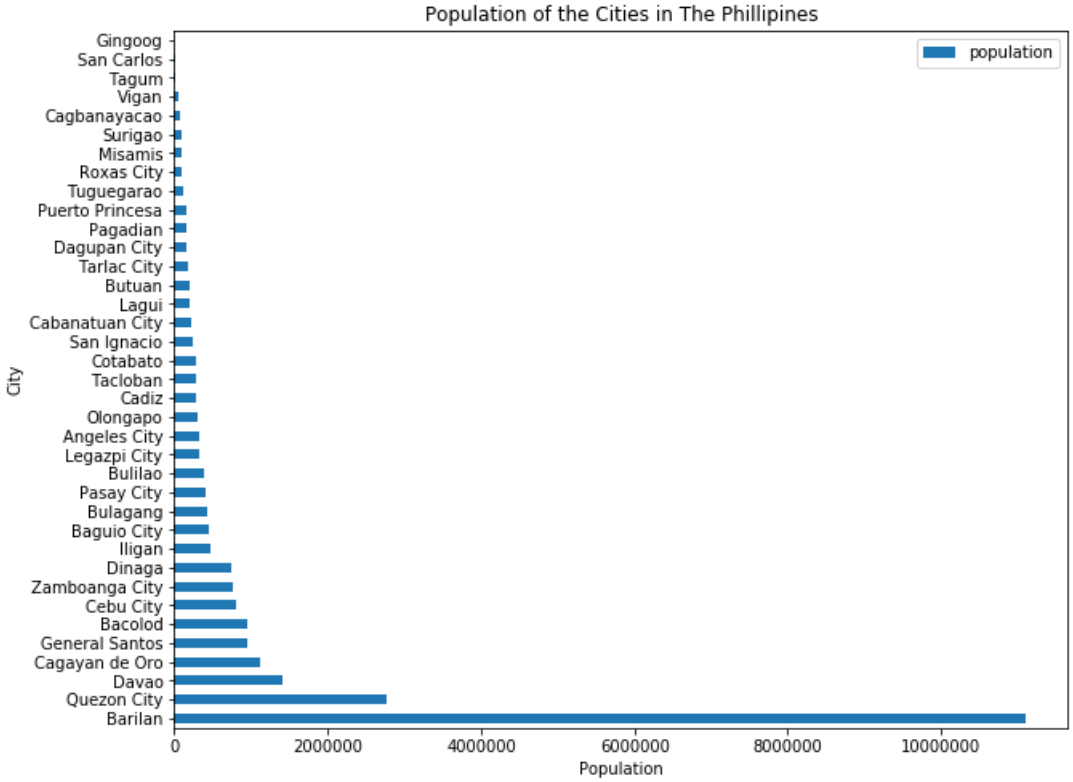
2.1 Analyzing the Data

Upon downloading the .csv file, we import it on the Jupyter Notebook. We remove columns 'country' & 'capital' as it is not needed. We remove locations that have no population data as population will influence where Jane will open up her inn.

Listing 1: Importing Data

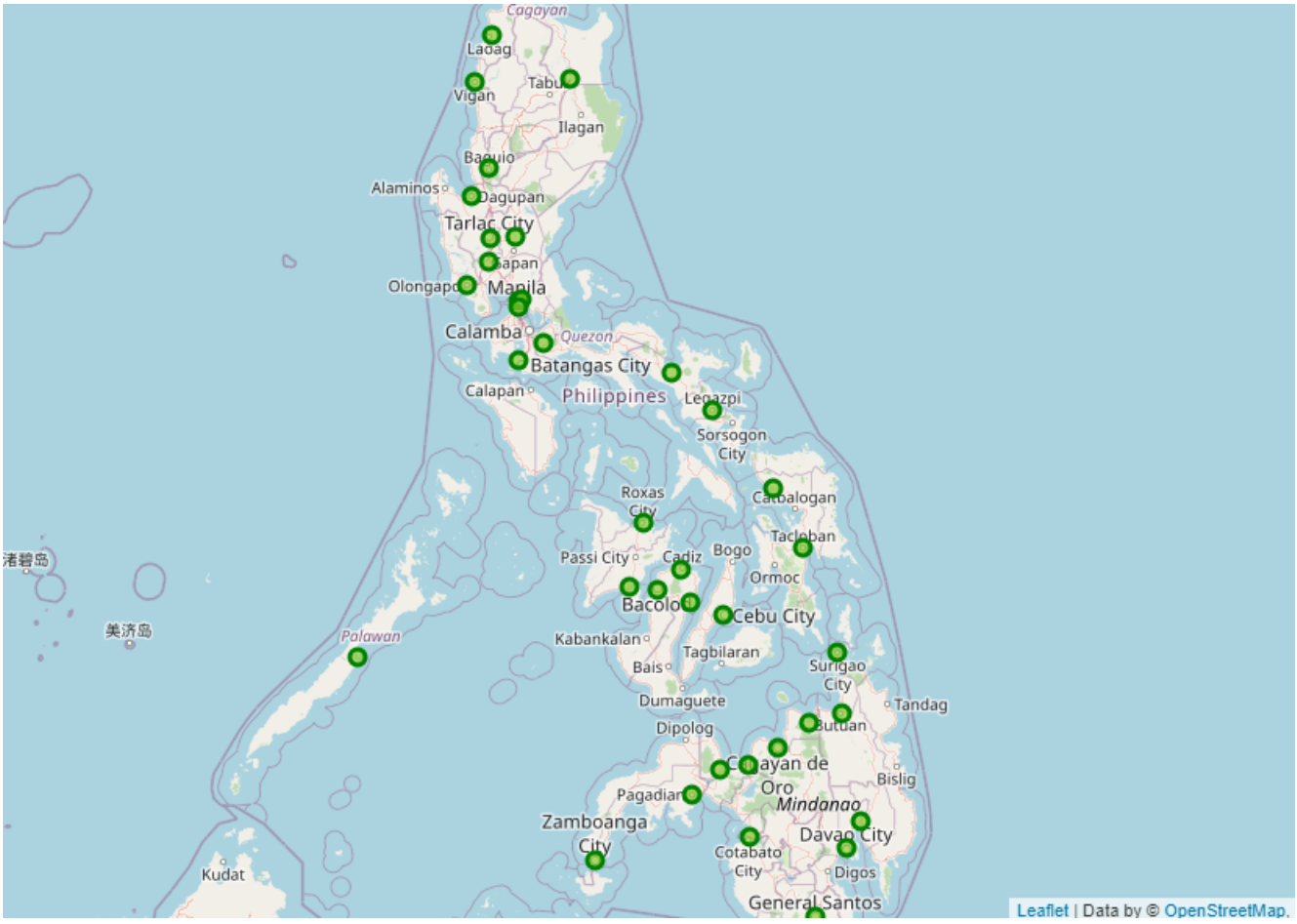
```
ph = pd.read_csv('G:\My_Drive\IBM_DATA_SCIENCE_COURSERA\ph.csv')
ph.sort_values(by = 'population', ascending = False)
ph.drop(columns = ['country','capital'], inplace = True)
# lets drop rows where population is NaN
ph_na = ph[ph['population'].notna()]
ph_na
```

The updated data set includes 37 cities & 35 provinces.



Above is a bar chart of the population of the cities. Clearly, the city of Barilan is the most populated, and may be a location of interest for Jane.

Using the *Folium* library, we display the cities as markers on the country map.



2.2 Initializing Foursquare API credentials

The Foursquare API can be used to explore the districts of The Philippines in more depth. We can obtain valuable information of the venues in the cities, which will be a factor in opening an inn within tourist destinations.

Listing 2: Initializing Foursquare

```
CLIENT_ID = 'XXXXXXXXXXXXXXXXX' # your Foursquare ID
CLIENT_SECRET = 'XXXXXXXXXXXXXXXXX' # your Foursquare Secret
VERSION = '20180604'
LIMIT = 30
print('Your credentails:')
print('CLIENT_ID: ' + CLIENT_ID)
print('CLIENT_SECRET: ' + CLIENT_SECRET)
```

After defining a function to return nearby venues, we create this dataframe consisting of venue categories & name of the venue. There were 119 different venue categories.

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Barilan	14.6333	121.0	Army Navy Burger + Burrito	14.632606	121.001846	Burrito Place
1	Barilan	14.6333	121.0	Tasty Dumplings	14.632874	121.002133	Chinese Restaurant
2	Barilan	14.6333	121.0	Kimpo Tea House	14.634365	121.001526	Chinese Restaurant
3	Barilan	14.6333	121.0	Trà Vinh	14.631662	120.999463	Vietnamese Restaurant
4	Barilan	14.6333	121.0	Café Monaco	14.633123	121.001650	Korean Restaurant

2.3 Modifying Data for K-means Clustering

As our venues are categorical, we assign dummy variables to them in preparation for the K-means clustering algorithm.

```
ph_onehot = pd.get_dummies(philipinnes_venues[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
ph_onehot['City'] = philipinnes_venues['City']

# move neighborhood column to the first column
fixed_columns = [ph_onehot.columns[-1]] + list(ph_onehot.columns[:-1])
ph_onehot = ph_onehot[fixed_columns]
ph_onehot.head()
```

	City	African Restaurant	American Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Badminton Court	Bakery	...	Tea Room	Theme Park Ride / Attraction	Track Stadium	Turkish Restaurant	Ur
0	Barilan	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
1	Barilan	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
2	Barilan	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
3	Barilan	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
4	Barilan	0	0	0	0	0	0	0	0	0	...	0	0	0	0	

5 rows × 120 columns

We then obtain the mean frequency of the venues in each city. The below image displays the top 5 venues in each city. We have that category BBQ Joint is the most common venue in Angeles City, which accounts to 29% of all venues in that city.

```
# determine the frequency in each city
# source: IBM Applied Data Science Capstone Lab Week 3
# Segmenting and Clustering Neighborhoods in New York City

num_top_venues = 5

for hood in ph_grouped['City']:
    print("----"+hood+"----")
    temp = ph_grouped[ph_grouped['City'] == hood].T.reset_index()
    temp.columns = ['venue','freq']
    temp = temp.iloc[1:]
    temp['freq'] = temp['freq'].astype(float)
    temp = temp.round({'freq': 2})
    print(temp.sort_values('freq', ascending=False).reset_index(drop=True).head(num_top_venues))
    print('\n')
```

```
----Angeles City----
                 venue  freq
0            BBQ Joint  0.29
1             Pharmacy  0.14
2  Fast Food Restaurant  0.14
3         Grocery Store  0.14
4    Chinese Restaurant  0.14
```

We then create a function that returns a dataframe of the top 10 venues of each city. This will be useful in handling K-means. The following is produced.

	City	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	Angeles City	BBQ Joint	Pharmacy	Grocery Store	Asian Restaurant	Chinese Restaurant	Fast Food Restaurant	Diner	Empanada Restaurant	Electronics Store	Dumpling Restaurant
1	Bacolod	Gym	Food & Drink Shop	Chinese Restaurant	Deli / Bodega	Restaurant	Middle Eastern Restaurant	Fast Food Restaurant	Grocery Store	Hotel	Fried Chicken Joint
2	Baguio City	Fast Food Restaurant	Convenience Store	Hotel	Flea Market	Campground	Restaurant	Café	Steakhouse	Coffee Shop	Fountain
3	Barilan	Chinese Restaurant	Bubble Tea Shop	Coffee Shop	Spa	Grocery Store	Food & Drink Shop	Dumpling Restaurant	Japanese Restaurant	Café	Pizza Place
4	Bulagang	Chinese Restaurant	Restaurant	Beach	Yoga Studio	Fast Food Restaurant	Cosmetics Shop	Cupcake Shop	Deli / Bodega	Department Store	Dessert Shop

2.4 K-means Clustering

We group the cities in to 5 clusters. We include the following code to remove NaN clusters and then converting to int as it is needed for plotting purposes.

Listing 3: Fixing variable type for cluster plotting

```
# remove NaN cluster labels & converting to int to initialize plotting
ph_merged = ph_merged.dropna(axis='rows')
ph_merged['Cluster Labels'] = ph_merged['Cluster Labels'].astype(int)
ph_merged.head()
```

We now visualize our clusters using the following code:

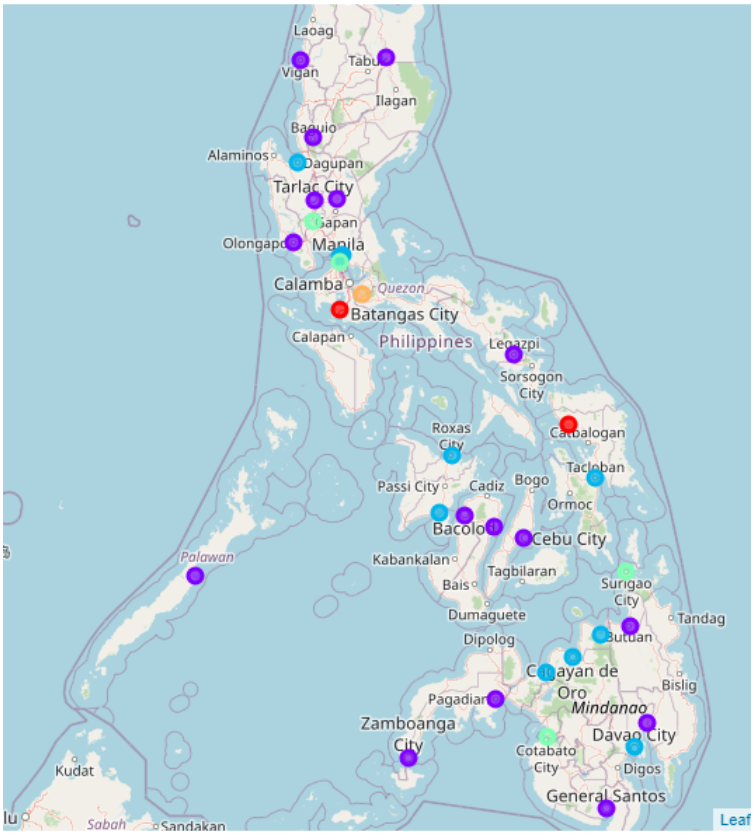
Listing 4: Visualize clusters

```
# create map
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=5)

# set color scheme for the clusters
x = np.arange(kclusters)
ys = [i + x + (i*x)**2 for i in range(kclusters)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lon, poi, cluster in zip(ph_merged['lat'], ph_merged['lng'], ph_merged['city'], pl
    label = folium.Popup(str(poi) + ' Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        color=rainbow[int(cluster)-1],
        fill=True,
        fill_color=rainbow[int(cluster)-1],
        fill_opacity=0.7).add_to(map_clusters)

map_clusters
```



The first cluster, the red markers, include only 2 cities. It has various venue categories that will compliment an inn. Hotels are not popular between both cities.

city	admin	population	Cluster Labels	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
Bulagang	Batangas	424508.0	0	Chinese Restaurant	Restaurant	Beach	Yoga Studio	Fast Food Restaurant	Cosmetics Shop	Cupcake Shop	Deli / Bodega	Department Store	Dessert Shop
Cagbanayacao	Samar	67921.0	0	Chinese Restaurant	Restaurant	Spa	Yoga Studio	Empanada Restaurant	Convenience Store	Cosmetics Shop	Cupcake Shop	Deli / Bodega	Department Store

The second cluster, the purple markers, has a good amount of cities. The top 3 most populated cities in this cluster have a variety of restaurants, but the $1st$ and $3rd$ most populated have hotels among their top categories. Analyzing this dataframe further, it seems that hotels are popular among other cities

city	admin	population	Cluster Labels	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
General Santos	General Santos	950530.0	1	Convenience Store	Hotel	Plaza	Coffee Shop	Cupcake Shop	Asian Restaurant	Fast Food Restaurant	Restaurant	Diner
Bacolod	Bacolod	949354.0	1	Gym	Food & Drink Shop	Chinese Restaurant	Deli / Bodega	Restaurant	Middle Eastern Restaurant	Fast Food Restaurant	Grocery Store	Hotel
Cebu City	Cebu	815000.0	1	Coffee Shop	Spa	Hotel	Pizza Place	Seafood Restaurant	BBQ Joint	New American Restaurant	Mexican Restaurant	Massage Studio
Zamboanga City	Zamboanga	773000.0	1	Hotel	Fast Food Restaurant	Convenience Store	Mobile Phone Shop	Department Store	Café	Burrito Place	Seafood Restaurant	Burger Joint
Baguio City	Baguio	447824.0	1	Fast Food Restaurant	Convenience Store	Hotel	Flea Market	Campground	Restaurant	Café	Steakhouse	Coffee Shop
Legazpi City	Albay	320081.0	1	Hotel	Filipino Restaurant	Diner	Restaurant	Bakery	Fast Food Restaurant	Park	Stadium	Cocktail Bar
Olongapo	Olongapo	304388.0	1	Convenience Store	Diner	Coffee Shop	Grocery Store	Fast Food Restaurant	Bakery	Restaurant	Supermarket	Department Store
Cabanatuan City	Nueva Ecija	220250.0	1	Fast Food Restaurant	Asian Restaurant	Bubble Tea Shop	Breakfast Spot	Food	Hotel	Flea Market	Filipino Restaurant	Bookstore
Butuan	Butuan	190557.0	1	Hotel	Pharmacy	Pizza Place	Fast Food Restaurant	Chinese Restaurant	Convenience Store	Donut Shop	Market	Karaoke Bar
Tarlac City	Tarlac	183930.0	1	Fast Food Restaurant	Pizza Place	Filipino Restaurant	Café	Spa	Bookstore	Lounge	Shopping Mall	Racetrack
Pagadian	Zamboanga del Sur	159590.0	1	Fast Food Restaurant	Hotel	Tea Room	Pizza Place	Coffee Shop	Asian Restaurant	Fried Chicken Joint	Electronics Store	Karaoke Bar
Puerto Princesa	Puerto Princesa	157144.0	1	Pizza Place	Hostel	Hotel	Bed & Breakfast	Dessert Shop	Yoga Studio	Donut Shop	Empanada Restaurant	Electronics Store

The third cluster, the blue markers, contains the 2 most populated cities in the Philippines. These cities have a good selection of restaurant categories and entertainment venues. Hotels are not popular among these cities.

city	admin	population	Cluster Labels	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
Barilan	Quezon	11100000.0	2	Chinese Restaurant	Bubble Tea Shop	Coffee Shop	Spa	Grocery Store	Food & Drink Shop	Dumpling Restaurant	Japanese Restaurant	Café	Pizza Place
Quezon City	Quezon	2761720.0	2	Filipino Restaurant	BBQ Joint	Coffee Shop	Japanese Restaurant	Karaoke Bar	Gym	Diner	Donut Shop	Fast Food Restaurant	Spa
Davao	Davao	1402000.0	2	Café	Chinese Restaurant	Coffee Shop	Bar	Hotel	Spa	Convenience Store	Dessert Shop	Diner	Park
Cagayan de Oro	Cagayan de Oro	1121561.0	2	Restaurant	Spa	Filipino Restaurant	Fast Food Restaurant	Asian Restaurant	Turkish Restaurant	BBQ Joint	Café	Bakery	Coffee Shop
Iligan	Iligan	464599.0	2	Café	Convenience Store	Pharmacy	Food & Drink Shop	Clothing Store	Chinese Restaurant	Plaza	Diner	Donut Shop	Bubble Tea Shop
Bulilao	Iloilo	387681.0	2	BBQ Joint	Filipino Restaurant	Coffee Shop	Café	Tea Room	Soup Place	Chinese Restaurant	Gastropub	Hotel	Donut Shop
Tacloban	Tacloban	280006.0	2	Asian Restaurant	Café	Hotel	Cupcake Shop	Bubble Tea Shop	Italian Restaurant	Diner	Coffee Shop	Mexican Restaurant	Donut Shop
Dagupan City	Dagupan	163676.0	2	Chinese Restaurant	Café	Fast Food Restaurant	Spa	Bus Station	Coffee Shop	Convenience Store	Plaza	Pizza Place	High School
Roxas City	Capiz	102688.0	2	Café	Chinese Restaurant	Dessert Shop	Coffee Shop	Department Store	Restaurant	Spa	Fast Food Restaurant	Beer Garden	Pub
Gingoog	Misamis Oriental	218.0	2	Harbor / Marina	BBQ Joint	Restaurant	Filipino Restaurant	Beach	Speakeasy	Frozen Yogurt Shop	Gas Station	Cosmetics Shop	Cupcake Shop

The fourth and fifth clusters, cyan and orange markers respectively, are similar to the first cluster. The cities in these clusters don't have the highest populations in the country, but have a wide selection of venues.

city	admin	population	Cluster Labels	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
Pasay City	Pasay	403064.0	3	Fast Food Restaurant	Shopping Mall	Chinese Restaurant	Basketball Stadium	Deli / Bodega	Plaza	Breakfast Spot	Stadium	Intersection	Convenience Store
Angeles City	Angeles	314493.0	3	BBQ Joint	Pharmacy	Grocery Store	Asian Restaurant	Chinese Restaurant	Fast Food Restaurant	Diner	Empanada Restaurant	Electronics Store	Dumpling Restaurant
Cotabato	Cotabato	279519.0	3	Fast Food Restaurant	Chinese Restaurant	Pizza Place	Shopping Mall	BBQ Joint	French Restaurant	Arts & Crafts Store	Park	Cosmetics Shop	Cupcake Shop
Surigao	Surigao del Norte	87832.0	3	Fast Food Restaurant	Spa	Athletics & Sports	Seafood Restaurant	Coffee Shop	Convenience Store	Cosmetics Shop	Cupcake Shop	Deli / Bodega	Department Store

city	admin	population	Cluster Labels	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
San Ignacio	Laguna	240830.0	4	Farm	Filipino Restaurant	Convenience Store	Fast Food Restaurant	Cosmetics Shop	Cupcake Shop	Deli / Bodega	Department Store	Dessert Shop	Diner

3 Discussion and Conclusion

3.1 Where Should Jane Open an Inn?

This clustering was based on Foursquare API. In the first cluster, the most common venue was Chinese Restaurant. In the second cluster, Fast Food Restaurant was most common. The third cluster, a mix of food venues were most popular. The fourth cluster, it was Fast Food Restaurant. The fifth cluster was a Farm.

Automatically, we eliminate the fifth cluster with the only city, as tourists don't really visit a country for a farm. The cities in the second cluster places hotels as popular venues. Jane wouldn't want other hotels to compete with her inn business, especially if they are common in these cities. The third cluster includes the two most populated cities in the country, Barilan and Quezon City. They both are in the same province of Quezon. Looking at the venues between the two, it seems that hotels are not common venues, as the most popular attractions are restaurants. Jane should travel to Barilan in the province of Quezon to open an inn. This city is the most populated city in The Philippines, with plenty of venues that will draw in tourists. With the amount of tourists eating at these restaurants, visiting spas, and working out at the gyms, Jane's inn business in the province of Quezon will definitely be profitable in the long run.

3.2 Conclusion and Recommendations

Data Science and Machine Learning helped us recommend the optimal location for Jane to open an inn business in the Philipinnes. After importing data based on cities in The Philippines, applying exploratory data analysis using *Folium*, extracting venue information using Foursquare, and applying K-means clustering, we concluded that the city of Barilan in the province of Quezon was the best location for Jane.

Some recommendations for a more accurate report would be importing data on the number of tourists visiting the Philippines each month, year. Since hotels and inns are more oriented for people visiting from other cities and countries, including data on the amount of people visiting a specific location would be beneficial in deciding where to open an inn business. The quality of life in each city would also be helpful in this study; people pick their travel destinations not only on the attractions a location offers, but based on the living conditions, security, and costs.