Top places-to-go in Taiwan

Introduction

'Isla Formosa' – Taiwan, is a beautiful island that is worth to discover. There are so many attractions to explore, from incredible natural scenery, lively traditions and cultures to delicious food. Tourists can have variety options to choose what places to visit based on their interest. Given there are numerous things tourists can do, my objective for this project is to cluster the venues in Taipei city into different segments which can help tourists to identify where to go based on their interests.

Data

The data I am going to use are Foursquare data for Taipei city. I am going to use Foursquare API to pull the venues data in a json file then clean-up the data into pandas data frame as shown below:

	name	categories	lat	lng
0	白石58莊園	Farm	25.106392	121.594388
1	Drift Wood Cafe Da	Café	25.104982	121.591988
2	龍船岩(石船)	Mountain	25.100527	121.597000

I am also going to use Taipei city district geo location data for the latitude and longitude so we can merge with foursquare venues data for analysis. Below is an example of the geo location data:

	District	City	District (Chinese)	Latitude	Longitude
0	Zhongzheng	Taipei City	中正區	25.032400	121.518000
1	Datong	Taipei City	大同區	25.065986	121.515514
2	Zhongshan	Taipei City	中山區	25.064361	121.533468
3	Songshan	Taipei City	松山區	25.049698	121.577206
4	Daan	Taipei City	大安區	25.026389	121.534444

Then I merged two datasets into one data frame as shown below. This will be my analysis dataset to do the clustering for finding similar places for each district.

	District	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Zhongzheng	Kinfen Braised Pork Rice (金峰魯肉飯)	25.032194	121.518534	Taiwanese Restaurant
1	Zhongzheng	虎記商行	25.031744	121.519284	Café
2	Zhongzheng	樂田麵包屋 Gakuden Boulangerie	25.032757	121.517534	Bakery
3	Zhongzheng	National Theater (國家戲劇院)	25.035197	121.518188	Theater
4	Zhongzheng	豆味行甜不辣、豆花、芋圓	25.031303	121.517232	Snack Place

Methodology

- Exploratory Analysis:
  Before we apply any machine learning method, we explore the dataset so we can better understand the data which can help us to decide what model to use later on. When looking at the descriptive statistics, there are 39 districts in Taipei city and Shilin, Xinyi and Tamsui are the

top districts with most venues in it. This indicates those districts could be most attractive for tourists who like to explore variety of venues.

Top 5 Districts with most venues

	District
Shilin	64
Xinyi	56
Tamsui	43
Banqiao	36
Zhongshan	30

To further look into what's is the top venue categories among each district, we can see that Xinyi district is pretty different than Tamsui and Shilin districts in which there are more fashion or consumer goods related stores in Xinyi district where as Tamsui and Shilin have more restaurants.

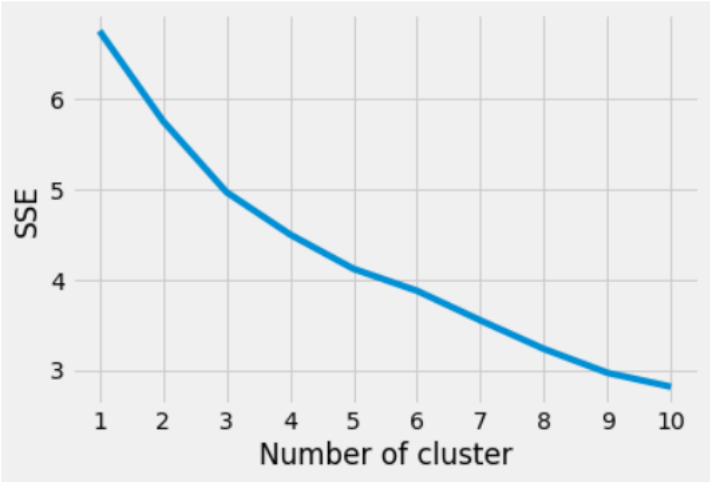
Top 3 categories for top district

	District	Venue Category	Venue
73	Xinyi	Department Store	5
76	Xinyi	Electronics Store	3
85	Xinyi	Lounge	3
62	Tamsui	Taiwanese Restaurant	7
43	Tamsui	Chinese Restaurant	4
45	Tamsui	Coffee Shop	3
5	Shilin	Café	7
34	Shilin	Taiwanese Restaurant	5
3	Shilin	Breakfast Spot	3

Then we expand the exploration to all districts to find out what's the top category among each district. When we look at the result, café, restaurant and bakery are the top categories that show up in the result. One of the most famous and best known thing about Taiwan is the delicious food. Therefore, the result supports the perception and it is expected that the top category shows up are most related with food.

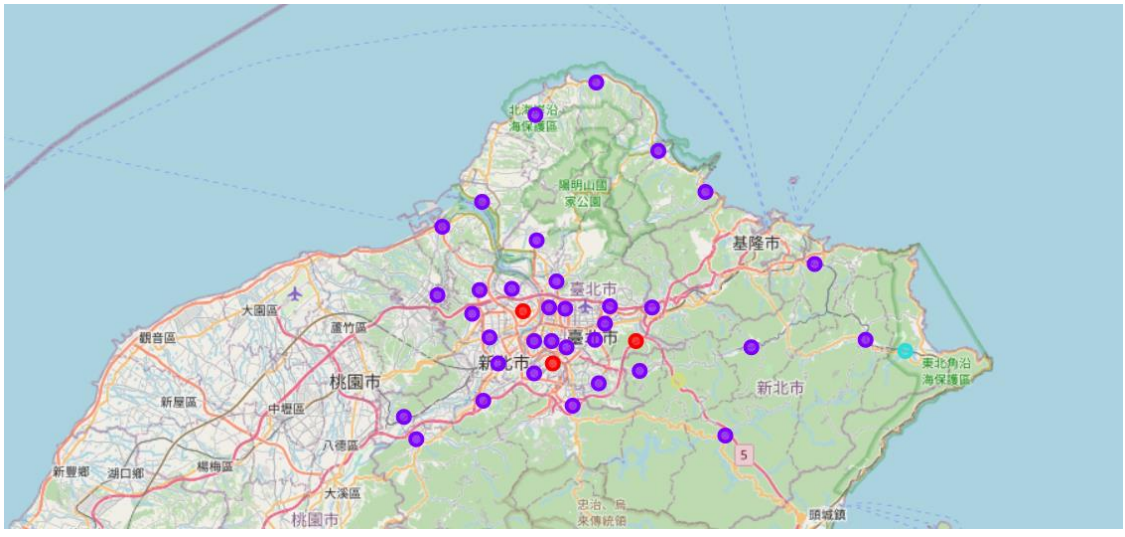
- Machine Learning:
  Now that we have an idea about the district and the most popular categories, we are ready to cluster the districts so it would be easier for tourists to know where to go based on their interested. I am going to use K-mean clustering technique to perform this analysis.

  Before I run the model, I run a list of k from 1 to 10 and use the elbow method on the SSE values generated from each model to decide the optimal point of K. Below chart plot the SSE for each cluster, since it's a little bit hard to determine the elbow point, we use the function to programmatically find out the best K = 4. Therefore, we are segmenting the 39 districts into 4 clusters.



Results

The map shows the cluster for each district. We can further analyze the result to examine each cluster and determine what characteristics and common theme within each cluster so we can introduce tourists on the places to visit based on their interest.



Cluster 1 (Residential Area): There are 3 districts in cluster 1. The common theme for the cluster is convenience store or supermarket. This area may not fit tourists need but more for local residents.

	District	District (Chinese)	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
10	Nangang	南港區	0	Convenience Store	Supermarket	Water Park	Electronics Store	Food & Drink Shop
25	Yonghe	永和區	0	Convenience Store	Breakfast Spot	Coffee Shop	Noodle House	Fast Food Restaurant
31	Sanchong	三重區	0	Convenience Store	Italian Restaurant	Restaurant	Water Park	Electronics Store

Cluster 2 (Food Explorer): This is the largest cluster we have; there are 31 districts belong to this cluster. When looking at a sample of the districts. We can see the common theme are café and restaurants. This cluster is good for tourist who like to explore different kind of food.

	District	District (Chinese)	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Zhongzheng	中正區	1	Café	History Museum	Monument / Landmark	Noodle House	Bakery
1	Datong	大同區	1	Taiwanese Restaurant	Café	Coffee Shop	Convenience Store	Night Market
2	Zhongshan	中山區	1	Hotel	Seafood Restaurant	Convenience Store	Chinese Restaurant	Massage Studio
3	Songshan	松山區	1	Convenience Store	Seafood Restaurant	Taiwanese Restaurant	Gym / Fitness Center	Bookstore
4	Daan	大安區	1	Café	Convenience Store	Tea Room	Coffee Shop	Vietnamese Restaurant
5	Wanhua	萬華區	1	Taiwanese Restaurant	Dessert Shop	Chinese Restaurant	Convenience Store	Coffee Shop

Cluster 3 (Culture and Outdoor Place). There is only 1 district for this cluster. The top venue for this is to visit the traditional train station or water park. This district is good for tourists like to enjoy the culture or outdoor place.

	District	District (Chinese)	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
21	Gongliao	貢寮區	2	Train Station	Water Park	Food Stand	Dessert Shop	Dim Sum Restaurant

Cluster 4 (Shopping area). There is only 1 district for this cluster. This district have restaurants, dessert, department store which fits for tourists who like shopping and eat.

Cluster 4

```
taipei_merged.loc[taipei_merged['Cluster Labels'] == 3, taipei_merged.columns[[0,2] + list(range(5, taipei_merged.shape[1]))]]
```

	District	District (Chinese)	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
17	Shiding	石碇區	3	Chinese Restaurant	Dessert Shop	Department Store	Dim Sum Restaurant	Diner

Discussion

From the above result, we can see there are 4 segments we can grouped based on the 39 districts we have originally. The four segments are:

1. Residential: The top venues for this cluster is convenience store and supermarket. The conveniency is good for local people where they can easier get the grocery they need.

2. Food Explorer: This is the largest segment k-mean cluster identified. The top venues associated with the segment are the various restaurants, from Taiwaness to seafood. These districts could be attractive for tourists who like food and open to explore different kind of cuisine.

3. Culture and Outdoor: The top venues show up for this district is the train station (which is a historical old train station full of culture and history) and the water park. The district can attract tourist who like to learn the culture and enjoy outdoor activity.

4. Shopping: The top venues for the district are restaurant, department store. This could be a good place to visit for tourists who like to shop and enjoy having a good meal after shopping.

Conclusion

In this project, the objective is to make tourists easier to find what place to go based on the characteristics show in each cluster. Therefore, I used k-means clustering to segment the 39 districts into 4 segments and identified residential area segment, food explorer segment, culture and outdoor segment and shopping segment. Based on tourists interest, they can quickly learn which districts they can visit.