**Final Report | Capstone Project – The Battle of Neighborhoods Finding A Place To Visit in Penang**

**1. Introduction:**

Penang, the city the author lives in, is a Malaysian state located on the northwest coast of Peninsular Malaysia, by the Malacca Strait, attracts a large number of tourists every year. For tourists, finding the right place to eat and hang around can be a challenge, though. Some crave for delicious food while other may prefer sightseeing. Thus, with this exercise, I wanted to give a simple reference and recommendation to tourists in Penang. Which district in Penang should food hunter visit? Which areas are best for coffee lovers? If a tourist like sunbath a lot, which hotel should he/she stays?

**2. Data Section**

I will use the overview of districts/city parts of Northeast Penang from Wikipedia: <https://en.wikipedia.org/wiki/Northeast_Penang_Island_District>

**Foursquare API:**

As requested by the assignment task, this project would use Four-square API as its prime data gathering source about Penang Northeast Region as it has a database of millions of places, especially their places API which provides the ability to perform location search, location sharing and details about a business.  
I will use foursquare data such as the venue name, ID, location and category(vegetarian, Italian etc.).

After finding the list of neighborhoods, then connect to the Foursquare API to gather information about venues inside each and every neighborhood. For each district, we have chosen the radius to be 500 meter.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

1. Neighborhood

2. Neighborhood Latitude

3. Neighborhood Longitude

4. Venue

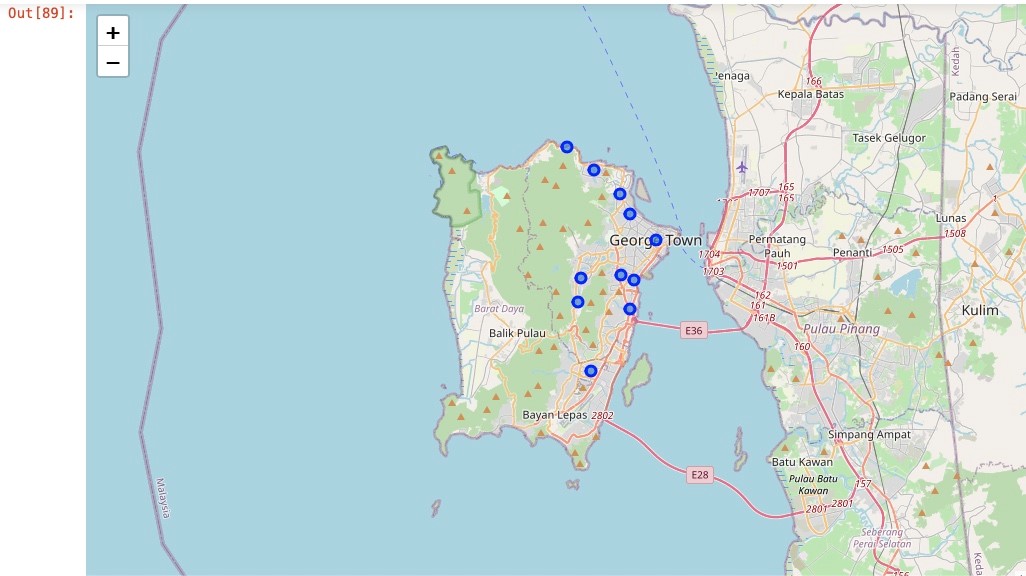
5. Name of the venue e.g. the name of a store or restaurant

6. Venue Latitude

7. Venue Longitude

8. Venue Category

**Map of Northeast Penang**

****

**3. Methodology Section**

**Work Flow:**

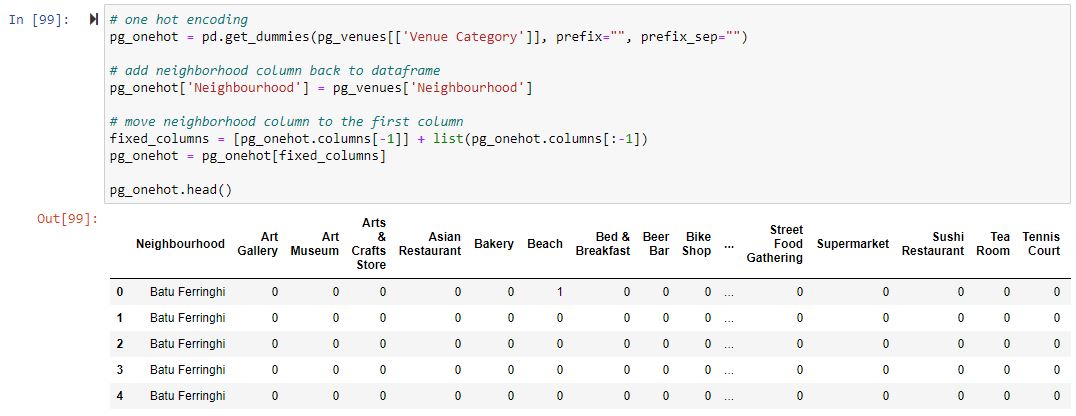
Using credentials of Foursquare API features of near-by places of the neighborhoods would be mined. Due to http request limitations the number of places per neighborhood parameter would reasonably be set to 100 and the radius parameter would be set to 500.

**Clustering Approach:**

To compare the similarities of the districts, i will explore the neighborhoods, segment them, and group them into clusters to find similar neighborhoods in each district. To be able to do that, we need to cluster data which is a form of unsupervised machine learning: k-means clustering algorithm.

**Using One-hot encoding:**

To find clusters of restaurant types in the different city districts, I first transformed the data frame with the venues information, associated to city districts, by one-hot encoding (0/1), as seen in the picture below

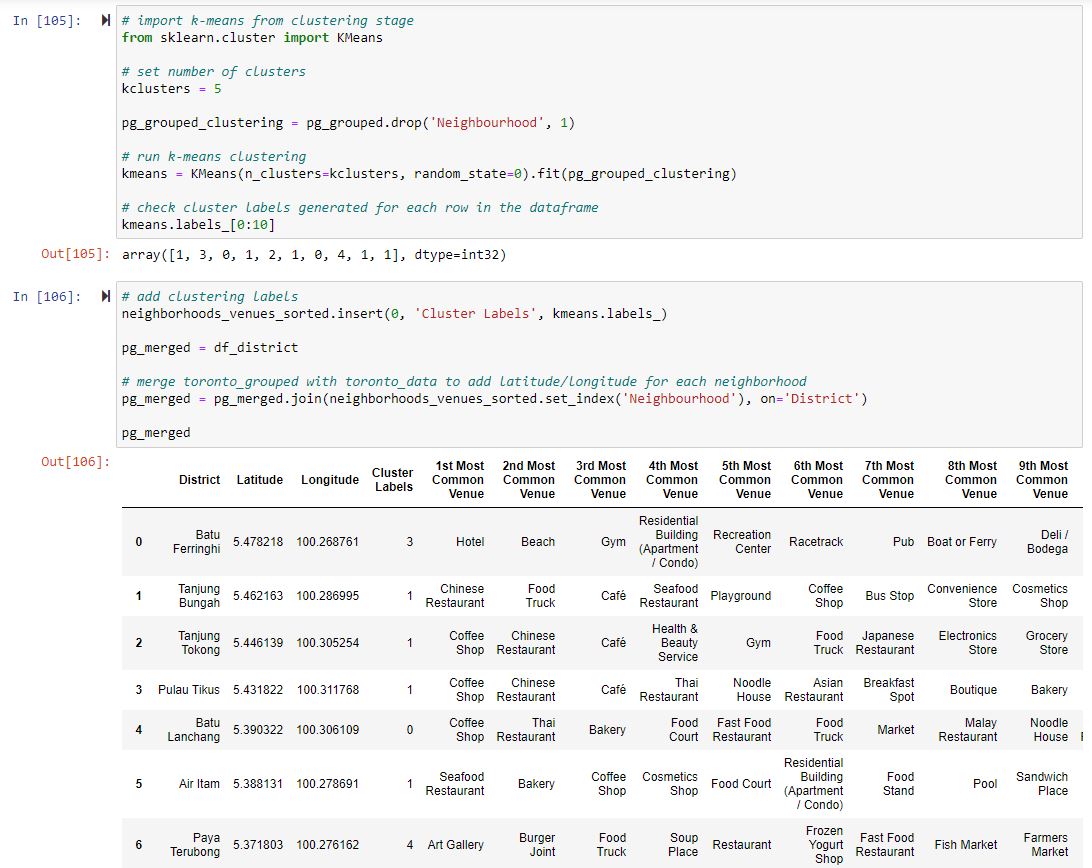
****

Next, I used grouping to show the frequency of each category of venues in each city district.

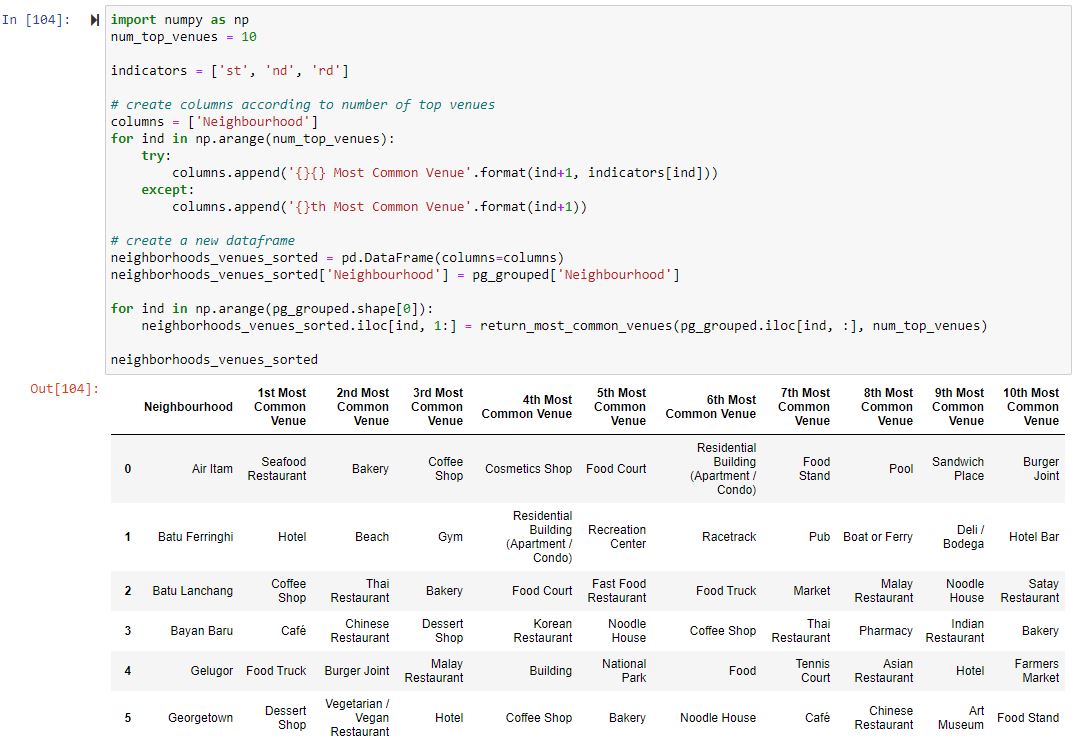


**Using K-Means Clustering Approach:**

With the data, k-means clustering algorithm from the scikit-learn package is used. One could use the elbow method to systematically define the k value, but I simply chose k to be 5, having been inspired by one of the coursera courses to do so.



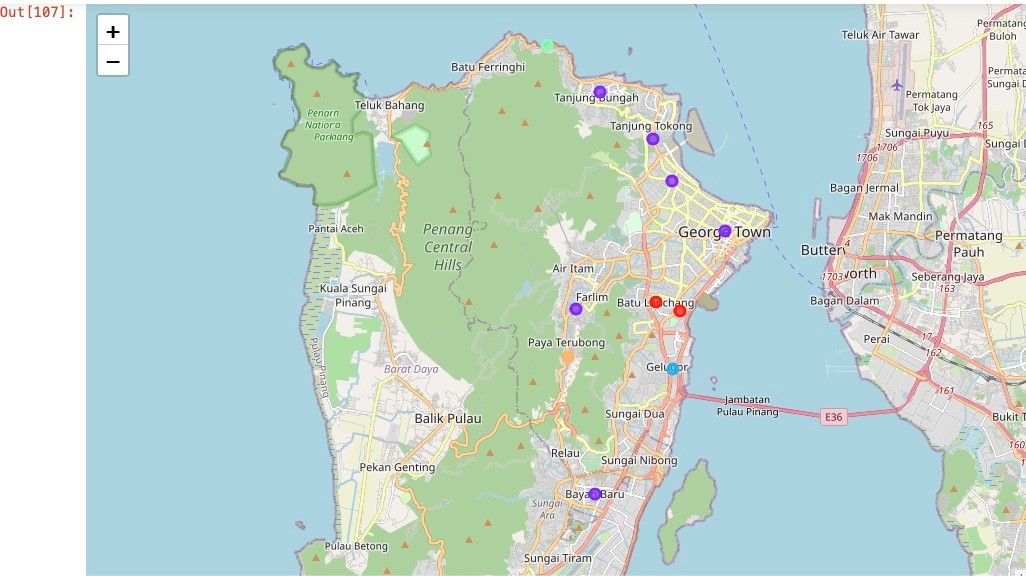
**Most Common venues near Neighborhood:**

****

**4. Results Section**

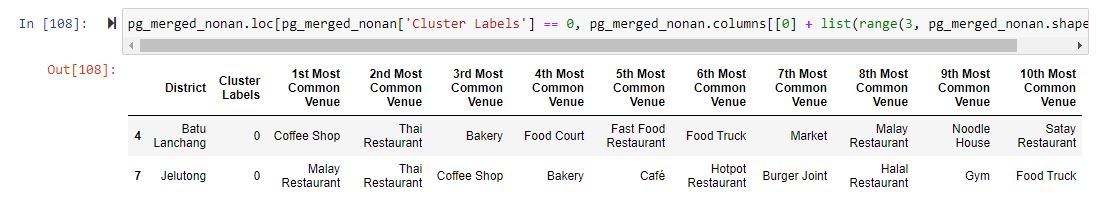
What we see in the map are the city districts and their most common venues, and they now have been assigned five different cluster labels from 0 to 4.We can now use the cluster labels to show the city districts marked with a cluster-specific color on a map, again using folium:

**Map of Clusters in Northeast Penang**

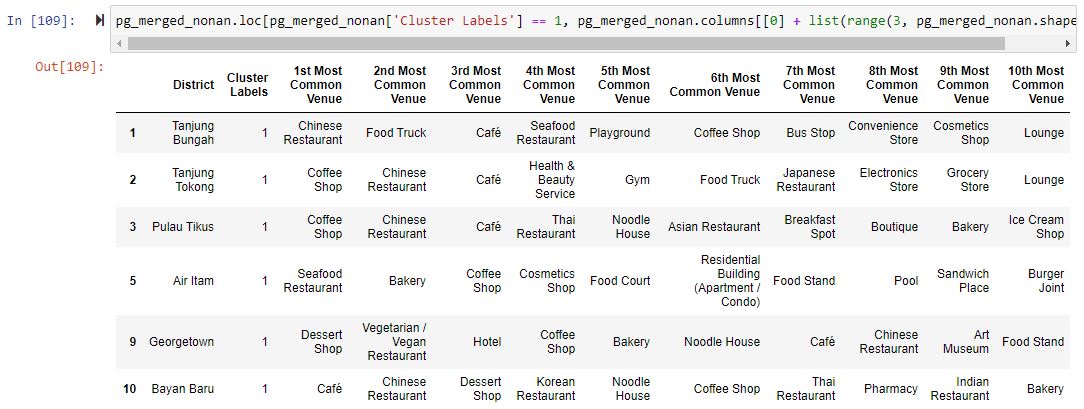
****

You will see 5 colors for the city districts, for the five different clusters, which I named according to the venues concentration the data shows. There is 1 light green dot on top of the Island and 1 light orange in the middle if you can't see them.

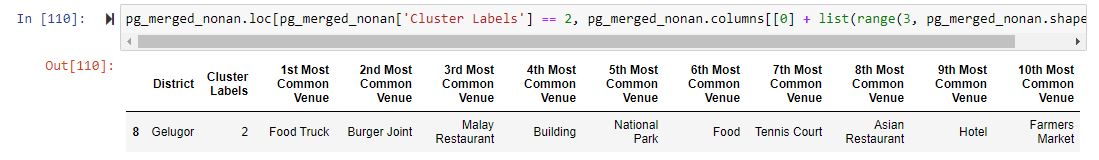
**Cluster 0 - The Thai Food Cluster**



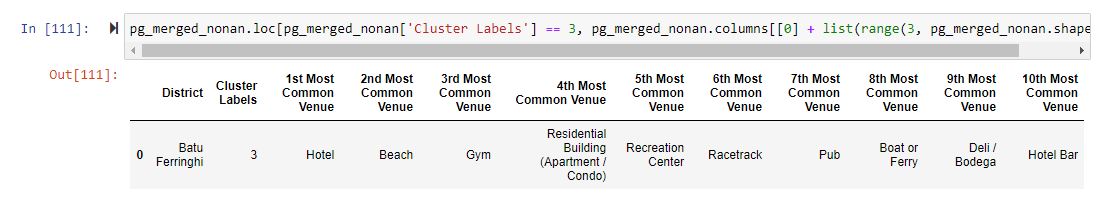
**Cluster 1 - The Chinese Food and Coffee shop Cluster**

****

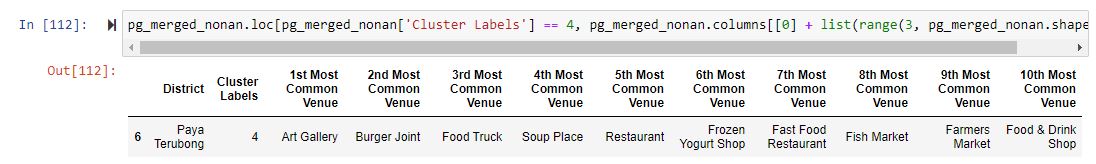
**Cluster 2 - The Food Truck Cluster**

****

**Cluster 3 - The Hotel and Beach Cluster**

****

**Cluster 4 - The Art Cluster**



**5. Discussion Section**

**Problem Which Tried to Solve:**

The major purpose of this project, is to suggest a place to visit in Penang for different tourist with specific interests. Tourists can visit the clusters based on their desires. Also, a simple laptop is enough to run the packages and provide comprehensive clustering for recommendation. What make this interesting is we could simply change the district name, from any location on earth, similar clustering can be done for other city.

**6. Conclusion Section**

Tourists can see in the results which city districts best match their desires. This is just one example of fantastic data science uses cases one can realize applying technology which is available for free today! I feel rewarded with the efforts and believe this course with all the topics covered is well worthy of appreciation. This project has shown me a practical application to resolve a real situation that has impacting personal and financial impact using Data Science tools. The mapping with Folium is a very powerful technique to consolidate information and make the analysis and decision better with confidence.