The Battle of the Neighborhoods
Pizza Places in Selangor and Kuala Lumpur

1. Introduction/Business Problem

1.1. I love pizza. Well, who doesn't? And yet, there are only a few pizza places are available in where I live and work, which is the state of Selangor, Malaysia. There are also just a handful of pizza places in Kuala Lumpur (where I work).

1.2. Perhaps, this is due to my ignorance of the available choices. Or maybe, there are really few choices available.

1.3. Hence, the problem that this project attempts to solve would be:

   i. What is the distribution of pizza places in Selangor and Kuala Lumpur (referred to as "my hometown" from now on); and

   ii. Where would be the best location to open a new pizza place in my hometown.

1.4. Target audience for this project are as follows:

   i. Myself. I travel a lot within my hometown for my work., therefore, the result of this project should improve my pizza meal choices moving forward.

   ii. Existing and future restaurant owners. The project should somewhat benefit this group. Food & beverage industry is very competitive, particularly for Western cuisine restaurants. A Southeast Asian country, rice-based meals are the preferred option for the locals, hence, location is a critical factor for Western cuisine restaurants.

2. Data

Throughout this project, the following data will be exploited through data cleaning, data wrangling, exploratory data analysis and machine learning techniques:

2.1. List of neighborhoods in my hometown

No known single source, hence, we will be using these sources from Wikipedia:

* https://en.wikipedia.org/wiki/Category:Townships_in_Selangor; and

* https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur

Web scraping techniques are used by using Python requests and BeautifulSoup packages.

2.2. Coordinates of the listed neighborhoods

Python Geocoder package should be able to provide us with the latitude and longitude of the listed neighborhoods.

2.3. List of pizza places in the listed neighborhoods

Foursquare API will be utilized as datasource, which should be able to provide to us with all kind of information in relation to venues, which include but not limited to venue name, category, menu, as well as location. Particularly, "Pizza Places" should be available as part of the information.

3. Methodology

Please refer to the working file via this link:
https://github.com/atfy-izad/Coursera_Capstone/blob/master/The%20Battle%20of%20Neighborhoods.ipynb

Step 1: Import relevant libraries.

This is self-explanatory.

Step 2: Scrape list of neighborhood data from sources and transform into a Dataframe.

Quite significant time was spent on data cleaning and data wrangling, considering that no single source is available for list of neighborhoods in my hometown.

FIrstly, we get data from Wiki pages mentioned earlier. Then, using BeautifulSoup package, the data from the html is then parsed and then stored as a list. Next, a dataframe is created using the list.

Once the neighborhood dataframe is created, the neighborhoods are then converted into geographical coordinates via Geocoder package. A separate coordinate dataframe was created and then merged with the existing neighborhood dataframe.

The same process is applied to both list of townships in Selangor as well as list of suburbs in Kuala Lumpur.

Step 3: Merge both Selangor and Kuala Lumpur dataframes.

Both dataframes are then concatenated into a single dataframe.

For completeness, another column for "State" is also added to easily differentiate both territories. Additionally, As a sanity check, the merged dataframe is then downloaded for quick review in Excel for completeness and errors.

Step 4: Create a map for data visualization.

This step utilizes Folium package. Since Kuala Lumpur is located in the middle of Selangor, Kuala Lumpur is chosen as map centre.

This step also serves as a preliminary sanity check of data gathered so far.

Step 5: Use Foursquare API to explore neighborhoods.

Using Foursquare API, we first extracted top 50 venues that are within a radius of 10,000 metres. – 50 to limit places rank and 10,000 metres to reflect delivery range of typical pizza places.

The information are then filtered specifically for "Pizza Places" and merged with the earlier neighborhood dataframe.

Step 6: Analyze each neighborhood.

This is where we apply k-means clustering to further segmentise the pizza places in my hometown.

The neighborhoods were clustered into 3 based on the frequency of occurrence for "Pizza Place". This allows us to appreciate the concentration of pizza places in each neighborhood.

4. Results

The map below depicts the visualization if the k-means clustering result, which categorizes the following:

Cluster	Description	Colour
0	Neighborhoods with zero to low concentration of pizza places	Purple
1	Neighborhoods with medium concentration of pizza places	Light green
2	Neighborhoods with high concentration of pizza places	Red
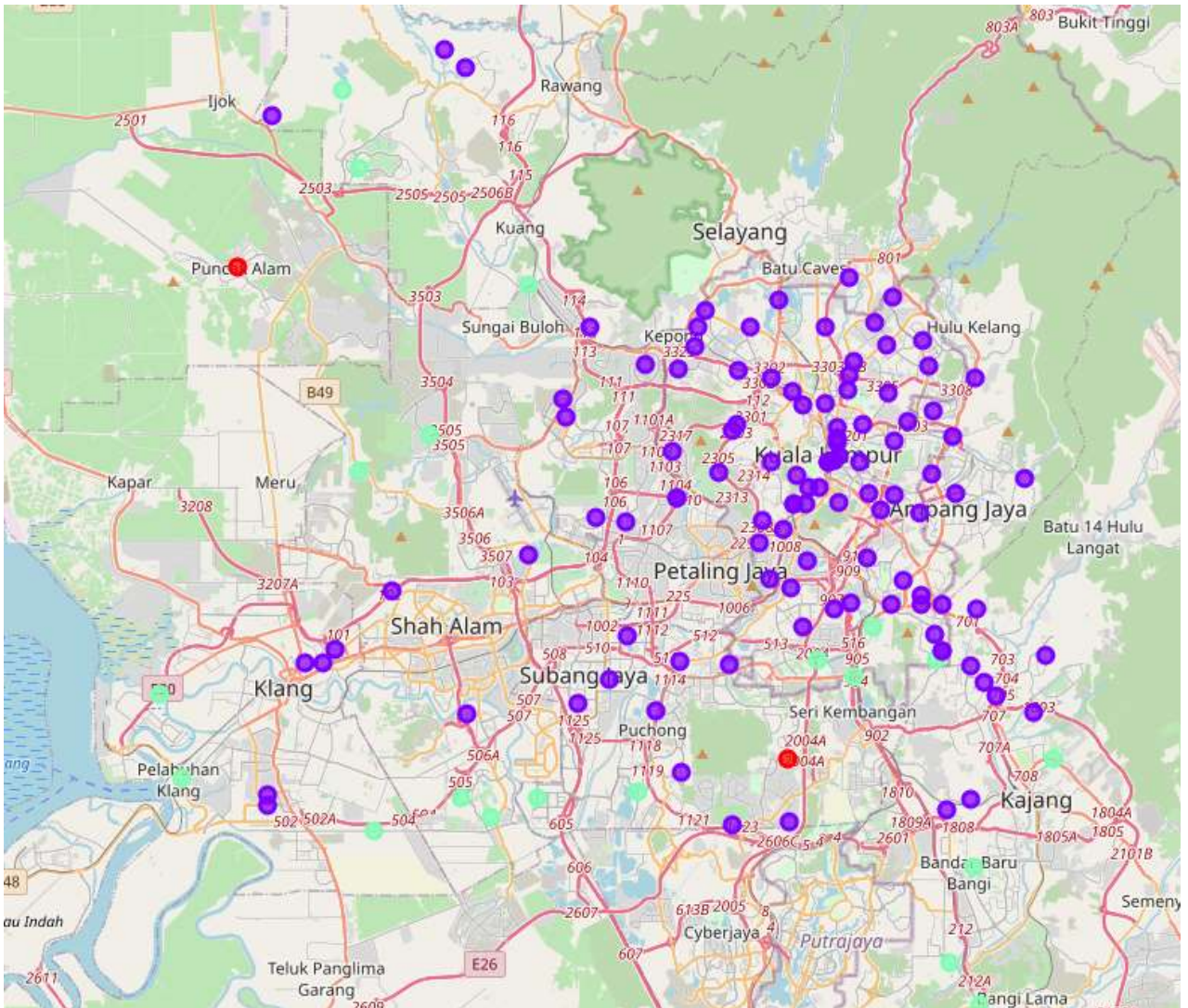
As you can see, most of the neighborhoods within my hometown are in cluster 0. An interesting observation, even for affluent neighborhoods such as the ones within Kuala Lumpur, have zero or low concentration of pizza places.

Another observation to note that 3 neighbourhoods – Puncak Alam, Puncak Jalil, and Bukit Beruntung – has high concentration of pizza places. A quick glance show that these neighbourhoods are pretty secluded and do not have much competition in food & beverage industry.

5.   Discussion

Of course, this result is still preliminary. The result could mean that there is a good prospect for a pizza place practically anywhere in my hometown. On the other hand, it could also mean that pizza is not really chosen food for Malaysians to dine.

Refining the result would require data on population as well as income level per capita for each neighborhood. Additionally, the project can also benefit by further segregating fast food pizza as well as gourmet pizza restaurants to provide better clarity to target audience.

6.   Conclusion

The result of the project confirmed my hypothesis that there aren't many pizza places in my hometown – a sad observation indeed.

The lack of pizza places could be a missed opportunity by restaurateurs in these neighborhoods, which in my opinion, should be affluent and modern enough to have pizza more frequently.

Further evaluation and refinement is necessary to further evaluate the rationale behind this finding.