IBM Data Science Capstone Project

Danh Thai Hoang
danhthaihoang@gmail.com

April 2020

Opening Italian Restaurants Business in Kuala Lumpur

1. Introduction

Starting a Restaurant Business in Malaysia

We can see that more and more restaurants are popping up in the city. The phenomenon proves one thing: the restaurant business is one of the common business ventures in today's business world. That said, building a restaurant from scratch is not easy – all of those advantages come at a price.

Snapshot of Advantages of a Restaurant Business

Though starting a restaurant business is a hard and expensive process, there are advantages you can get when running a restaurant business. As restaurants are in huge demand, a good restaurant will always be filled with customers- that means it will generate revenues. The

probability of earning big profits (and encountering less loss) is high in opening a restaurant business. In other words, the right restaurant at the right location can be a lucrative business.

Things to Consider Before Starting a Restaurant Business in Malaysia. Having a business plan is crucial to success. A carefully planned business plan will steer the direction of your restaurant business thus ensuring lucrative return. Let us dive in to have a look at what you should know before starting a restaurant business in Malaysia.

Concept:

Know and determine your restaurant's concept before you launch it. If you are unsure, you can always run some research; consult the professionals to streamline your idea. Generally, the concept is like the road map that includes your restaurant's menu, investment's capital, and source of financial aids, expected profit, and marketing plan.

Location:

The location of your restaurant business will affect its success nearly as much as the menu. As such, you may want to research an ideal location with high traffic, easy access, and good visibility. Other factors that could

determine your restaurant's location is the target market of your business as well as the competitors nearby.

Menu:

The menu of a restaurant is another big consideration when you are planning a restaurant business. When you are deciding what to include in the menu, the first thing you should know is the target market, the current food trend and what you plan to serve. A good restaurant menu is the key to any restaurant's marketing plan.

Marketing:

Marketing plan is a plan of how to get the word out about your new restaurant. Nowadays the marketing plan sees a lot of variety when you can take advantage of the many social media platforms. Thinking it through with brilliant ideas will actually help your restaurant to succeed and generate huge profits.

Licensing and permits:

Before you can open the doors to your restaurant, you need to make sure you apply for the proper licenses. In Malaysia, a Malaysia business license is issued on various terms depending on whether the investor is local or foreign. For restaurant business, a Signboard License, Alcohol License (if you are planning to sell alcohol in your

restaurant), Halal License (if applicable) MACP and PPM (for restaurants where they will play music in the premises). If the restaurant is a foreign-owned restaurant, an approved Wholesale, Retail Trade License is needed. Please note that such license is only granted for restaurants with unique concepts (like Arabian restaurants, French cuisine) and a minimum space of 1,500 sq. ft.

The problem statement

As the final assignment of IBM Data Science professional certificate on the Coursera, the project is intended to apply possible knowledge and skills in Data Science gained from different courses along the certificate to resolve the problem in the real world.

Because the location of the Restaurant – This is the key to the future of your business the project tries to answer the questions: **“Should an entrepreneur start a Italian restaurant business in Kuala Lumpur City, Malaysia?” and “Which location is suitable and recommended to open it”**. With applying various data science methodology and machine learning techniques from linear regression to clustering, classification the project work aims to answer these questions.

Since there are lots of restaurants in Kuala Lumpur we will try to detect **locations** and that are not already crowded with restaurants. If there is no such location, then should not open a restaurant business there. We are also particularly interested in **areas** with no restaurants in vicinity. We would also prefer locations as close to city center as possible, assuming that the first two conditions are met.

We will use our data science powers to generate a few most promising neighborhoods based on this criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

Target audience:
Any investors or entrepreneurs who are interested in open Italian restaurants business in Kuala Lumpur.

2. Data

Data Requirements

Because the location is the most important factor for further analysis, visualizing, modelling and evaluation as

well so it requires the Kuala Lumpur's **neighborhoods data**. This defines the scope of this project which is confined to the city of Kuala Lumpur, the capital city of the country of Malaysia in SouthEast Asia.

Latitude and longitude coordinates of those neighbourhoods are also required to visualize and plot the map and also to get the venue data.

Venues data, particularly data related to restaurants. This is the crucial data for clustering on the neighbourhoods.

Data collection

- The list of Kuala Lumpur's **neighborhoods data** is collected from the wiki page: https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur with more than 70 **neighborhoods** in the list. We will crawl the data from the page by using BeautifulSoup library, Python requests and Pandas. The collection process looks like this:

Data Collection

Wiki page to get the list of Kular Lumpur

```
In [10]: url = 'https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur'
```

Call the bs4 lib to start crawling the data

```
In [14]: nData = requests.get(url).text
```

Parse the data collected from wiki page

```
In [15]: nSoup = BeautifulSoup(nData, 'html.parser')
```

Create the list of neighborhoods

```
In [20]: neighborhoodList = []
         for row in nSoup.find_all("div", class_="mw-category")[0].findAll("li"):
             neighborhoodList.append(row.text)

Out[20]: ['Alam Damai',
          'Ampang, Kuala Lumpur',
          'Bandar Menjalara',
          'Bandar Sri Permaisuri',
          'Bandar Tasik Selatan',
          'Bandar Tun Razak',
          'Bangsar',
```

- **The latitude and longitude coordinates** of the neighbourhoods are retrieved using Google Maps Geocoding API. The geometric location values are then stored into the initial dataframe.

Get the geographical coordinates

Define a function to get lat and long

```
In [47]: def get_latlng(neighborhood):
             # initialize your variable to None",
             lat_long_coords = None
             #loop until you get the coordinates",
             while(lat_long_coords is None):
                 g = geocoder.arcgis('{}, Kuala Lumpur, Malaysia'.format(neighborhood))
                 lat_long_coords = g.latlng
             return lat_long_coords
```

Call the function to get lat and long list and store in the new list (include the neigborhoods and their coordinates

```
In [48]: coords = [get_latlng(neighborhood) for neighborhood in kuLu_df["Neighborhood"].tolist()]
```

Merge the neigborhoods data into coordinates data into one dataframe

```
In [49]: neigh_coors_df = pd.DataFrame(coords, columns=['Latitude', 'Longitude'])
         neigh_coors_df.head()
```

Out[49]:

	Latitude	Longitude
0	3.057690	101.743880
1	3.148494	101.696729
2	3.190350	101.625450
3	3.103910	101.712260
4	3.072750	101.714610

- From the location data obtained after Web Scraping and Geocoding, the venue data is found out by passing in the required parameters to the **FourSquare API (https://developer.foursquare.com/)**, and creating another DataFrame to contain all the venue details along with the respective neighbourhoods.

3. Methodology

This project will direct our efforts on detecting areas of Kuala Lumpur that have low restaurant density, particularly

those with low number of Italian restaurants. We will limit our analysis to areas ~6km around the city center.

In the first step we have collected the required data: location and type (category) of every restaurant within 6km from Kuala Lumpur center (KLCC). We have also identified Italian restaurants (according to Foursquare categorization).

Second step in our analysis will be calculation and exploration of 'restaurant density' across different areas of Kuala Lumpur - we will use heatmaps to identify a few promising areas close to the center with low number of restaurants in general (and no Italian restaurants in vicinity) and focus our attention on those areas.

In the third and final step we will focus on the most promising areas and within those create clusters of locations that meet some basic requirements established in discussion with stakeholders: we will take into consideration locations with no more than two restaurants in a radius of 250 meters, and we want locations without Italian restaurants in radius of 400 meters. We will present a map of all such locations but also create clusters (using k-means clustering) of those locations to identify general zones / neighborhoods / addresses which should be a

starting point for final 'street level' exploration and search for optimal venue location by stakeholders.

Exploratory Analysis

Let's create latitude & longitude coordinates for centroids of our candidate neighborhoods. We will create a grid of cells covering our area of interest which is approx. 12x12 kilometers centered around Kuala Lumpur city center.

Let's first find the latitude & longitude of Kuala Lumpur city center, using specific, well known addresses and Google Maps geocoding API.

Output:

```
Coordinate of Suria KLCC, Kuala Lumpur, Malaysia: [3.1580207, 101.7116671]
```

Now let's create a grid of area candidates, equally spaced, centered around the city center and within ~6km from Suria KLCC. Our neighborhoods will be defined as circular areas with a radius of 300 meters, so our neighborhood centers will be 600 meters apart.

To accurately calculate distances we need to create our grid of locations in the Cartesian 2D coordinate system which allows us to calculate distances in meters (not in

latitude/longitude degrees). Then we'll project those coordinates back to latitude/longitude degrees to be shown on the Folium map. So let's create functions to convert between WGS84 spherical coordinate system (latitude/longitude degrees) and UTM Cartesian coordinate system (X/Y coordinates in meters).

```
Kuala Lumpur center longitude=101.7116671, latitude=3.1580207
2219179.12425
Kuala Lumpur center UTM X=2219179.12425, Y=19634000.569
Kuala Lumpur center longitude=101.7116671, latitude=3.1580207
```

Let's visualize the data we have so far: city center location and candidate neighborhood centers:

```
Reverse geocoding check
-----------------------
Address of [3.1580207, 101.7116671] is: Lot LC-402, 404, Jalan Ampang,
Kuala Lumpur City Centre, 50088 Kuala Lumpur, Wilayah Persekutuan Kuala
Lumpur, Malaysia
```

Let's perform some basic exploratory data analysis and derive some additional info from our raw data. First let's count the number of restaurants in every area candidate:

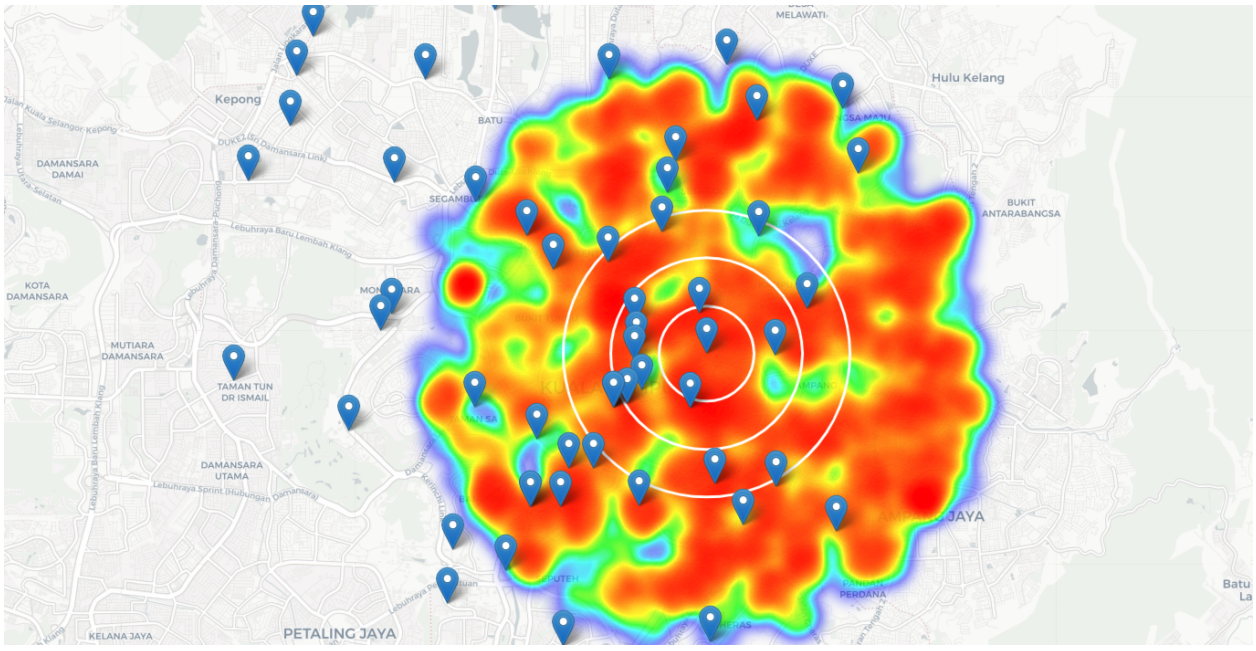
```
('Average number of restaurants in every area with radius=300m:',
7.164835164835165)
```

Now let's calculate the distance to the nearest Italian restaurant from every area candidate center (not only

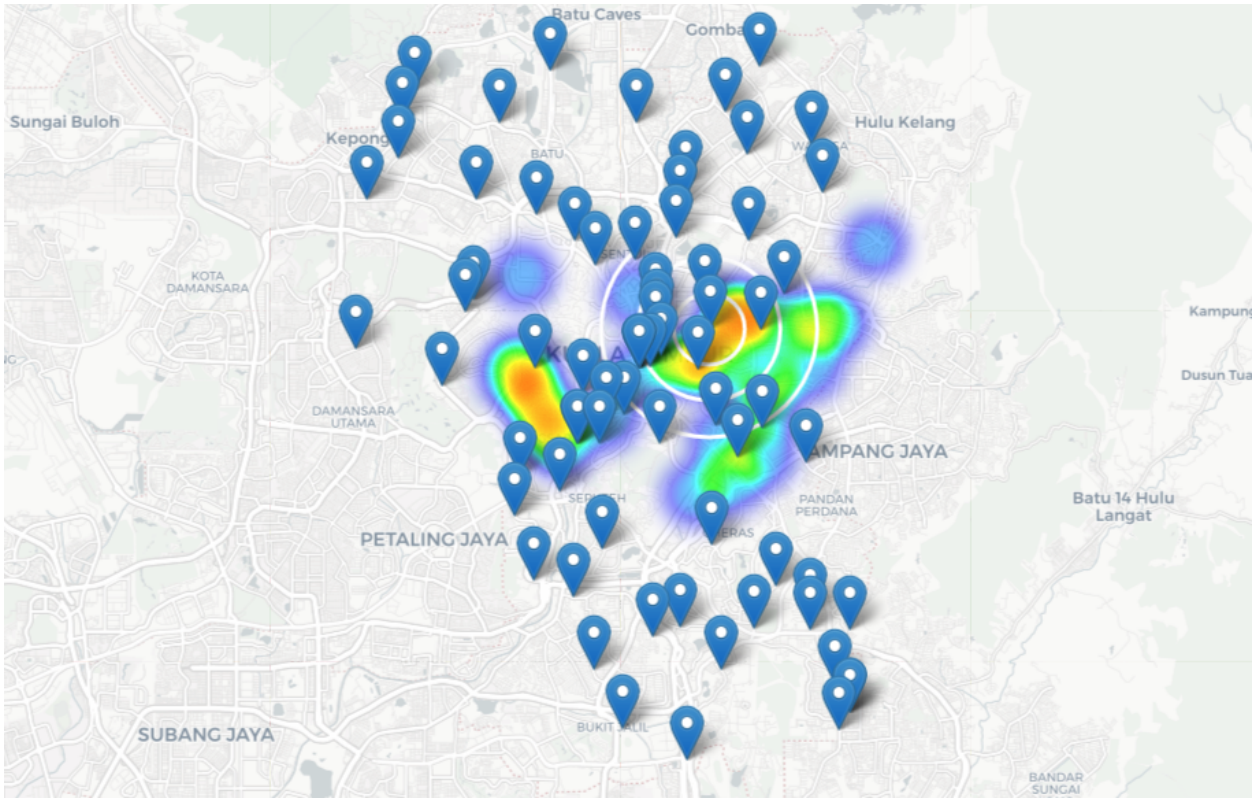
those within 300m - we want distance to closest one, regardless of how distant it is).

```
('Average distance to closest Italian restaurant from each area center:',
1460.6965522299477)
```

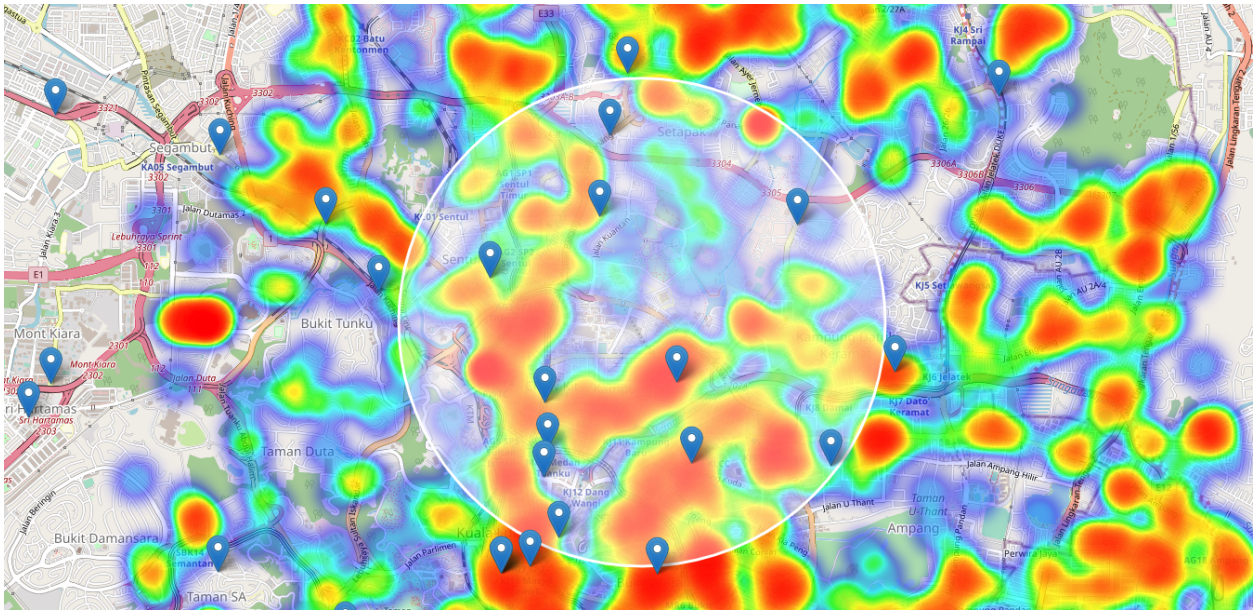
So on average Italian restaurant can be found within ~1460m from every area center candidate. That's fairly close, so we need to filter our areas carefully!
Let's create a map showing heatmap / density of restaurants and try to extract some meaningful info from that. Also, let's show markers of Kuala Lumpur boroughs on our map and a few circles indicating distance of 1km, 2km and 3km from Suria KLCC.



Let's create another heatmap map showing heatmap/density of Italian restaurants only.



Let's define a new, more narrow region of interest, which will include low-restaurant-count parts closest to Suria KLCC.



Not bad - this nicely covers all the pockets of low restaurant density in closest to Kuala Lumpur center

Now let's calculate two most important things for each location candidate: number of restaurants in vicinity (we'll use a radius of 250 meters) and distance to closest Italian restaurant.

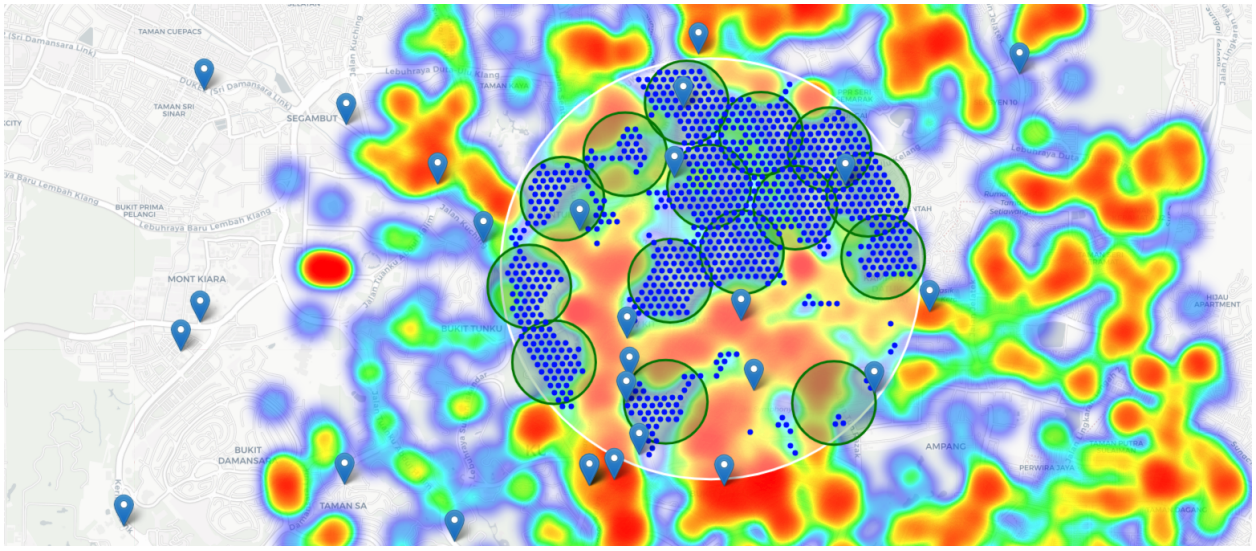
	Distance to Italian restaurant	Latitude	Longitude	Restaurants nearby	X	Y
0	3424.963359	3.192860	101.707235	6	2.219629e+06	1.963000e+07
1	3389.552037	3.192846	101.706367	4	2.219729e+06	1.963000e+07
2	3261.375329	3.192177	101.712021	7	2.219079e+06	1.963009e+07
3	3269.031296	3.192164	101.711153	11	2.219179e+06	1.963009e+07
4	3279.719834	3.192151	101.710285	11	2.219279e+06	1.963009e+07
5	3293.411418	3.192137	101.709417	12	2.219379e+06	1.963009e+07
6	3310.068783	3.192124	101.708549	12	2.219479e+06	1.963009e+07
7	3329.647417	3.192111	101.707681	12	2.219579e+06	1.963009e+07
8	3338.242594	3.192097	101.706813	5	2.219679e+06	1.963009e+07
9	3289.592130	3.192084	101.705945	0	2.219779e+06	1.963009e+07

Let us now filter those locations: we're interested only in locations with no more than two restaurants in a radius of 250 meters, and no Italian restaurants in a radius of 400 meters.

```
('Locations with no more than two restaurants nearby:', 969)
('Locations with no Italian restaurants within 400m:', 1927)
('Locations with both conditions met:', 937)
```

What we have now is a clear indication of zones with low number of restaurants in vicinity, and no Italian restaurants at all nearby.

Let us now cluster those locations to create centers of zones containing good locations. Those zones, their centers and addresses will be the final result of our analysis.



Not bad - our clusters represent groupings of most of the candidate locations and cluster centers are placed nicely in the middle of the zones 'rich' with location candidates.

Addresses of those cluster centers will be a good starting point for exploring the neighborhoods to find the best possible location based on neighborhood specifics.

Finally, let's reverse geocode those candidate area centers to get the addresses which can be presented to stakeholders.

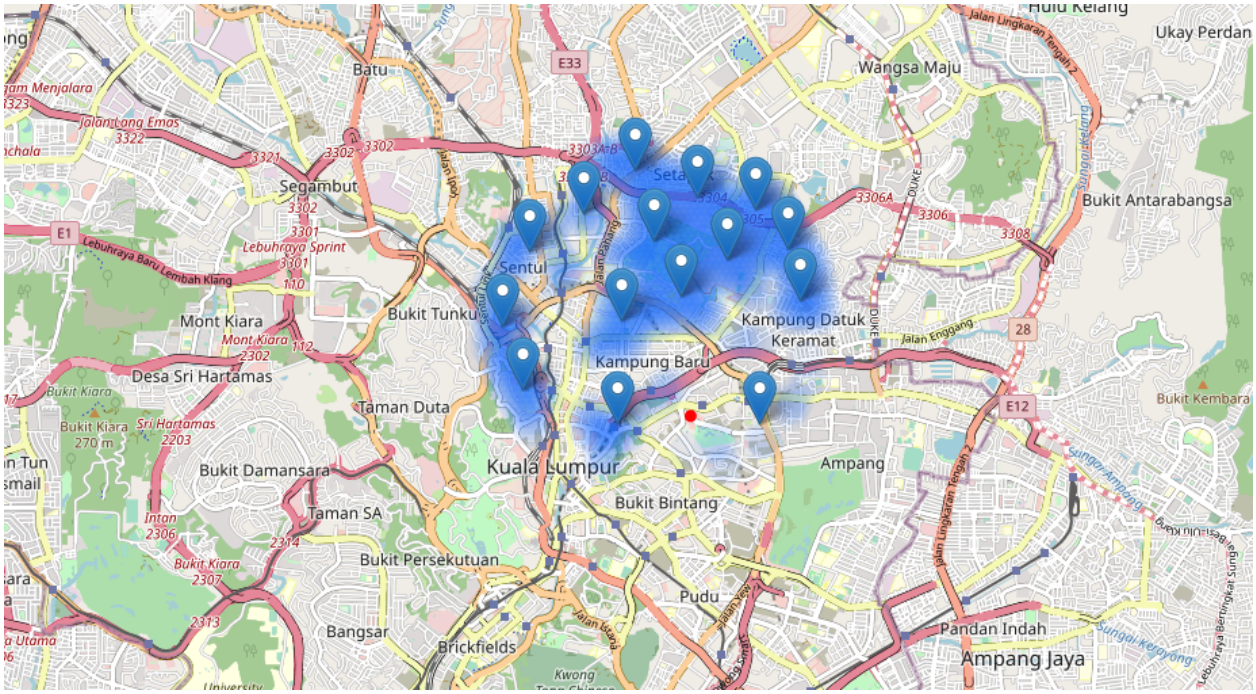
```
============================================================
Addresses of centers of areas recommended for further analysis
============================================================

2218187.45758
3, Lorong Kuda, Kuala Lumpur, 50450 Kuala Lumpur, Wilayah Persekutuan
Kuala Lumpur, Malaysia => 1.0km from Suria KLCC
2219036.26711
B-32-3A Bennington Residensi, Taman Ayer Panas, 53200 Kuala Lumpur,
Wilayah Persekutuan Kuala Lumpur, Malaysia => 3.1km from Suria KLCC
2220182.9704
1111, Jalan Dr Latif, 50586 Kuala Lumpur, Wilayah Persekutuan Kuala
Lumpur, Malaysia => 1.7km from Suria KLCC
2217555.08579
Markas Latihan TD, Kampung Datuk Keramat, 54000 Kuala Lumpur, Federal
Territory of Kuala Lumpur, Malaysia => 2.3km from Suria KLCC
2218629.12425
Jalan Johor, Pusat Latihan Polis (PULAPOL), 54100 Kuala Lumpur, Wilayah
Persekutuan Kuala Lumpur, Malaysia => 2.3km from Suria KLCC
2221495.79092
139, Jalan Union, Sentul, 51000 Kuala Lumpur, Wilayah Persekutuan Kuala
Lumpur, Malaysia => 3.3km from Suria KLCC
2219687.7264
```

```
Taman Tasik Titiwangsa, Tasik, Titiwangsa, 53200 Kuala Lumpur, Wilayah
Persekutuan Kuala Lumpur, Malaysia => 2.6km from Suria KLCC
2221919.50886
Jalan 11, Jalan Ipoh, 51200 Kuala Lumpur, Wilayah Persekutuan Kuala
Lumpur, Malaysia => 3.0km from Suria KLCC
2220260.4968
Dang Wangi, Kuala Lumpur, 50300 Kuala Lumpur, Federal Territory of Kuala
Lumpur, Malaysia => 1.1km from Suria KLCC
2219947.87425
416, Jalan Pahang, Setapak, 53000 Kuala Lumpur, Wilayah Persekutuan Kuala
Lumpur, Malaysia => 3.6km from Suria KLCC
2219294.7869
Jalan Asrama, Titiwangsa, 53200 Kuala Lumpur, Wilayah Persekutuan Kuala
Lumpur, Malaysia => 1.7km from Suria KLCC
2217726.70489
Unnamed Road, Kampung Padang Tembak, 54100 Kuala Lumpur, Federal Territory
of Kuala Lumpur, Malaysia => 2.8km from Suria KLCC
2218196.98139
2, Lorong Ayer Leleh, Taman Ayer Panas, 53200 Kuala Lumpur, Wilayah
Persekutuan Kuala Lumpur, Malaysia => 3.1km from Suria KLCC
2220713.12425
43-0-8, Jalan 1/48A, Sentul Perdana, Bandar Baru Sentul, WP Kuala Lumpur,
Bandar Baru Sentul, 51000 Kuala Lumpur, Federal Territory of Kuala Lumpur,
Malaysia => 3.3km from Suria KLCC
2221629.12425
Club House, Jalan Tun Ismail, Kuala Lumpur, 50480 Kuala Lumpur, Federal
Territory of Kuala Lumpur, Malaysia => 2.5km from Suria KLCC
```

This concludes our analysis. We have created 15 addresses representing centers of zones containing locations with low number of restaurants and no Italian restaurants nearby, all zones being fairly close to the city center (all less than 4km from Suria KLCC, and about half of those less than 2km from Suria KLCC). Although zones are shown on map with a radius of ~500 meters (green circles), their shape is actually very irregular and their centers/addresses should be considered only as a starting

point for exploring area neighborhoods in search for potential restaurant locations.



4. Results and Discussion

Our analysis shows that although there is a great number of restaurants in Kuala Lumpur (**~2800** in our initial area of interest which was **12x12km** around Suria), there are pockets of low restaurant density fairly close to the city center. Highest concentration of restaurants was detected north and west from West, so we focused our attention to areas south, south-east and east.

After directing our attention to this more narrow area of interest (covering approx. **5x5km** south-east from Suria KLCC) we first created a dense grid of location candidates (spaced **100m** apart); those locations were then filtered so that those with more than two restaurants in a radius of **250m** and those with an Italian restaurant closer than **400m** were removed.

Those location candidates were then clustered to create zones of interest which contain the greatest number of location candidates. Addresses of centers of those zones were also generated using reverse geocoding to be used as markers/starting points for more detailed local analysis based on other factors.

Result of all this is **15 zones** containing the largest number of potential new restaurant locations based on number of and distance to existing venues - both restaurants in general and Italian restaurants in particular. This, of course, does not imply that those zones are actually optimal locations for a new restaurant! Purpose of this analysis was to only provide info on areas close to Kuala Lumpur center but not crowded with existing restaurants (particularly Italian) - it is entirely possible that there is a very good reason for small number of restaurants in any of those

areas, reasons which would make them unsuitable for a new restaurant regardless of lack of competition in the area. Recommended zones should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only no nearby competition but also other factors taken into account and all other relevant conditions met.

5. Conclusion

Purpose of this project was to identify Kuala Lumpur areas close to the center with a low number of restaurants (particularly Italian restaurants) in order to aid stakeholders in narrowing down the search for the optimal location for a new Italian restaurant. By calculating restaurant density distribution from Foursquare data we have first identified general boroughs that justify further analysis), and then generated extensive collection of locations which satisfy some basic requirements regarding existing nearby restaurants. Clustering of those locations was then performed in order to create major zones of interest (containing greatest number of potential locations) and addresses of those zone centers were

created to be used as starting points for final exploration by stakeholders.

Final decision on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.

6. References

Wiki page
Google API
Foursquare Developer Portal