Recommendation System For Opening A New Restaurant In Kualalumpur, Malaysia.

By: Teguh Badrusalam

Abstract- This paper describes the techniques used to provide recommendations location-based to open a new restaurant with the help of machine learning clustering algorithms. One such algorithm used in this project is the k-means clustering. The basic idea behind this paper, is to firstly create a tool that helps us scap data from Malaysia wikipedia. Once the zip-codes and its corresponding latitudes & longitudes are available, we had to identify the venues popular in those areas. Associating the features/amenities for the various locations gives us our dataset and with this data we can run the clustering algorithm to find what location in kualalumpur similar to one another, thus providing us with a neighbourhood recommendation system.

Keywords: *clustering, K-means, Foursquare API, zip-codes, latitudes & longitudes..*

INTRODUCTION

Eating is a human activity that will never stop at any time because that restaurant is a place that will never go extinct as long as humans need food. For investor, the central location and the large crowd at the restaurant provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more restaurant to cater to the demand. As a result, there are many restaurant in the city of kuala lumpur and many more are being built. Opening restaurant allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new restaurant requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the restaurant is one of the most important decisions that will determine whether the restaurant will be a success or a failure.

METHODOLOGY

This section gives the detailed description of our work. The whole task was divided into two parts. They are (i) Data pre-processing, (ii) Model building. Figure 1 illustrates our overall workflow.

BUSINESS PROBLEM

The objective of this capstone project is to analyse and select the best locations in the city of kuala lumpur, malaysia to open a new restaurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: in the city of kuala lumpur, malaysia, if a property developer is looking to open a new restaurant, where would you recommend that they open

it?

DATA

To solve the problem, we will need the following data:

- list of neighbourhoods in kuala lumpur. 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 south east asia.
- latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.
- venue data, particularly data related to restaurant. We will use this data to perform clustering on the neighbourhoods.

SOURCES OF DATA AND METHODS TO TEXTRACT THEM

This wikipedia page (https://en.wikipedia.org/wiki/category:suburbs_in_kuala_lumpur) contains a list of neighbourhoods in kuala lumpur, with a total of 70 neighbourhoods. We will use web scraping techniques to extract the data from the wikipedia page, with the help of python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighbourhoods using python geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use foursquare api to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare api will provide many categories of the venue data, we are particularly interested in the restaurant category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (wikipedia), working with api (foursquare), data cleaning, data wrangling, to machine learning (k-means clustering) and map visualization (folium). In the next section, we will present the methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

DATA PREPROCESSING

The main challenge in this section is to generate an organized data set, which is not available online. The data processing task consists of fetching data from two APIs. That is Google Geocode API and FourSquare API. The following part contains the description of our Data preprocessing.

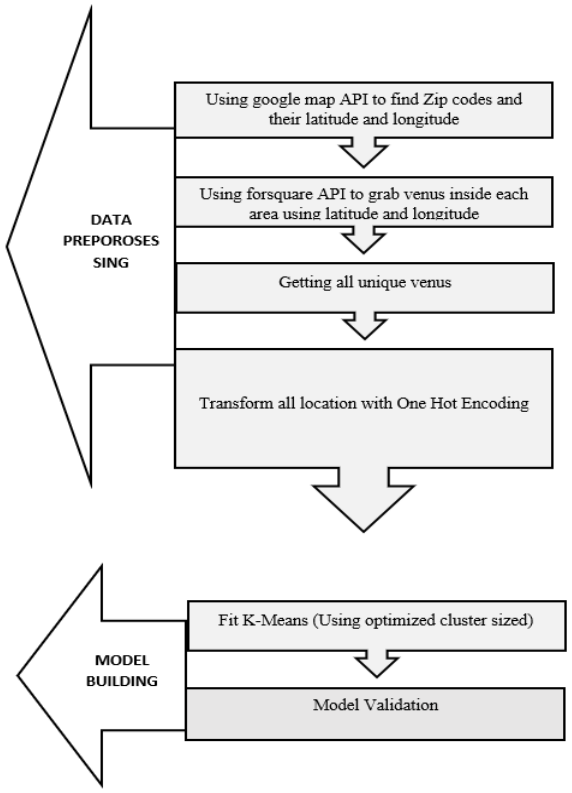


Figure 1.

(i) **Use of Google map API to find zip codes and corresponding area latitudes and longitudes:** here we came up with a tool that can be used to find out latitudes and longitudes of a given zip code. We used google Geocode Javascript API to find out the zip codes and their corresponding latitude and longitude. Finally we provided options for saving the data in a CSV file. Figure 2 shows the code snippet of fetching latitude and longitude. Once we have the zip codes, city names and their corresponding latitude & longitude we can download them in a csv file. Figure 3a and 3b shows the UI of our tool. From the downloaded CSV file, we use the latitudes and longitudes to fetch venues from FourSquare API.

(ii) **Using FourSquare API to identify venues related to zipcode:** We use FourSquare API to search for a particular place and filter by food, shops and outdoor places. Foursquare API (Python API)gives us access to a huge database of different venues from all around the world. It includes a rich variety of information such as the

place's addresses, popularity, tips and photos. The access to the API is available for free and provides an easy setup. It also enables the user to search for places using a simple input box and this information is fetched from the Foursquare API and displayed.

Before getting started, we need to create account for accessing the APIs: in order to access the places information. To get the API key from FourSquare, we first need to create an account on their website, click on "Create a new app" and fill in the details. At the new page, we will be granted with a Client ID and a Client Secret, both needed in order to be able to perform request to the API. We start by implementing the search function. First, we need to figure out what kind of request we have to do in order to get the data we desire. We will do a simple HTTP request which will return a response. The URL is in the form:

https://api.foursquare.com/v2/venues/4ccc574a063a721e83df8d9a?client_id=E3IEHLNNW5CRC3WA2K5DLRZBNVP2LKNTIOD4GN0YHWUKUBA5&client_secret=FHVGA3OGTSWQ4ORRHKYN1VQTWRT4GPIWTMACTMBFLYH0CNWJ&v=20190421

The parameters are as following: **client_id**: the generated client ID from FourSquare. **client_secret**: the generated secret from Foursquare. **v**: the version of the API that we are using. The venue ID is retrieved from the search query executed in the mainview controller. When we push the page, the ID is given as a parameter and used inside the details controller. After reading the ID, a query to the API will be executed. After the query has finished, we can parse the desired data out of the returned JSON object. Following this above query method, we collected top 50 venues within 2 miles radius of each of the zip-codes.

(iii) **Getting all unique venues**

(iv) **One Hot transformation:** A one hot encoding or transformation is a representation of categorical variables as binary vectors.This first requires that the categorical values be mapped to integer values. Then, each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1.

I. Model Building:

(i) **Optimal Cluster Number:** K-means is a simple unsupervised machine learning algorithm that groups a dataset into a user-specified number (*k*) of clusters. The algorithm is somewhat naive--it clusters the data into *k* clusters, even if *k* is not the right number of clusters to use. Therefore, when using k-means clustering, users need some way to determine whether they are using the right number of clusters.

(ii) **Run K-means clustering algorithm:**

Once we have the optimal number of clusters. We use python scikit learn API and k-means to fit our processed data

(iii) **Model Validation:** To validate the model, we followed a different approach. We seperated a portion of the dataset for testing. After getting the K-means model, we find the average distance of centroid to points for each cluster. And then we calculated the distance of each test data from the 16 clusters. Comparing the calculated distance with the average distance, we get an idea of which test point belongs to which cluster. So, we have manually calculated the clusters of the test points. Finally, we again used our model to predict the clusters of the test point. Comparing the manually calculated clusters for the test point with the predicted clusters by our model, we got an overview of the strength of our model.

RESULTS

In this section we describe our experimental work. Figure 2 shows the correlation between the 304 features of our dataset.

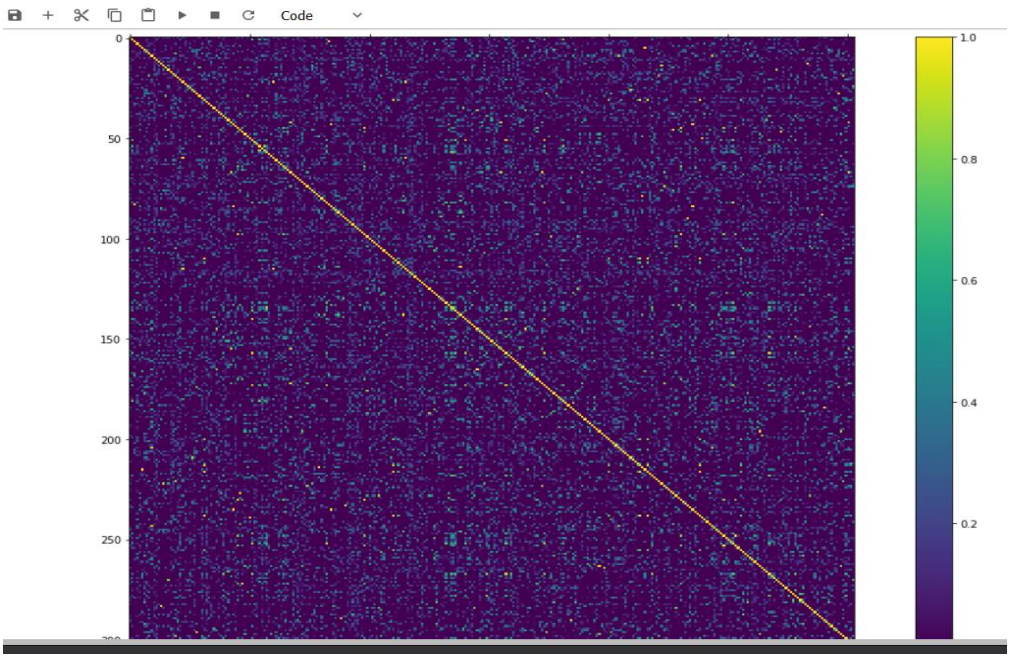
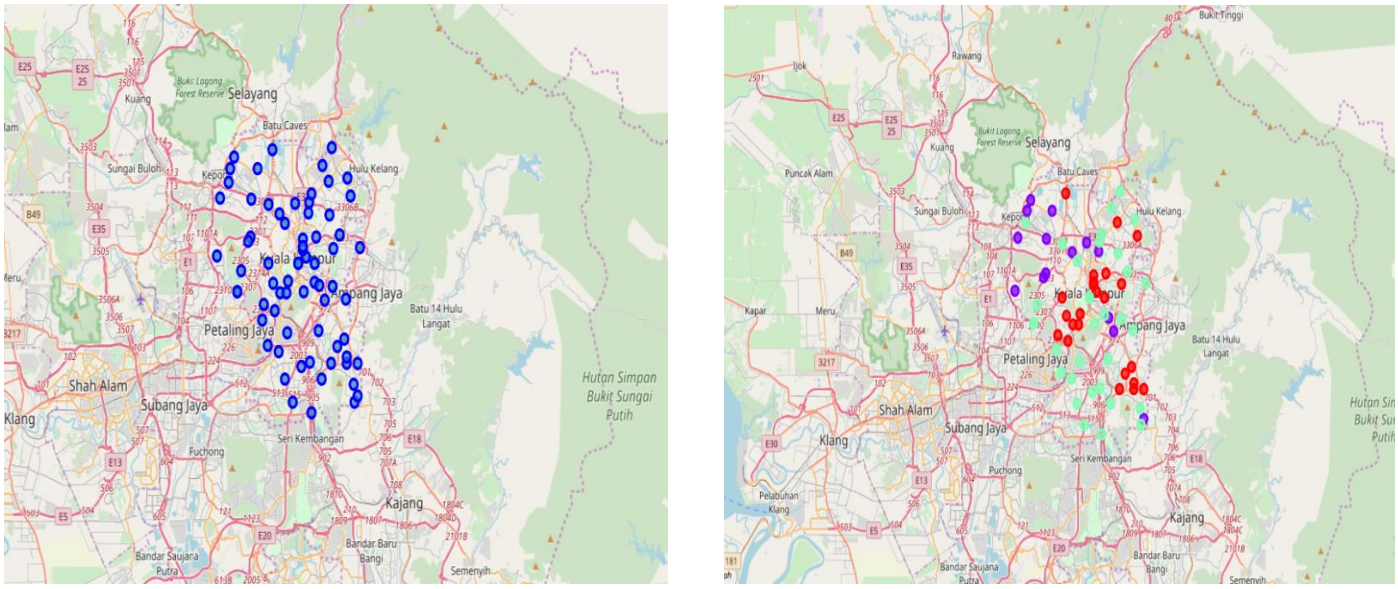


Figure 2. Feature correlation heatmap (425 features). Violet means no correlation, whereas yellow means high correlation



(a) ***Before Clustering*** ***(b) After Clustering***

Figure 9. Kuala Lumpur Zip codes with the assigned cluster label

DISCUSSION AND CONCLUTION

Most of the restaurant mall are concentrated in the central area of Kuala Lumpur city, with the highest number in cluster 2 and moderate number in cluster 0. On the other hand, cluster 1 has very low number to totally no restaurant in the neighborhoods. This represents a great opportunity and high potential areas to open new restaurant as there is very little to no competition from existing restaurant. Meanwhile, shopping malls in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of restaurant. From another perspective, this also shows that the oversupply of restaurant mostly happened in the central area of the city, with the suburb area still have very few restaurant. Therefore, this project recommends property developers to capitalize on these findings to open new restaurant in neighborhoods in cluster 1 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new restaurant in neighborhoods in cluster 0 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of restaurant and suffering from intense competition.

CONCLUSION

Answer to business question: The neighborhoods in cluster 1 are the most prefered locations to open a new restaurant. Findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new restaurant