# IBM Data Science Capstone Project – Report

## Introduction/Business Problem

In this project, I plan to identify the ideal area to open a Malaysian restaurant in San Diego, California. I have decided on this problem because despite the number of Malaysians who live in San Diego, there has not been a Malaysian restaurant opened in the area. I believe that opening one would be a good idea, and this project will identify where in San Diego would be best.

## Data

In order to find the ideal area to open up the Malaysian restaurant, I first looked up some of the most popular Malaysian/Southeast Asian restaurants in the United States. The restaurants chosen were **Kopitiam** in New York City, **Lukshon** in Culver City, California, **Kedai Makan** in Seattle, Washington, and **Pok Pok** in Portland, Oregon. Analyzing the neighbourhoods/areas that these already successful restaurants are located in helped decide which area in San Diego was most similar to those.

This project utilized the **Foursquare** API to explore venues in the specified areas. It was used to analyze the neighbourhoods of the already successful restaurants, as well as explore the cities in the San Diego county to find the city that is most similar.

Once the locations of the restaurants were found, the Foursquare API gave us a list of the venues around each restaurant, allowing us to analyze these neighbourhoods. Pandas dataframes were used to store and analyze the data. After collecting and cleaning the data, dataframes such as the table below were created to contain the information of the venues.

Table 1. First 5 rows of the dataframe containing venues around Lukshon

| | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|
| 0 | Lukshon | 34.029960 | -118.384247 | Asian Restaurant |
| 1 | Room & Board | 34.030223 | -118.384531 | Furniture / Home Store |
| 2 | Arcana: Books on the Arts | 34.030146 | -118.383500 | Bookstore |
| 3 | The COOLHAUS Shop | 34.030329 | -118.381487 | Ice Cream Shop |
| 4 | H.D. Buttercup L.P. | 34.030752 | -118.385067 | Furniture / Home Store |

From these dataframes, the **Folium** Python package was used to create maps showing the location and distribution of the venues around the restaurants.

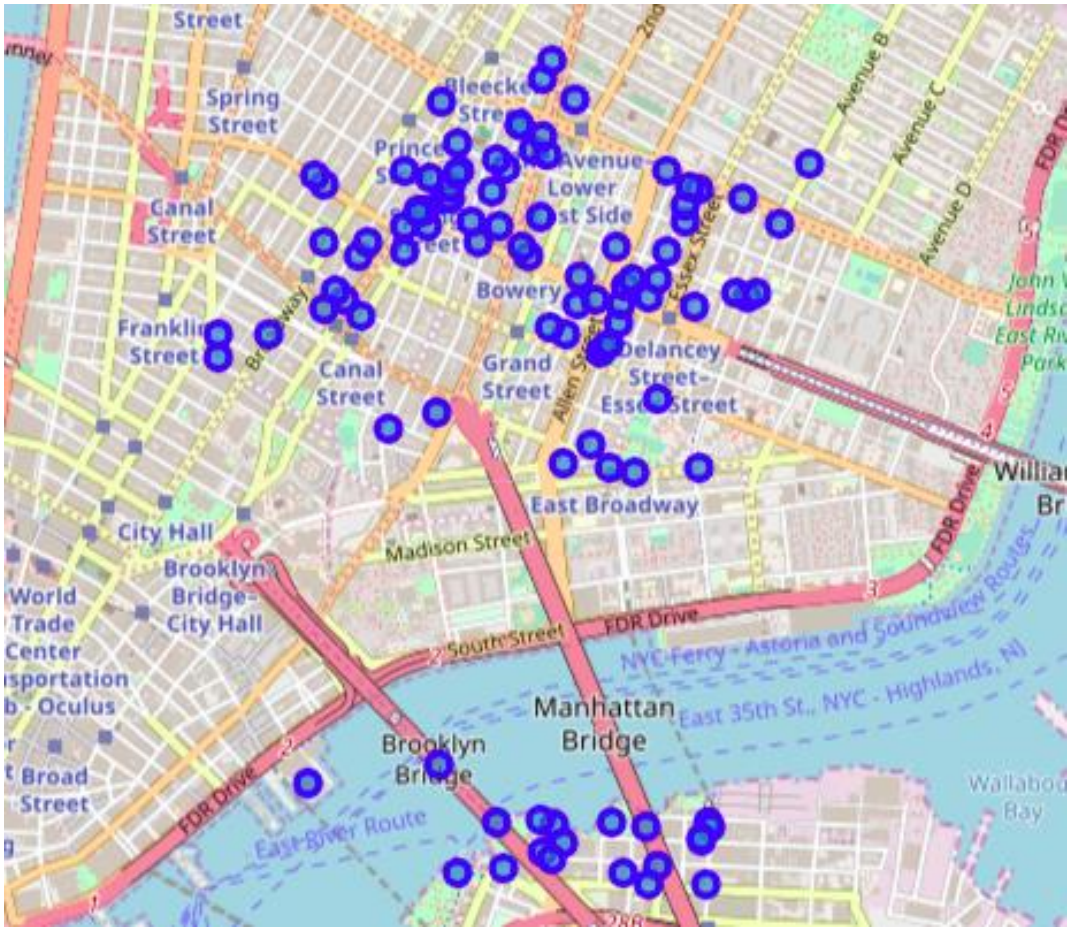

Figure 1. Markers showing the locations of venues around Kopitiam in New York City

The Foursquare API was also used to get the venues in the different cities of the San Diego county. Similar dataframes were created, and a map of San Diego was created, showing the regions which would be explored.

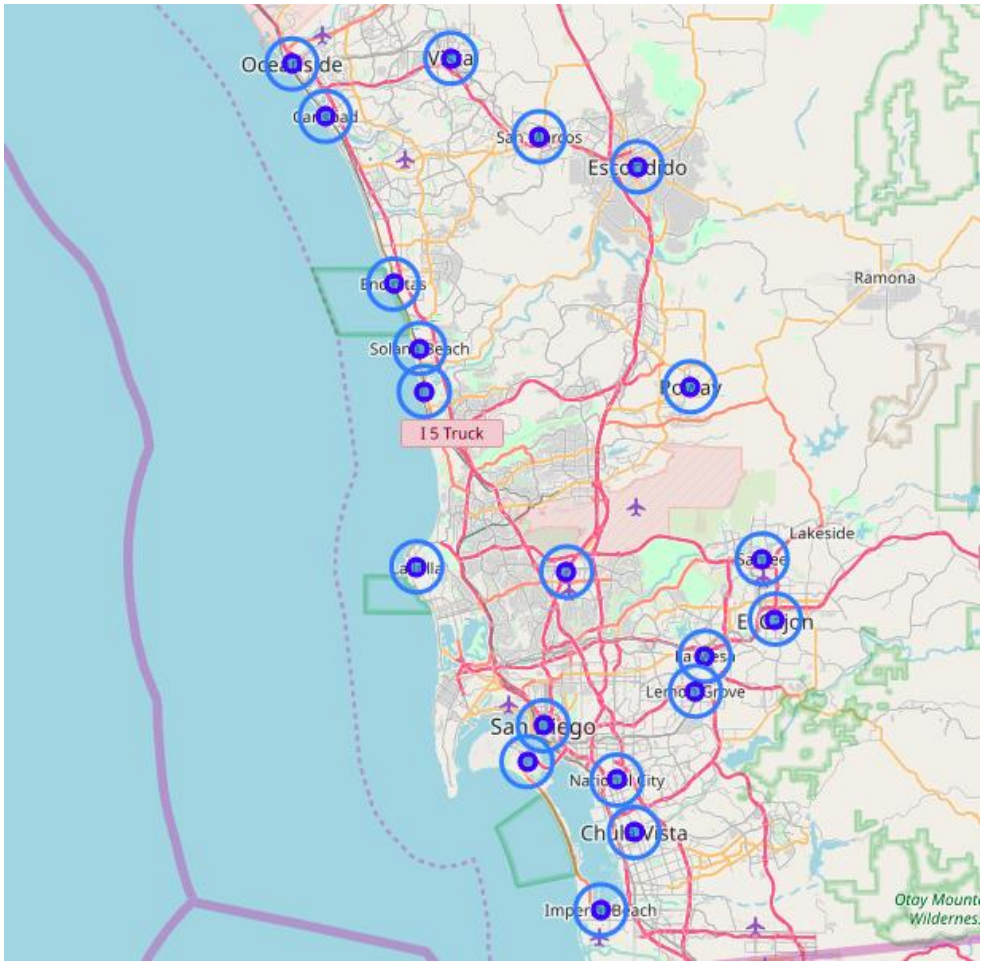


Figure 2. Map showing the centers of cities of the San Diego county, with a 2000 meter radius around each center.

## Methodology

Now that we have most of the data that we need, we can start analyzing the neighbourhoods. We have the data on the venues around some of the most popular/highly rated Malaysian/Southeast Asian restaurants in the United States. We looked at venues within a 2000

meter radius, because that is the radius that we will be using when analyzing the cities of San Diego as well, to reduce overlap of the cities.

Now, using the data that we have collected, we will analyze the different categories of venues that are around the restaurants. Once we have found the different frequencies of the categories for each location, we will take the average of that to come up with our *"comparison guideline"*. We will be using this to compare with the cities in San Diego, finding the city which is most similar to this guideline. In this project, we define *most similar* simply as the city whose frequencies of categories of venues differ the least.

Then, once we have completed the analysis of the cities of San Diego, we will perform **k-means clustering** to cluster the cities of San Diego. We perform this clustering to find the cities that are similar to the one that we found to be the closest to our comparison guideline, giving us a few options for cities to open our restaurant in.

## Analysis

We perform **one hot encoding** to analyze the frequencies of venue categories. We find the most frequent venue categories for each location, and from the frequencies of venues around the 4 already established restaurants, we get an aggregate of the frequencies, with which we will use to compare with the cities of San Diego.

Table 2. The 10 most frequent venue categories in the aggregate created from all 4 locations

```
----Grouped----
                      venue  freq
0               Coffee Shop  0.08
1                      Café  0.03
2                   Theater  0.02
3               Cocktail Bar  0.02
4                     Hotel  0.02
5           Ice Cream Shop  0.02
6       Italian Restaurant  0.02
7                 Juice Bar  0.02
8       Mexican Restaurant  0.02
9  New American Restaurant  0.02
```

The same analysis was done on the venues in the different cities of the San Diego county. The table below shows some of the frequencies for a few of the cities.

Table 3. Frequencies of venue categories for a few cities in San Diego County

| | Neighbourhood | ATM | Accessories Store | American Restaurant | Antique Shop | Arcade | Art Gallery | Arts & Crafts Store | Asian Restaurant | Athletics & Sports | ... | Vegetarian / Vegan Restaurant | Video Game Store | Video Store |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | San Diego | 0.0 | 0.000000 | 0.040000 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.0 | ... | 0.0 | 0.000000 | 0.000000 |
| 1 | Carlsbad | 0.0 | 0.000000 | 0.044444 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.011111 | 0.0 | ... | 0.0 | 0.000000 | 0.000000 |
| 2 | Chula Vista | 0.0 | 0.000000 | 0.022472 | 0.0 | 0.011236 | 0.0 | 0.0 | 0.000000 | 0.0 | ... | 0.0 | 0.011236 | 0.011236 |
| 3 | Coronado | 0.0 | 0.000000 | 0.000000 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.010870 | 0.0 | ... | 0.0 | 0.000000 | 0.000000 |
| 4 | Del Mar | 0.0 | 0.015385 | 0.076923 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.0 | ... | 0.0 | 0.000000 | 0.000000 |

The frequencies of each city in San Diego were compared with the grouped aggregate of the 4 established locations. The comparison was done simply by calculating the absolute values of the differences of the grouped frequency and the frequency of the same venue category in each city. The following graph shows the calculated differences.

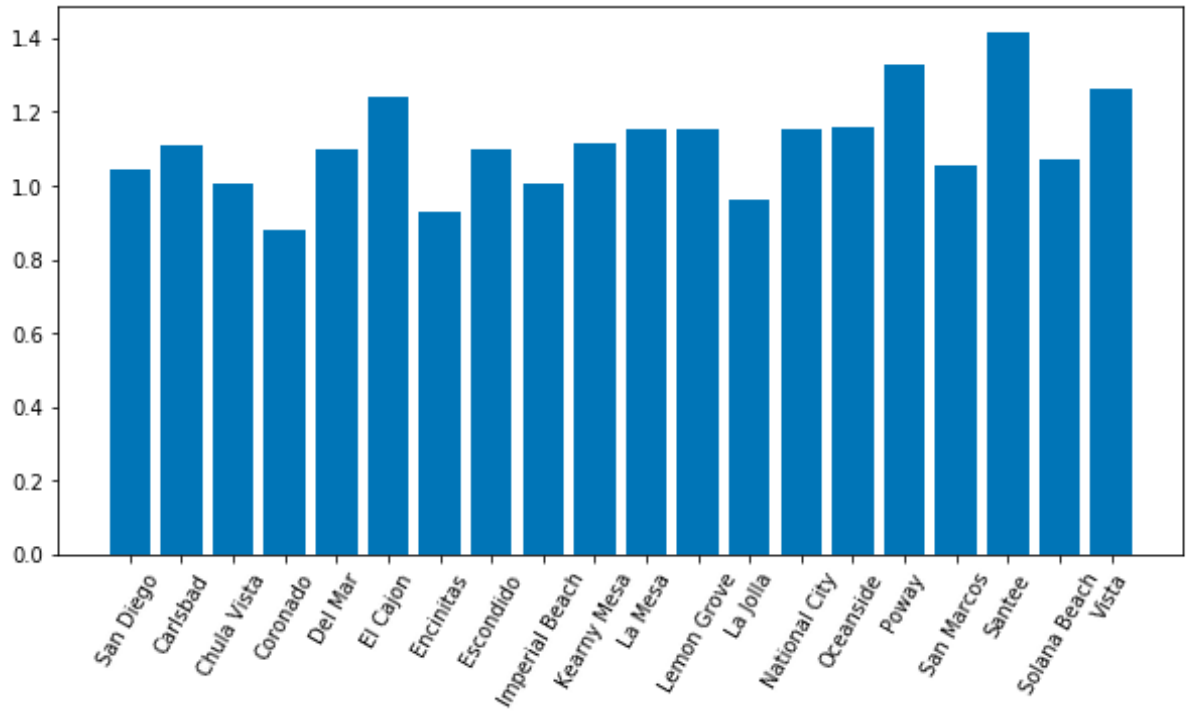


Figure 3. Bar chart of differences between cities of San Diego and calculated grouped frequency

As can be seen in the bar chart, **Coronado** was determined to be the city with closest venue category frequencies to the aggregate. Following this**, k-means clustering** was used to cluster the cities in San Diego, in order to find the cities in the same cluster as Coronado, to provide additional options of possible locations for the restaurant.

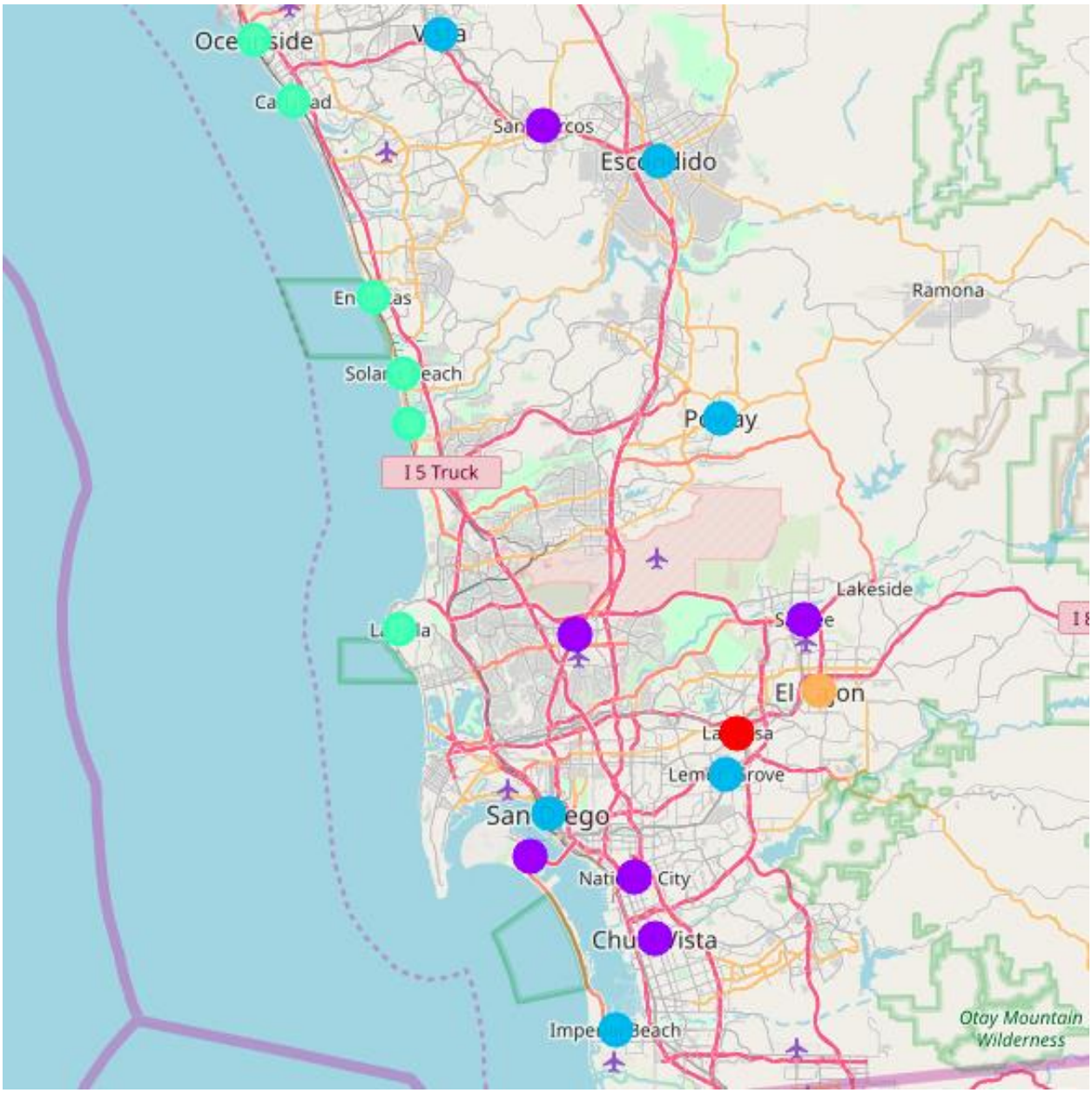


Figure 4. Map showing clustering of cities, with different colored markers indicating clusters

Table 4. Cluster containing Coronado and its similar cities

| | Neighbourhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Chula Vista | Mexican Restaurant | Grocery Store | Seafood Restaurant | Sandwich Place | Taco Place | Hotel | Pizza Place | Burger Joint | Pharmacy | Brewery |
| 1 | Coronado | Park | Seafood Restaurant | Hotel | Ice Cream Shop | Mexican Restaurant | Burger Joint | Pizza Place | Surf Spot | Italian Restaurant | Resort |
| 2 | Kearny Mesa | Japanese Restaurant | Sushi Restaurant | Bubble Tea Shop | Korean Restaurant | Mexican Restaurant | Sandwich Place | Vietnamese Restaurant | Chinese Restaurant | Noodle House | Ramen Restaurant |
| 3 | National City | Mexican Restaurant | Chinese Restaurant | Filipino Restaurant | Pizza Place | Park | Bakery | Sushi Restaurant | Auto Dealership | Tea Room | Pub |
| 4 | San Marcos | Mexican Restaurant | Pizza Place | Sushi Restaurant | Brewery | Baseball Field | Deli / Bodega | Gym / Fitness Center | Park | Diner | Hotel |
| 5 | Santee | Mexican Restaurant | Brewery | Breakfast Spot | Sushi Restaurant | Japanese Restaurant | Vietnamese Restaurant | Shoe Store | Deli / Bodega | Gymnastics Gym | American Restaurant |

## Results and Discussion

It's interesting to see that the cities in this cluster tend to have a lot of different restaurant categories as their most common venues. Taking a closer look, we can also see that a good amount of the restaurants are Asian restaurants, similar to Malaysian restaurants. Even more interestingly, this is more apparent in **Kearny Mesa**, where all of its top 10 most common venues are food or drink venues, with 8 out of 10 of them being Asian themed venues.

This raises an interesting discussion to go into deeper analysis to find which out of the 6 cities would be the most ideal location. Perhaps it would be impossible to determine the answer to that question without gathering other data, such as the ratings and popularities of other venues and restaurants in those areas, the demographic in those areas as well as their general preferences in food. Maybe some areas are closer to university campuses with international students, which would make it more appealing for them to travel to the restaurant. Also, would it be best to open the restaurant in an area that is already saturated with restaurants - specifically Asian restaurants - or one that is more sparse of restaurants. No matter what the actual ideal location is, our analysis of the data available to us has given us great insights, and has already given us some very good choices of location.

## Conclusion

Our aim of this project was the find the ideal location to open a Malaysian restaurant in San Diego. We did this by first looking at some already popular and highly rated Malaysian restaurants in other parts of the United States, and analyzed the locations that they are in. We used this data to then compare it to cities in San Diego to find the ideal locations, in the end giving us 6 possible cities: Chula Vista, Coronado, Kearny Mesa, National City, San Marcos, and Santee. These cities were determined by finding the city that was closest to an aggregate of the locations of the already established restaurants, which we found to be Coronado, and then performing clustering to find the other cities that are similar to Coronado.

This study has been insightful, giving good choices for potential locations. In order to narrow down our selection to one ideal city/location would require additional data such as the demographic data, success of restaurants and other venues in those locations, preferences of people who frequent the locations, etc.