

Comparing Universities

Introduction/Business Problem:

Malaysia is one of the top countries that receive big number of new coming students every year from all over the world. By 2019 there were 130,110 international students in Malaysia from 136 countries and Gross Enrolment Ratio in 2016 of 44% is higher than most of the Asian countries, and higher than the world average of 37%.

As an international student in Malaysia for 3 years I would try to build a solution for the problem I had and most students are having choosing a university and here is the problem:

In the university campus you can find restaurants but they serve only Malay food although some of the university has international food but it is not that quality, so most of students would go outside for eating or order food from outside using grab food app. Same goes for other facilities like barbershops, universities don't have these ones inside the campus.

The solution I would like to build would solve the problem from 2 aspects:

- Give the students a clue about the area around the universities so that they can choose the one with more facilities around and easy to be reached or order services from. This would be an area of 5km around the campus, which is a very good area for service delivery, or to go to by yourself, since this will cost like 2.5USD using Grab car or even 20 cents using bus
- Give people who are interested in business a good idea about the area around the university, so that they can choose the best business to invest their money in, since there will be so many students coming to these places because it is near to the campus, so a stakeholder may find a business that complete the missing item in the area. Let's say we have: 2 restaurants, 1 grocery shop, but we don't have a barbershop in that area, so having one would be a good idea.

Data:

We are going to have 2 datasets merged in one dataset.

We have a Wikipedia page that has 2 tables of public and private universities in Malaysia provided with the: name in English, name in Malay, Acronym, Foundation year, Location, link to the university website.

These 2 tables will be scraped using Pandas in python.

The picture below shows a sample of one of these 2 tables:

Universities [edit]					
Name in English	Name in Malay	Acronym	Foundation	Location	Link
University of Malaya	Universiti Malaya ^[8]	UM	1905	Kuala Lumpur Nilam Puri, Kelantan	[1]↗
International Islamic University Malaysia	Universiti Islam Antarabangsa Malaysia ^[9]	IIUM	1983	Gombak, Selangor	[2]↗
National University of Malaysia	Universiti Kebangsaan Malaysia ^[10]	UKM	1970	Bangi, Selangor	[3]↗
University of Malaysia Kelantan	Universiti Malaysia Kelantan ^[11]	UMK	2007	Pengkalan Chepa, Kelantan Jeli, Kelantan Bachok, Kelantan	[4]↗
University of Malaysia Pahang	Universiti Malaysia Pahang ^[12]	UMP	2002 (as KUKTEM)	Pekan, Pahang	[5]↗
University of Malaysia Perlis	Universiti Malaysia Perlis ^[13]	UniMAP	2001 (as KUKUM)	Arau, Perlis	[6]↗
University of Malaysia Sabah	Universiti Malaysia Sabah ^[14]	UMS	1994	Kota Kinabalu, Sabah	[7]↗
University of Malaysia Sarawak	Universiti Malaysia Sarawak ^[15]	UNIMAS	1992	Kota Samarahan, Sarawak	[8]↗

We will also be using Foursquare calls that will return to us a JSON file that we can read it as a panda data frame, in these calls we are going to focus on “explore” call and the category of each venue around the campus in 5km radius.

Data was scraped from the following link:

https://en.wikipedia.org/wiki/List_of_universities_in_Malaysia

Methodology:

- **Data wrangling:** if we look at the data in the above link we will notice that the website gives the data in 2 tables: Public universities and private universities, and each one of these tables has two sub-tables in it, so the methodology that I follow was to read the first main table “Public universities” table into Pandas data frames then split it into its 2 sub-tables and do the cleaning then merge then into their original data frames but in a consistent way, then I did the same thing to the second table (split – clean – merge), so now I will discuss the cleaning steps I made during data wrangling phase:
 - o Remove the unwanted rows, and these are the rows that have values not related to universities like the picture below:

0	Universities[edit]	Universities[edit]	Universities[edit]	Universities[edit]	Universities[edit]	Universities[edit]
1	Name in English	Name in Malay	Acronym	Foundation	Location	Link
2	University of Malaya	Universiti Malaya[8]	UM	1905	Kuala Lumpur Nilam Puri, Kelantan	[1]
3	International Islamic University Malaysia	Universiti Islam Antarabangsa Malaysia[9]	IIUM	1983	Gombak, Selangor	[2]
4	National University of Malaysia	Universiti Kebangsaan Malaysia[10]	UKM	1970	Bangi, Selangor	[3]

as you can see, the headers of the original tables were imported as rows’ values

- o The next step was to give the columns meaningful names because they were just numbers:

	0	1	2	3	4	5
2	University of Malaya	Universiti Malaya[8]	UM	1905	Kuala Lumpur Nilam Puri, Kelantan	[1]
3	International Islamic University Malaysia	Universiti Islam Antarabangsa Malaysia[9]	IIUM	1983	Gombak, Selangor	[2]
4	National University of Malaysia	Universiti Kebangsaan Malaysia[10]	UKM	1970	Bangi, Selangor	[3]
5	University of Malaysia Kelantan	Universiti Malaysia Kelantan[11]	UMK	2007	Pengkalan Chepa, Kelantan Jeli, Kelantan Bacho...	[4]

- o Then we edit the name in each cell of name so that we remove the brackets and numbers

- Next I remove the unwanted columns, and I chose the features to be: MalayName , Acronym , Location. I chose the Malay name for the sake of consistency with other tables , because other ones have only Malay name without English name, the acronym was selected also because some universities can be found by its acronym
- These were the steps for the first sub-tables of the first table then I repeated the same steps for the second sub-table and I merged them together in one table, I also edited the index to start from 1 , and this was the result after cleaning:

	MalayName	Acr	Loc
1	Universiti Malaya	UM	Kuala Lumpur Nilam Puri, Kelantan
2	Universiti Islam Antarabangsa Malaysia	IIUM	Gombak, Selangor
3	Universiti Kebangsaan Malaysia	UKM	Bangi, Selangor
4	Universiti Malaysia Kelantan	UMK	Pengkalan Chepa, Kelantan Jeli, Kelantan Bacho...
5	Universiti Malaysia Pahang	UMP	Pekan, Pahang
6	Universiti Malaysia Perlis	UniMAP	Arau, Perlis
7	Universiti Malaysia Sabah	UMS	Kota Kinabalu, Sabah
8	Universiti Malaysia Sarawak	UNIMAS	Kota Samarahan, Sarawak
9	Universiti Malaysia Terengganu	UMT	Kuala Terengganu, Terengganu
10	Universiti Pendidikan Sultan Idris	UPSI	Tanjung Malim, Perak
11	Universiti Pertahanan Nasional Malaysia	UPNM	Kuala Lumpur
12	Universiti Putra Malaysia	UPM	Serdang, Selangor Bintulu, Sarawak
13	Universiti Sains Islam Malaysia	USIM	Nilai, Negeri Sembilan
14	Universiti Sains Malaysia	USM	George Town, Penang
15	Universiti Sultan Zainal Abidin	UniSZA	Kuala Terengganu, Terengganu
16	Universiti Teknikal Malaysia Melaka	UTeM	Durian Tunggal, Malacca
17	Universiti Teknologi Malaysia	UTM	Skudai, Johor Jalan Semarak, Kuala Lumpur
18	Universiti Teknologi MARA	UiTM	Shah Alam, Selangor
19	Universiti Tun Hussein Onn Malaysia	UTHM	Batu Pahat, Johor

- I repeated the same steps for the second table with its 2 sub-tables and after cleaning it I merged with the previous tables then we had one data frame consists of 121 rows and 3 columns
- **Data Expanding:** Now that we have our dataset cleaned, we need to prepare the data to be used by foursquare APIs:
 - First of all, we add new 2 columns to the dataset and these are: latitude and longitude information for each row or university:


```
unis["lat"] = float("NaN")
unis["long"] = float("NaN")
unis.astype({'lat': 'float64'} , inplace=True).dtypes
unis.astype({'long': 'float64'},inplace = True).dtypes
```

- We used “geolocator” package to retrieve the coordinates of each university and we tried it first on one of the universities to check the results:

```
#testign the function on one uni
geolocator = Nominatim(user_agent="kk")
location = geolocator.geocode("Universiti Malaysia Sarawak")
print(location.address)
print(location.latitude, location.longitude)
```

Universiti Malaysia Sarawak, Jalan Datuk Mohamad Musa, Taman Samarindah, Kampung Mangka, Sarawak, 94300, Malaysia
1.4648777 110.4270042

however, there were some issues here, for example when we pass the location parameter to the geolocator it should be without ‘,’ , instead there should be a space. The other issue was that not always the combination of the university name and its location would give u back the coordinates, instead it will give back a None value, in this case we need to try different combination of address, like using the university name only, or the university name with the first part of address, or using the university acronym.. etc. However, we handled this problem using two type of combination only and in case there was a Value of None returned back by the locator then we will drop the row of that specific university. Following this methodology gave us very good results, it returned back 103 coordinates out of 121:

Looping to get the coordinates:

```
i = 0
for index , data in unis.iterrows():
    loc = data["Loc"].replace(',',' ')
    add = data['MalayName'] + " " + loc
    location = geolocator.geocode(data['MalayName'])
    if (location == None):
        location = geolocator.geocode(loc)
        if (location == None):
            unis.drop(index , axis=0 , inplace=True )
            unis.reset_index()
            continue

    print(location.latitude , location.longitude )
    unis.iloc[index , unis.columns.get_loc('lat')] = location.latitude
    unis.iloc[index , unis.columns.get_loc('long')] = location.longitude

    i += 1
print(i)
```

Checking the number of universities that are able to be used in foursquare:

```
In [73]: unis.shape
```

```
Out[73]: (103, 5)
```

The final data frame:

	MalayName	Acr	Loc	lat	long
0	Universiti Malaya	UM	Kuala Lumpur Nilam Puri, Kelantan	3.122674	101.653561
1	Universiti Islam Antarabangsa Malaysia	IIUM	Gombak, Selangor	3.253190	101.735714
2	Universiti Kebangsaan Malaysia	UKM	Bangi, Selangor	2.924087	101.781385
3	Universiti Malaysia Kelantan	UMK	Pengkalan Chepa, Kelantan Jeli, Kelantan Bacho...	5.995766	102.402764
4	Universiti Malaysia Pahang	UMP	Pekan, Pahang	3.722724	103.122982
5	Universiti Malaysia Perlis	UniMAP	Arau, Perlis	6.447307	100.244720
6	Universiti Malaysia Sabah	UMS	Kota Kinabalu, Sabah	6.038435	116.120267
7	Universiti Malaysia Sarawak	UNIMAS	Kota Samarahan, Sarawak	1.464878	110.427004
8	Universiti Malaysia Terengganu	UMT	Kuala Terengganu, Terengganu	5.409421	103.088738

- **Foursquare jobs:** so here we will start to retrieve the venues around each university for 5km radius around.
 - o First, I set up my API call credentials
 - o Then I defined one function that goes into the dataset and take each university and fetch the venues around it, limited to 100 venues, as it is shown below:

```
def getNearbyVenues(names, latitudes, longitudes, radius=5000):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)

        # create the API request URL
        url = "https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&v={}&ll={}&radius={}&limit={}".format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)

        # make the GET request
        results = requests.get(url).json()["response"]["groups"][0]["items"]

        # return only relevant information for each nearby venue
        venues_list.append([
            name,
            lat,
            lng,
            v["venue"]["name"],
            v["venue"]["location"]["lat"],
            v["venue"]["location"]["lng"],
            v["venue"]["categories"][0]["name"] for v in results])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    nearby_venues.columns = ['Neighborhood',
                             'Neighborhood Latitude',
                             'Neighborhood Longitude',
                             'Venue',
                             'Venue Latitude',
                             'Venue Longitude',
                             'Venue Category']

    return(nearby_venues)
```

- o Then we call the function on the dataset, and we got a new dataset that contains universities with its 100 venues around:

```
: uni_venues.shape
```

```
76]: (7945, 7)
```

After doing group by function on the dataset to see a summary of this new dataset:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
	Kolej Universiti Bandar Utama	29	29	29	29	29	29
	Kolej Universiti Islam Antarabangsa Selangor	100	100	100	100	100	100
	Kolej Universiti Islam Melaka	100	100	100	100	100	100
	Kolej Universiti Komunikasi Han Chiang	100	100	100	100	100	100
	Kolej Universiti Linton	5	5	5	5	5	5
	Kolej Universiti New Era	20	20	20	20	20	20
	Kolej Universiti Saito	100	100	100	100	100	100
	Kolej Universiti TATI	90	90	90	90	90	90
	Kolej Universiti Teknologi Sarawak	39	39	39	39	39	39
	Kolej Universiti Widad	100	100	100	100	100	100
	Kolej Universiti Yayasan Sabah	100	100	100	100	100	100

- After performing one hot encoding and data frame transformations we can get each university with its venues ranked from 1st to 10th place, here we can see an example of this dataset:

	University	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	Kolej Universiti Bandar Utama	Malay Restaurant	Boat or Ferry	Beach	Food Truck	Pier	Shopping Mall	Hotel	Buffet	Laundromat	Surf Spot
1	Kolej Universiti Islam Antarabangsa Selangor	Café	Hotel	Dessert Shop	Chinese Restaurant	Bakery	Vegetarian / Vegan Restaurant	Food Truck	Noodle House	Public Art	Coffee Shop
2	Kolej Universiti Islam Melaka	Malay Restaurant	Café	Ice Cream Shop	Dessert Shop	Convenience Store	Hotel	Chinese Restaurant	Coffee Shop	Shopping Mall	Supermarket
3	Kolej Universiti Komunikasi Han Chiang	Hotel	Café	Chinese Restaurant	Shopping Mall	Boutique	Indian Restaurant	Korean Restaurant	Middle Eastern Restaurant	Bar	Spa
4	Kolej Universiti Linton	French Restaurant	Breakfast Spot	African Restaurant	Airport	Hotel	Yoga Studio	Flower Shop	Field	Filipino Restaurant	Fish & Chips Shop

- **Clusters and Machine learning:** So now we need to build I model that looks into the data and finds out the similarities between each of the universities depending on their near venues and group the similar ones in a group:
 - We are using K-means algorithm with 5 clusters then are fitting the model with dataset and getting the cluster for each row, then appending these cluster labels to each row in the dataset:

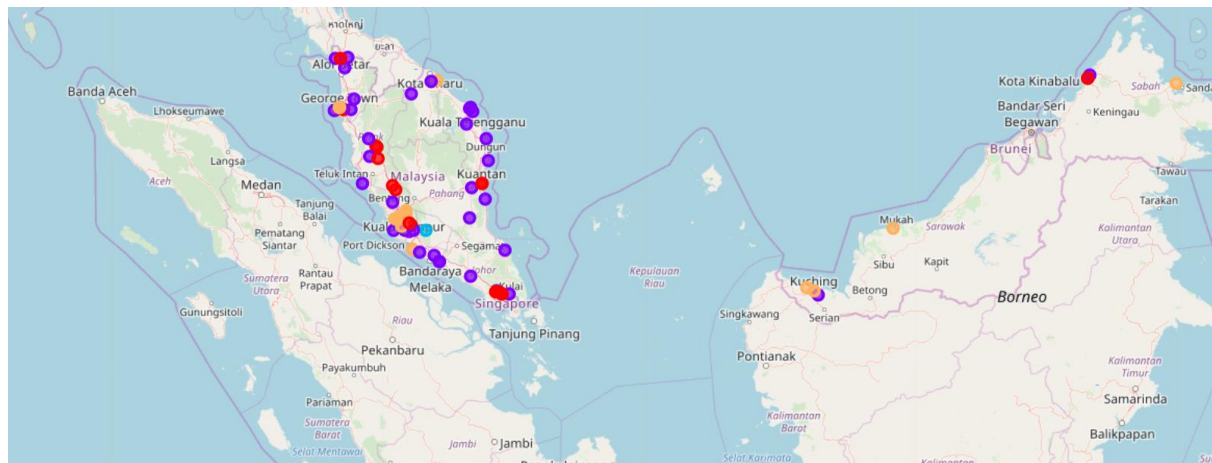
	MalayName	Acr	Loc	lat	long	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Universiti Malaya	UM	Kuala Lumpur Nilam Puri, Kelantan	3.122674	101.653561	2	Café	Indian Restaurant	Malay Restaurant	Shopping Mall	Ice Cream Shop	Convenience Store	Coffee Shop	Hotel	Spa	Steakhouse
1	Universiti Islam Antarabangsa Malaysia	IUM	Gombak, Selangor	3.253190	101.735714	2	Malay Restaurant	Indonesian Restaurant	Café	Sandwich Place	Asian Restaurant	Burger Joint	Thai Restaurant	Gym	Spa	Middle Eastern Restaurant
2	Universiti Kebangsaan Malaysia	UKM	Bangi, Selangor	2.924087	101.781385	2	Malay Restaurant	Coffee Shop	Restaurant	Asian Restaurant	Café	Burger Joint	Hotel	Soccer Field	Indonesian Restaurant	Japanese Restaurant
3	Universiti Malaysia Kelantan	UMK	Pengkalan Chepa, Kelantan Jeli, Kelantan Bacho...	5.995766	102.402764	2	Beach	Restaurant	Grocery Store	Resort	Breakfast Spot	Soup Place	Soccer Field	Other Great Outdoors	Caribbean Restaurant	Fast Food Restaurant
4	Universiti Malaysia Pahang	UMP	Pekan, Pahang	3.722724	103.122982	1	Malay Restaurant	Chinese Restaurant	Light Rail Station	Thai Restaurant	Asian Restaurant	Breakfast Spot	Bus Station	Cafeteria	Metro Station	Baseball Field

- **Building the map:** Our job now is to locate these universities on a map showing their location and their clusters along with their names:
 - o Since we are looking for the universities in all Malaysia, we are going to build full Malaysia map, so we need to get the coordinates of Malaysia:

```
address = 'Malaysia Kuala lumpur'

geolocator = Nominatim(user_agent="ny_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
```

- o Then we start to build or Folium map using looping through the universities and their clusters, then shows group these universities by colouring them depending on their clusters, and this is the result we get on the map:



- **Exploring the clusters:** Now that we have 5 clusters, we need to explore each cluster of them so we can gain more understanding of the structure or properties:

○ Cluster 0:

```
uni_merged.loc[uni_merged['Cluster Labels']==0]
```

7]:

	MalayName	Acr	Loc	lat	long	Cluster Labels	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	8th Most Common Venue
5	Universiti Malaysia Perlis	UniMAP	Arau, Perlis	6.447307	100.244720	0	Malay Restaurant	Asian Restaurant	Café	Thai Restaurant	Convenience Store	Pizza Place	Burger Joint	Seafood Restaurant	Diner
7	Universiti Malaysia Sarawak	UNIMAS	Kota Samarahan, Sarawak	1.464878	110.427004	0	Malay Restaurant	Food Court	Café	Asian Restaurant	Fast Food Restaurant	Shopping Mall	Coffee Shop	Bakery	Clothing Store
8	Universiti Malaysia Terengganu	UMT	Kuala Terengganu, Terengganu	5.409421	103.088738	0	Malay Restaurant	Seafood Restaurant	Breakfast Spot	Café	Hotel	Beach	Fast Food Restaurant	Asian Restaurant	Restaurant
9	Universiti Pendidikan Sultan Idris	UPSI	Tanjung Malim, Perak	3.696201	101.519946	0	Malay Restaurant	Asian Restaurant	Café	Coffee Shop	Chinese Restaurant	Convenience Store	Food Court	Restaurant	Breakfast Spot
14	Universiti Sultan Zainal Abidin	UniSZA	Kuala Terengganu, Terengganu	5.396161	103.082711	0	Malay Restaurant	Seafood Restaurant	Breakfast Spot	Café	Fast Food Restaurant	Hotel	Restaurant	Asian Restaurant	Beach

○ Cluster 1:

```
uni_merged.loc[uni_merged['Cluster Labels']==1]
```

8]:

	MalayName	Acr	Loc	lat	long	Cluster Labels	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
4	Universiti Malaysia Pahang	UMP	Pekan, Pahang	3.722724	103.122982	1	Malay Restaurant	Chinese Restaurant	Light Rail Station	Thai Restaurant	Asian Restaurant	Breakfast Spot	Bus Station	Cafeteria	Metro Station	Baseball Field
6	Universiti Malaysia Sabah	UMS	Kota Kinabalu, Sabah	6.038435	116.120267	1	Malay Restaurant	Café	Chinese Restaurant	Asian Restaurant	Fast Food Restaurant	Indian Restaurant	Indonesian Restaurant	Restaurant	Coffee Shop	Halal Restaurant
10	Universiti Pertahanan Nasional Malaysia	UPNM	Kuala Lumpur	3.048244	101.724247	1	Chinese Restaurant	Coffee Shop	Asian Restaurant	Ice Cream Shop	Middle Eastern Restaurant	Café	Indian Restaurant	Noodle House	Burger Joint	Japanese Restaurant
24	Politeknik Sultan Abdul Halim Muadzam Shah	POLIMAS	Bandar Darul Aman, Kedah	4.588326	101.124961	1	Asian Restaurant	Bakery	Chinese Restaurant	Restaurant	Café	Coffee Shop	Malay Restaurant	Indian Restaurant	Ice Cream Shop	Hotel
28	Politeknik Kota Kinabalu	PKK	Kota Kinabalu, Sabah	1.631086	110.191285	1	Park	Trail	River	Asian Restaurant	Pool	Farm	Water Park	Lake	Snack Place	Swing

○ Cluster 2:

```
uni_merged.loc[uni_merged['Cluster Labels']==2]
```

9]:

	MalayName	Acr	Loc	lat	long	Cluster Labels	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
0	Universiti Malaya	UM	Kuala Lumpur Nilam Puri, Kelantan	3.122674	101.653561	2	Café	Indian Restaurant	Malay Restaurant	Shopping Mall	Ice Cream Shop	Convenience Store	Coffee Shop	Hotel	Sports Field
1	Universiti Islam Antarabangsa Malaysia	IUAM	Gombak, Selangor	3.253190	101.735714	2	Malay Restaurant	Indonesian Restaurant	Café	Sandwich Place	Asian Restaurant	Burger Joint	Thai Restaurant	Gym	Sports Field
2	Universiti Kebangsaan Malaysia	UKM	Bangi, Selangor	2.924087	101.781385	2	Malay Restaurant	Coffee Shop	Restaurant	Asian Restaurant	Café	Burger Joint	Hotel	Soccer Field	Indonesian Restaurant
3	Universiti Malaysia Kelantan	UMK	Pengkalan Chepa, Kelantan Jeli	5.995766	102.402764	2	Beach	Restaurant	Grocery Store	Resort	Breakfast Spot	Soup Place	Soccer Field	Other Great Outdoors	Caribbean Restaurant

○ Cluster 3:

```
uni_merged.loc[uni_merged['Cluster Labels']==3]
```

9]:

	MalayName	Acr	Loc	lat	long	Cluster Labels	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
43	Politeknik Jeli	PJK	Jeli, Kelantan	2.872797	112.022353	3	Athletics & Sports	Food & Drink Shop	Café	Beach	Yoga Studio	Fishing Store	Field	Filipino Restaurant	Fish & Chips Shop	Fish Market

○ Cluster 4:

```
uni_merged.loc[uni_merged['Cluster Labels']==4]
```

1]:

	MalayName	Acr	Loc	lat	long	Cluster Labels	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
46	Politeknik Mersing	PMJ	Mersing, Johor	2.840798	102.15408	4	Campground	River	Yoga Studio	Fishing Spot	Farm	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant	Fish & Chips Shop

Results: as we can see from the results above the first 3 clusters have many universities in them, but cluster 4 and 3 have only 1 university each, and we can see that these 2 last clusters don't have restaurants or cafés as their first 2 most common venues, instead they have entertainment places. For cluster 0 we notice that it's surrounded by Malay restaurants and Asian ones as the most, they have less places of international chains like pizza hut or Dominos. For cluster 2 we notice that they have restaurants with international dishes, like Indian and Indonesians dishes so this would be good for international students.

Observations: what I mainly observed during working with dataset on this type of project that Malaysia really needs to put more efforts on building datasets, because more of information and data were missing. My project would be more effective and wider if they have this type of dataset.

Conclusion: from the final results we can see that students who prefer international dishes to choose cluster 2. However, stakeholders who want to start international restaurants business, they would have high chance of success in cluster 3 and 4