Submitted by: Eric Ng Date: May 27, 2019

Applied Data Science Capstone Project – Week 2 Report

Discover "Michelin Quality" Restaurants in Hong Kong

I. Introduction and Background:

Michelin is probably one of the most influential companies in the restaurant industry in the world. Originally a company that sells tires, it started assigning ratings and awards to restaurants since 1926 in France, while putting together a guide which helps drivers find good food while they travel. The ratings are conducted by Michelin's own full-time food critics, who are passionate about food, have good eye for detail, and have a great taste memory to recall and compare types of foods. For many years now, the Michelin awards are most coveted by restaurant owners around the world, since being on the Michelin list will certainly bring prestige as well as more customers. The award system has changed over time, and currently the ratings are given as follows:

- **Bib gourmand**: A place that high-quality food and services and good value for the money. Bib is short for Bibendum the character of Michelin's logo.
- **One star:** A good place to stop on your journey, indicating a very good restaurant in its category, offering cuisine prepared to a consistently high standard.
- **Two stars:** A restaurant worth a detour, indicating excellent cuisine and skillfully and carefully crafted dishes of outstanding quality
- **Three stars:** A restaurant worth a special journey, indicating exceptional cuisine where diners eat extremely well, often superbly. Distinctive dishes are precisely executed, using superlative ingredients.

With the advance of mobile computing in the past 10 years, ratings to any restaurant can now be conducted by anyone conveniently through Apps such as Foursquare. Now that we have aggregated ratings and stats offered by these Apps, one may ask how the data would compare to those conducted by the Michelin judges. Is there opportunity to have the regular user data and the professional Michelin data leverage each other so it can help customers pick good restaurants?

II. Issue and Opportunity:

In every metropolitan city, there are thousands of restaurants. Being just one single company, Michelin is limited by its coverage. There are only so many restaurants a Michelin judge can go visit every year in a particular city, and as a result, many great restaurants may not ever get discovered or graded by Michelin. Now, data provided by the massive number of users that got aggregated by Foursquare may help. Foursquare covers thousands of restaurants in town and possess many stats. Based on these stats, we can try to discover the special characteristics of the restaurants that had

earned a Michelin award. After learning their special characteristics, we can then predict whether a non-Michelin restaurant possess "Michelin quality" too.

III. Objective – Discover "Michelin Quality" Restaurants

We would like to train a classification model that is able to identify between a "Michelin quality" and a "Not Michelin quality" restaurant. Using this trained model, we will then generate a list of restaurants that is currently not awarded by Michelin; yet has a high probability that their overall quality is comparable to those that have already earned the Michelin award.

We can then advise customers about this list, so they have more choices beyond the Michelin list. Alternatively, we can also provide this list to Michelin so it may send its food critics to these restaurants for a visit, just in case they missed them in their radar.

IV. City of choice and defining "Michelin Restaurants"

Being a Hong Kong citizen, I choose to use Hong Kong as the city of study, since I am more familiar with the restaurants and food being offered. I may also be able to make sense of the data and the result better. Hong Kong has a decent number of Michelin winners close to that of New York, so I shall have a workable sample size that helps me train the model.

Quantity of the Top 10 Most Michelin Starred Cities (including 1, 2, and 3 stars):

1. Tokyo, 230
2. Paris, 123
3. Kyoto, 103
4. Osaka, 97
5. New York, 76
6. London, 69
7. **Hong Kong, 63**
8. Singapore, 39
9. San Francisco, 38
10. Barcelona, 25

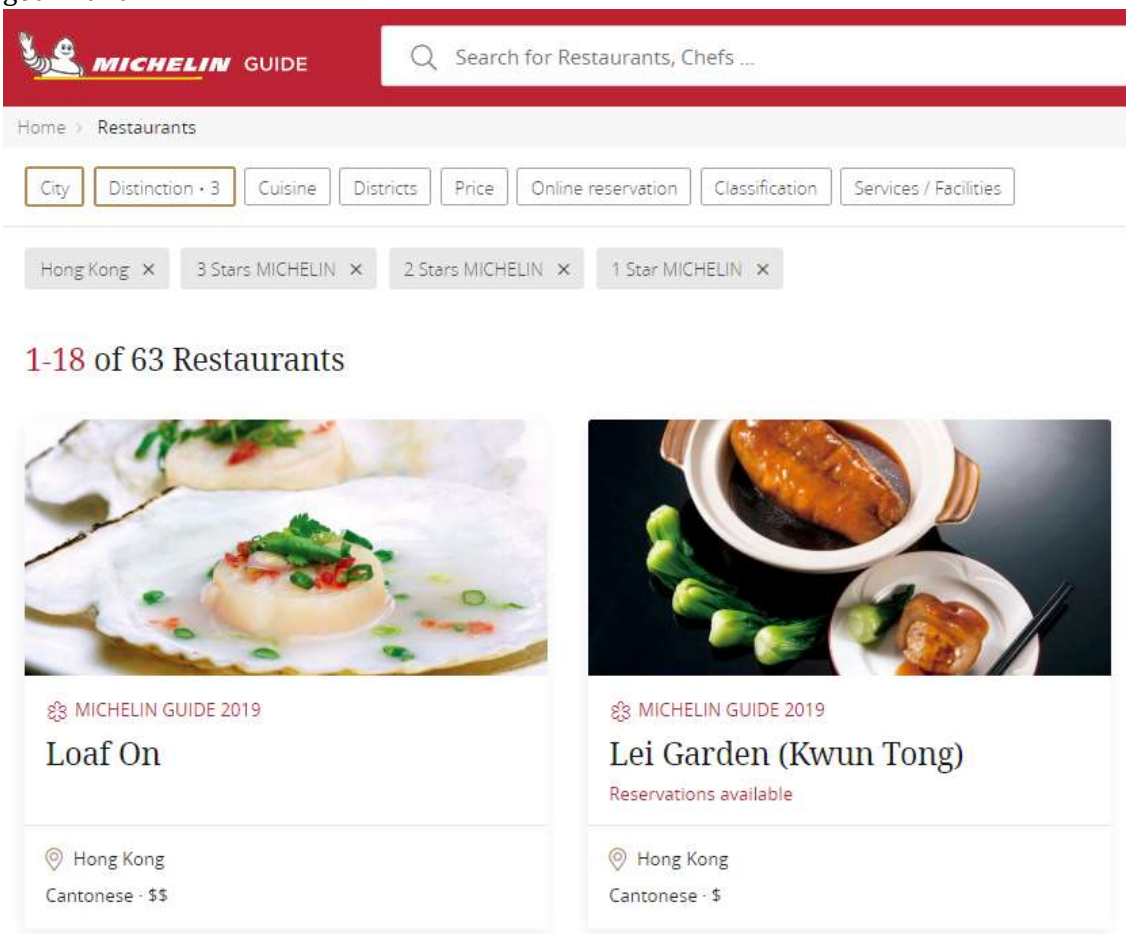
Besides the starred restaurants, we also add the 71 Bib gourment restaurants to the list, giving us a total of 134. The model is trained irrespective of whether the restaurants is 1 star, 2 star, 3 stars, or Bib Gourment. They are all grouped into the "Michelin Restaurants" category. These restaurants will be assigned with a value "1", while restaurants that did not earn the Michelin award will be assigned "0". The 1 and 0 will be the value of our "y", the dependent variable in our classification model (to be explained further).

V. Learn about Michelin restaurants

The data sourcing is straight forward, but the data cleaning is challenging especially when the Chinese language is involved.

1. Michelin winners data

   Starred restaurant data from Michelin website –
   https://guide.michelin.com/hk/en/hong-kong-region/hong-kong/restaurants/3-stars-michelin/2-stars-michelin/1-star-michelin
   Bib gourment restaurant data from Michelin website –
   https://guide.michelin.com/hk/en/hong-kong-region/hong-kong/restaurants/bib-gourmand



   The data we have obtained include these 2 most important info:

   a. Name of the winning restaurants
   b. Coordinates of the restaurants

   With these 2 pieces of data, we try to find the restaurants in Foursquare database, then learn about their venue ID's. Only when we have the venue ID's can we pull the ratings and stats from Foursquare.

   Below is the restaurant data scraped from Michelin's websites, a total of 134:

a. 63 Starred Michelin restaurants

(63, 5)

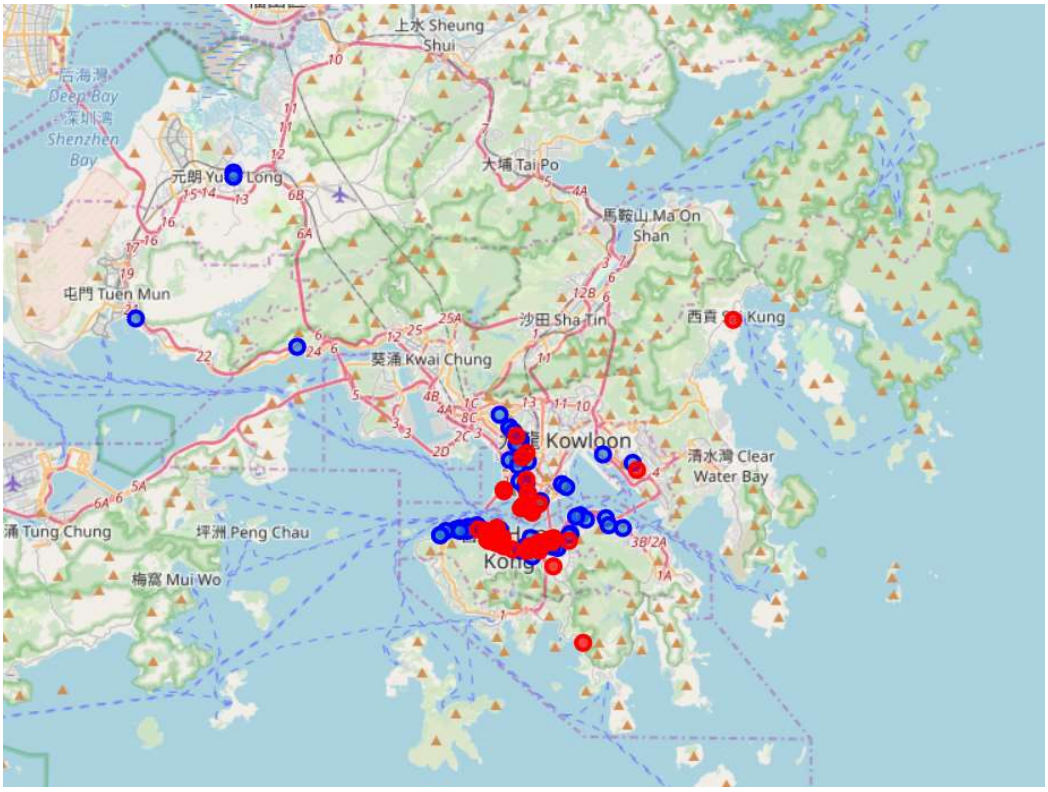
	Name	Cuisine	Latitude	Longitude	Award
0	Loaf On	Cantonese	22.3799	114.272	Star
1	Lei Garden (Kwun Tong)	Cantonese	22.3127	114.225	Star
2	Tim Ho Wan (Sham Shui Po)	Dim Sum	22.3281	114.167	Star
3	Lei Garden (Mong Kok)	Cantonese	22.3202	114.171	Star
4	Ming Court	Cantonese	22.3183	114.169	Star

b. 71 Bib Gourment restaurants

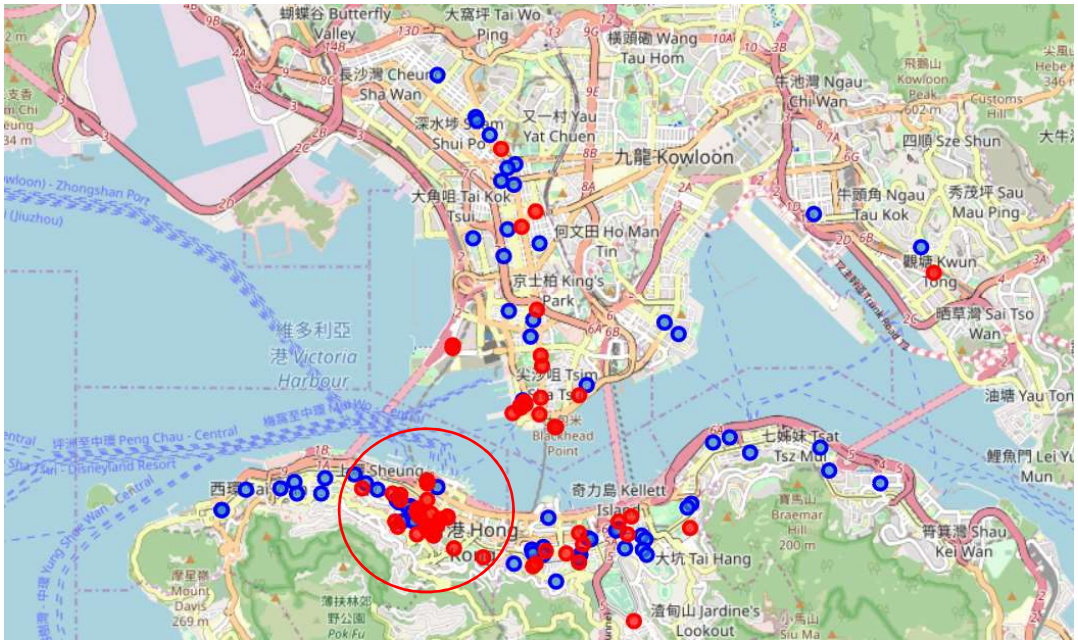
(71, 5)

	Name	Cuisine	Latitude	Longitude	Award
0	Tai Wing Wah	Cantonese	22.4462	114.029	Bib Gourment
1	Ho To Tai (Yuen Long)	Noodles	22.4438	114.029	Bib Gourment
2	Yue Kee	Cantonese	22.3682	114.06	Bib Gourment
3	Kwan Kee Bamboo Noodles (Cheun...	Noodles	22.3372	114.158	Bib Gourment
4	Lucky Indonesia	Indonesian	22.3158	114.223	Bib Gourment

For the audience's interest, here's how the Michelin awarded restaurants spread throughout Hong Kong. Red's are star restaurants, while Blue's are Bib restaurants.



The Michelin restaurants are concentrated in the Northern part of Hong Kong Island and Kowloon area. There is no big difference as to how Bib and Star restaurants are spread. Let's zoom in further the area where they are most concentrated.



The circled area is Central, Hong Kong's financial district, and this is also where one may find most concentration of Michelin restaurants. Below is a picture of Central provided by Michelin.

VI . Data sourcing and Pre-processing

a. Find Venue ID of Michelin Restaurants in the Foursquare database

The coordinates of the restaurants used by Michelin and Foursquare do not match. There are slight differences, and therefore, we will use the explore function in the Foursquare API to find the nearby venues using Michelin's coordinates, and hopefully find the matching restaurants.

(3145, 8)

	Neighborhood	Neighborhood_Latitude	Neighborhood_Longitude	Venue	Venue_Latitude	Venue_Longitude	Venue_ID
0	Loaf On	22.379858	114.27188	Loaf On (六福菜館)	22.379852	114.272226	4c1371bf583c9c74ee6d3fa4
1	Loaf On	22.379858	114.27188	The Bottle Shop	22.380181	114.272724	50b02fbbe4b0441cd8d87bfb
2	Loaf On	22.379858	114.27188	Sai Kung Market (西貢街市)	22.380541	114.272112	4c1db378fcf8c9b6b649ac0b
3	Loaf On	22.379858	114.27188	Little Cove Espresso	22.379613	114.271835	5641b617498e385faf1c93cb
4	Loaf On	22.379858	114.27188	AJ's Sri Lankan Cuisine (AJ's 斯里蘭卡菜館)	22.379363	114.271335	4b7a84d3f964a5209b2f2fe3

As you can see on the first line, the coordinates from Michelin (left side) is different from that of Foursquare (right side). The name is also not exactly the same. Upon further investigation, a straightforward name matching is not possible.

b. Name matching – manual work required due to language issues

In the Foursquare database, the names of the restaurant are mostly inconsistent with that of Michelin. Michelin uses English only for names, but in Foursquare, the language used is in consistent. The names can be:

a. English only
b. English + Chinese
c. Chinese only
d. Simplified Chinese and Traditional Chinese are also used interchangeably

Venue
Loaf On (六福菜館)
The Bottle Shop
Sai Kung Market (西貢街市)
Little Cove Espresso
AJ's Sri Lankan Cuisine (AJ's 斯里蘭卡菜館)

Because of this, the data matching had to be done by human manually on a spreadsheet instead of via python programming. Luckily there is only just 134 names to go through.

After matching the data, we uploaded file back into the program and call it 'Micheline_4Square_ID_HK.csv' . It is discovered that only 83 out of the 134 Michelin restaurants are listed or can be found in Foursquare. The reason is unknow, but since Foursquare is not as popular in Hong Kong as in the US, some restaurants may not be registered, or some may have changed their names or closed down recently.

And because of a smaller list with a smaller sample, it was decided that the Bib restaurants and Star restaurants will not be separately analyzed. It will be analyzed together as one category of "Michelin restaurants".

c. Find Venue ID of Non-Michelin restaurants

We also needed data from non-Michelin restaurants. To obtain their venue ID's, we leverage on the same API calls which we used to explore the coordinates provided by Michelin. We have screened out all restaurants nearby, from which we obtained a total 545 non-Michelin restaurants which we can use for model building.

(545, 5)

	Venue	Venue_Latitude	Venue_Longitude	Venue_ID
1	AJ's Sri Lankan Cuisine (AJ's 斯里蘭卡菜館)	22.379363	114.271335	4b7a84d3f964a5209b2f2fe3
2	Chuen Kee Seafood Restaurant (全記海鮮菜館)	22.380751	114.273345	4bb87186314e95214f72489d
3	Sing Kee Seafood Restaurant 勝記海鮮酒家	22.379198	114.270490	4c5beec37735c9b643cf8a72
4	CASA TAPAS BAR	22.381086	114.273251	520767dc11d2abbef5b4a8b9
5	Piccolos	22.381280	114.271950	4f1952b5e4b00583e7d64b04

d. Find ratings and stats of Michelin and Non-Michelin restaurants

Now that all venue ID's are available, we use the venue_ID API call to acquire details for each restaurant. Since we are using a free developer's package from Foursquare, we can only have visibility to limited type of data. For example, we are not able to get the no. of dislikes and no. of total check-ins. However, there are still a number of useful data we may use as the independent variables (x) for our model:

a. Ratings
b. Number of likes
c. Number of tips
d. Number of photos
e. Number listed (something like my favorite)
f. Price level (1, 2, 3 to choose from, where 3 being most expensive)
g. BeenHere counts
h. Number of Reasons

(640, 12)

	Name	Cuisine	Rating	Likes	Dislike	Tips	Photos	Listed	Price Level	BeenHere	Reasons	Area
0	Tam Chai Yunnan Noodles (譚仔雲南米線)	Chinese Restaurant	7.6	6	False	3	16	4	1	0	0	旺角
1	Market Hotpot (鮮入圍煮)	Hotpot Restaurant	7.3	16	False	6	89	28	0	0	0	香港
2	The Great Restaurant (一品雞煲火鍋)	Hotpot Restaurant	7.2	10	False	1	15	3	0	0	0	旺角
3	Peking Garden (北京樓)	Beijing Restaurant	7.5	30	False	12	110	5	0	0	0	香港
4	Minh & Kok	Vietnamese Restaurant	7.3	8	False	2	38	0	2	0	0	太古城

After further study, all Dislikes are “False”, BeenHere is zero, and Reasons is rarely used. So we decided to remove these data value as well. Cuisine and Area are discrete value and will not be meaningful to the modeling we chose.

During the GET process, some of the venue ID’s provided by Foursquare shows errors. The Michelin list was OK, but for Non-Michelin list experienced multiple errors and disruptions. As such, the data acquired need to be looped multiple times, and got saved and combined separately. Eventually, we are left with a list of 567 restaurants that included both Michelin’s and Non-Michelin’s.

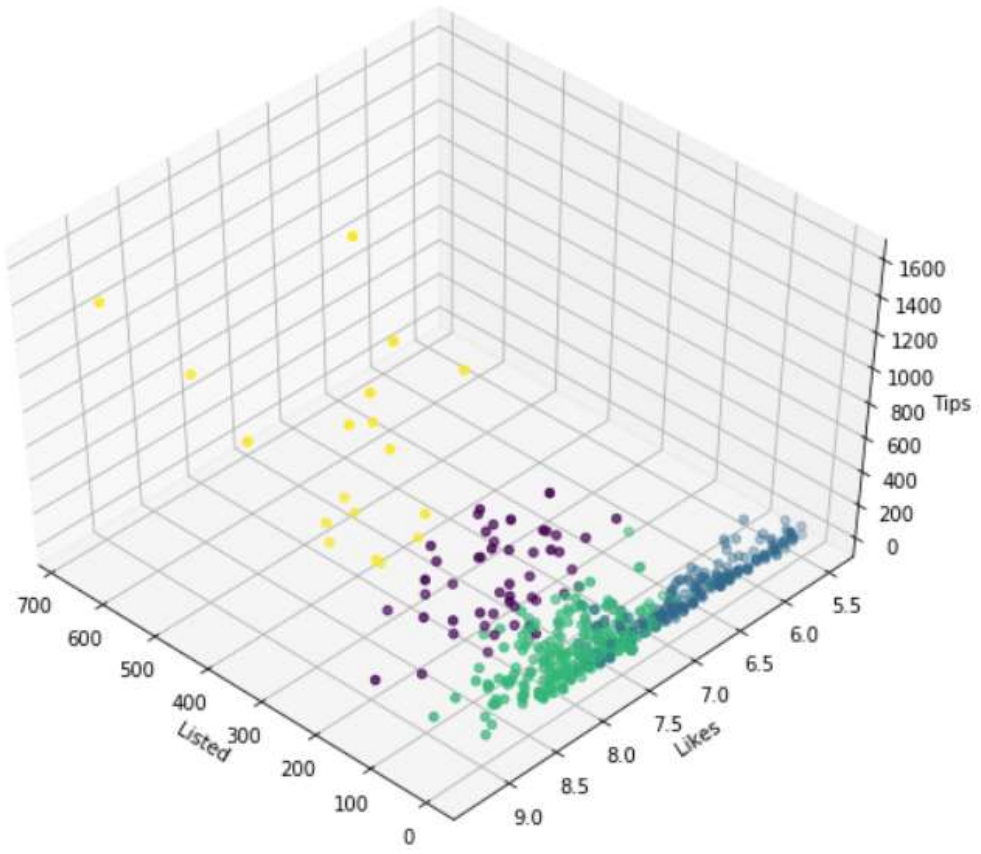
(567, 10)

	Name	Cuisine	Rating	Likes	Tips	Photos	Listed	Price Level	Area	Michelin
0	Loaf On (六福菜館)	Chinese Restaurant	8.5	42	22	129	127	1.0	Sai_Kung	1
1	Lei Garden Restaurant (利苑酒家)	Chinese Restaurant	7.4	14	5	41	7	1.0	Kwun_Tong	1
2	Tim Ho Wan (添好運)	Dim Sum Restaurant	8.2	300	135	902	804	2.0	Sham Shui Po	1
3	Ming Court (明閣)	Cantonese Restaurant	7.7	74	49	309	219	1.8	Mong Kok	1
4	Yat Tung Heen (逸東軒)	Cantonese Restaurant	6.9	18	13	77	19	1.8	Yau_Ma_Tei	1

VII . Model Selection and Building

Let’s start with a K-means Clustering model and see whether it can help us find the “Michelin Quality” restaurants. This unsupervised machine learning algorithm can help us find a cluster that includes a significant share of Michelin restaurants. Those non-Michelin restaurants in the cluster will then have a high chance of being “Michelin Quality” since they demonstrate characteristics that are in close proximity.

In determining the number of clusters required for the input, instead of testing just one cluster number, several numbers were tested. Different versions of the results were generated so that the one which gives the most distinct clusters and the highest share of Michelin restaurants can be found. Finally, it was found that having 4 clusters would give the most ideal result. As shown in below 3D chart, the clusters are clear and distinct.



We look at the mean value of the input variables, and it gives the followings:

Labels	Rating	Likes	Tips	Photos	Listed	Price Level	Michelin
0	8.048980	129.918367	62.285714	410.877551	319.367347	2.024490	0.387755
1	6.623721	11.600000	5.939535	55.469767	12.725581	1.477209	0.106977
2	7.809474	26.378947	12.343860	92.926316	47.617544	2.035789	0.126316
3	8.350000	356.833333	149.055556	869.222222	847.722222	1.755556	0.277778

There are clear difference in Ratings, but Likes, Tips, Photos, and Listed show even a bigger difference among the clusters. So we are confident that the system has done a good job. The number of restaurants in each cluster is as below. Cluster 0 has 49 restaurants, while Cluster 1 has 215, Cluster 2 has 285, and Cluster 3 has only 18.

```
0     49
1    215
2    285
3     18
```

Next, we need to find which cluster includes a significant share of Michelin restaurants.

```
Share of Michelin restaurant: Cluster 0 is 0.39, Cluster 1 is 0.11, Cluster 2 is 0.13, Cluster 3 is 0.28
```

Cluster 0 stands out in having the highest share of Michelin restaurants. It is 39% compared to 28%, 13%, and 11% for the rest. There is a total of 49 restaurants in cluster 0, and 19 of them are on the Michelin list . Therefore, we have strong evidence **that the rest 30 out of 49 restaurants are of "Michelin quality".**

Now, Cluster 0 includes only a small quantity of restaurants. But in Cluster 1,2,3, and beyond there are also Michelin quality restaurants that are undiscovered. In fact, many current Michelin restaurants are found in these clusters. For example, there are 29 Michelin restaurants out of 285 in Cluster 2. We now need another model to help us.

VIII. Logistics Regression for "Michelin Quality" discovery beyond Cluster 0

The logistics regression works well for classification problem, and it can provide a probability value for whether the restaurant will likely be Michelin quality. However, its usefulness depend highly on the soundness of its prediction, so a few evaluations models such as jaccard and confusion matrix will be used.

(567, 8)

	Name	Rating	Likes	Tips	Photos	Listed	Price Level	Michelin
0	Loaf On (六福菜館)	8.5	42	22	129	127	1.0	1
1	Lei Garden Restaurant (利苑酒家)	7.4	14	5	41	7	1.0	1
2	Tim Ho Wan (添好運)	8.2	300	135	902	804	2.0	1
3	Ming Court (明閣)	7.7	74	49	309	219	1.8	1
4	Yat Tung Heen (逸東軒)	6.9	18	13	77	19	1.8	1

The same clean dataset with 567 samples is used. The y, or dependent variable, will be on column "Michelin" which gives a value of 1 if it is of Michelin quality, and a value of 0 if it is not. The rest will be independent variables.

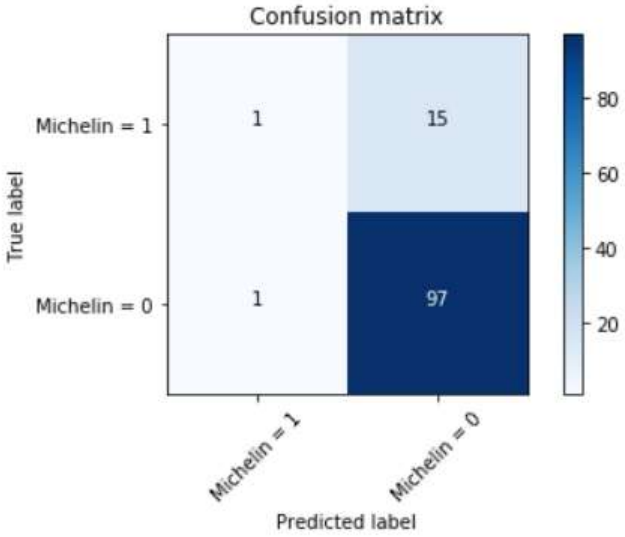
For training purpose, 20% of Foursquare's data was allotted for testing purpose. Several solvers, such as newton-cg, saga, sag, and liblinear, were applied to see what works best. Finally, solver lbfgs gives the best prediction. Below is the result of the evaluations:

i) Jaccard index : 0.86

ii) The F1 – score is 0.81

	precision	recall	f1-score	support
0	0.87	0.99	0.92	98
1	0.50	0.06	0.11	16
micro avg	0.86	0.86	0.86	114
macro avg	0.68	0.53	0.52	114
weighted avg	0.81	0.86	0.81	114

iii) The confusion matrix:



iv) Log loss: 0.38

Based on the above evaluation, the Logistics Regression is a decent model with over 80% probability of success in making a prediction. According to the Confusion matrix, it is very good at predicting restaurants that don't possess Michelin quality. However, the log loss is still rather high at 0.38, indicating certain weakness in this model.

IX. Results

Using the K-means clustering method, we able came up cluster 0 which has a high share of Michelin restaurants (39%). Therefore, we believe the Non-Michelin restaurants within the cluster share similar characteristics, so very much "Michelin" alike. There are 30 restaurants in this "Michelin Quality" restaurants list we can confidently recommend to customers and Michelin judges for review:

	Name	Rating	Likes	Tips	Photos	Listed	Price Level	Labels	Michelin	Area
124	Under Bridge Spicy Crab (Under Bridge Spicy Cr...	7.8	119	40	361	260	3.0	0	0	Wanchai
131	DimDimSum Dim Sum Specialty Store (點點心點心專門店)	8.3	201	99	470	344	2.0	0	0	Hong_Kong
139	Fook Lam Moon (福臨門)	7.5	81	44	266	295	1.8	0	0	Wanchai
142	Chuen Kee Seafood Restaurant (全記海鮮菜館)	7.4	88	21	508	61	3.0	0	0	Sai_Kung
167	Sea View Congee Shop (海景粥店)	8.1	104	38	381	218	1.8	0	0	Kowloon
168	DimDimSum Dim Sum Specialty Store (點點心點心專門店)	8.0	173	103	442	279	2.0	0	0	Mong_Kok
189	Din Tai Fung (鼎泰豐)	8.9	124	27	163	211	1.0	0	0	Tsim_Sha_Tsui
200	Tapas Bar	8.1	57	87	245	32	3.0	0	0	NaN
206	Ichiran (一蘭)	8.6	120	26	384	151	1.8	0	0	Tsim_Sha_Tsui
230	Ippudo (一風堂)	8.0	121	58	763	271	1.8	0	0	Tsim_Sha_Tsui
231	Crystal Jade La Mian Xiao Long Bao (翡翠拉麵小籠包)	7.9	98	32	311	115	1.8	0	0	Tsim_Sha_Tsui
233	Hu Tong (胡同)	7.6	101	53	469	279	1.0	0	0	Tsim_Sha_Tsui
238	Felix	7.5	94	57	342	167	1.8	0	0	Tsim_Sha_Tsui
239	Nobu	8.1	105	56	451	223	2.0	0	0	Tsim_Sha_Tsui
254	Delicious Kitchen (美味廚)	8.2	116	55	286	245	1.0	0	0	Causeway_Bay
285	Joy Hing Roasted Meat (再興燒臘飯店)	8.0	177	93	431	708	1.8	0	0	Wanchai
288	Pirata	7.8	107	47	238	187	3.0	0	0	Hong_Kong
325	Isola Bar & Grill	7.9	181	81	543	215	2.0	0	0	Central
331	Ham and Sherry	8.2	131	52	182	230	2.0	0	0	Hong_Kong
333	22 Ships	7.9	192	113	652	612	3.0	0	0	Hong Kong
344	Crystal Jade La Mian Xiao Long Bao (翡翠拉麵小籠包)	7.4	156	106	416	170	1.8	0	0	Hong_Kong
347	Peking Garden (北京樓)	8.0	112	31	266	167	1.8	0	0	Central, Central and Western District
356	Ronin	8.7	155	58	370	757	2.0	0	0	Central
359	Sang Kee Congee Shop (生記粥品專家)	8.3	97	40	233	399	1.0	0	0	Sheung Wan
361	Social Place (唐宮小聚)	8.1	122	49	366	219	2.0	0	0	Central, Central and Western District
384	La Vache!	8.2	189	94	322	283	2.0	0	0	Central
392	Carbone	8.5	136	64	234	301	2.0	0	0	Central
393	Sushi Kuu (壽司喰)	7.9	140	73	513	356	3.0	0	0	Central
415	Chôm Chôm	8.7	220	95	283	424	2.0	0	0	Central
417	Posto Pubblico	8.4	204	107	302	461	3.0	0	0	Central

For predicting restaurants in cluster 1, 2, 3, and beyond, we used a logistics regression model. The model seems to do well with an accuracy of over 80%, but with certain weakness in predicting True positive values and a not so low log loss (0.38).

X. Discussion

Due to the low popularity of Foursquare in Hong Kong, some Michelin restaurants did not appear in the database, and certainly the same issue for other non-Michelin restaurants too. The Foursquare data provided free of charge also has its limits. Some of the important data such as dislikes and no. of check-ins are not provided. An time series analysis is also important, for example, for how long was the no. of likes gathered? 100 likes gathered in one month over 100 likes gathered in one year certainly paints a different picture. If more critical data become available, the accuracy of the model can certainly improve significantly. The Foursquare database also has issues in handling non-English content, which leads to inconsistency in raw data and therefore alignments needed to be done

manually. When getting the API data using the venue ID provided by Foursquare, there's also unknown disruptions, which seems to either indicate there's some formatting inconsistency issue with some of the venue data (at least for the Hong Kong region). We would recommend Foursquare to fix this issue not just for developers, but for user experience of non-English speakers.

In this exercise, we also make a major assumption that the Michelin judges give the most reliable and accurate score to the restaurants. Some may also argue that the Michelin professionals have a "noble tongue" that is not representative of the laymen. But since the Michelin guide targets the regular consumers rather than a small group of food professionals, we should be confident that Michelin ratings are also well adjusted for the audience, and therefore not so far away from the tastes of the Foursquare users.

XI. Conclusion

This exercise shows there is a lot of opportunity in leveraging data provided by professionals and those of ordinary users. The data from the professionals is reliable and authoritative, while the data generated by regular users is massive (in this case Foursquare users). Based on this massive data, one can derive a relationship or pattern that reflect the professional judgments (in this case Michelin food critics), which then help one to make prediction that mimics that of a professional.

While the clustering method seems the work well. There is room for improvement for the logistics regression method. The 30 restaurants recommended based on the clustering method is highly reliable, while the prediction model through logistics regression need to be improved when there's more valuable (those that needs to be purchased from Foursquare) data available.