Opening a restauration business in Sydney

Research on best locations

MARCH 2020

Part of Coursera IBM data science project

Business problem

The audience for this project is owners who want to open a restauration business in a new location (here, Sydney). The aim of this project is to help an owner deciding on what location would present the most pros to open his business. The location of a restaurant can have a huge impact on the profit made by the business as it will impact the number of customers coming to the restaurant, as well as the type of customers.

Introduction to the problem

A person wants to open a restauration business in the city of Sydney, Australia. He is looking for a strategic place to open his business, where they would be a potential large number of customers. He also wants a place where they are some other restaurants around to attract customers but offering different type of food to limit the competition. Finally, he wants to be no more than 10 km away from where he lives (North Sydney).

The future owner wants to use the foursquare data to find the neighbourhoods with venues that have large amount of ratings, implying many potential customers. He also wants to use the foursquare data to put forward the location where the venues with large amount of ratings offers different type of food than what he plans to offer. In addition, he wants to promote the location where the venues have quite low ratings, to push the customer toward his new restaurant.

Data

The data used are the data from the foursquare API. For this project, they will include data on the restauration business, within a certain radius of the future business owner home and within the limitation of API calls to the foursquare API. The restaurant data selected will include:

1. The number of reviews from restauration business in Sydney. This will allow to determine where there is a high density of customers. We assume here that a high number of reviews involve a high density of customer.
2. The ratings of the restauration business in Sydney. This will allow to determine where the competition is the lowest. We assume here that the owner's business is more likely to succeed where the other restauration businesses have low ratings.
3. The type of food served in the restauration business. This will help the owner to make sure he doesn't sell the same type of food as his neighbours.
4. The distance from the restaurant to the business owner.

Methodology

In a first place, the data of 50 restauration businesses in a radius of 10 km of the business owner home are queried from the foursquare API. The data are cleaned and put into a pandas dataframe to execute first analysis. At this stage, the data include, for each restaurant, its name, category (type of food), distance from the owner's home, overall rating and number of ratings.

The restauration business without ratings are allocated an average overall rating, as they are not enough data to discard them and they are considered to have neither a positive or negative impact on the customer's behaviour (translated by the average rating).

The first analysis will include a spatial analysis of the number of ratings, the overall rating and the type of restauration businesses. This will be done by showing on a map the location of the business and an indicator of the above criteria, for example a colour or size of markers.

In a second time, machine learning will be used to cluster the different restauration business based on the following criteria:

- Overall rating
- Number of ratings
- Location of restaurant (latitude and longitude)

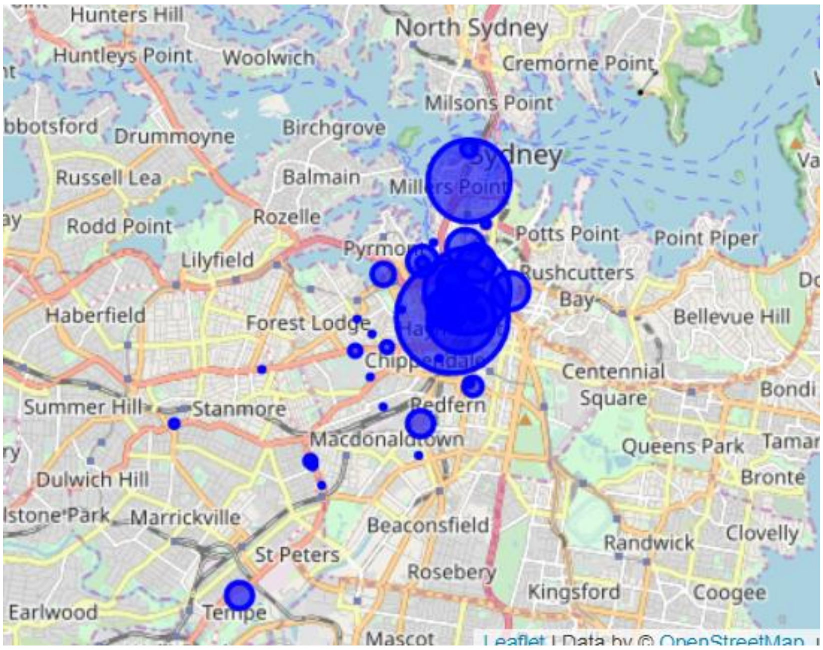
The clustering will allow to highlight the location of businesses based on their number of ratings and overall ratings, therefore helping the business owner to make a decision on its restaurant location. The k-means clustering method will be used, y identifying the optimal number of clusters and then using the algorithm to cluster the restaurants.

Results

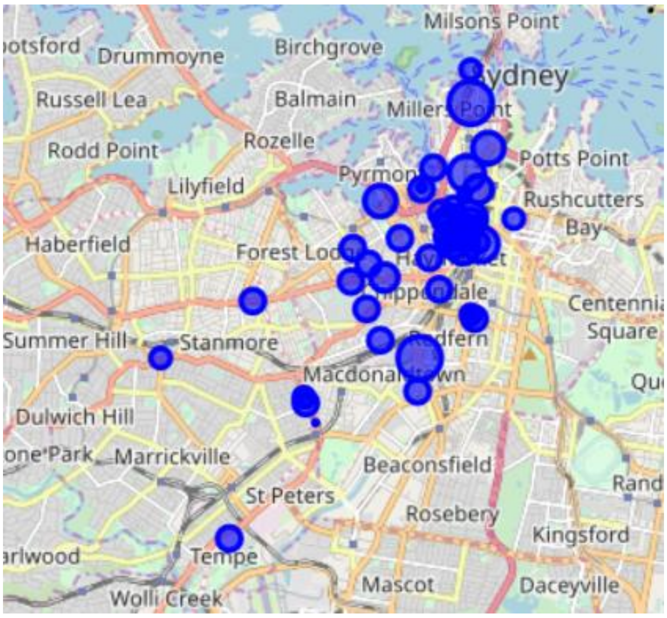
A. Preliminary analysis

In a first time, we analysed the spatial representation of overall rating, number of ratings and restaurant type.

Below is an example of the spatial representation of the number of rating per restaurant (ie showing customer density). As expected, the places closer to the cbd have higher number of ratings, there would therefore be a high potential for customers in these areas.



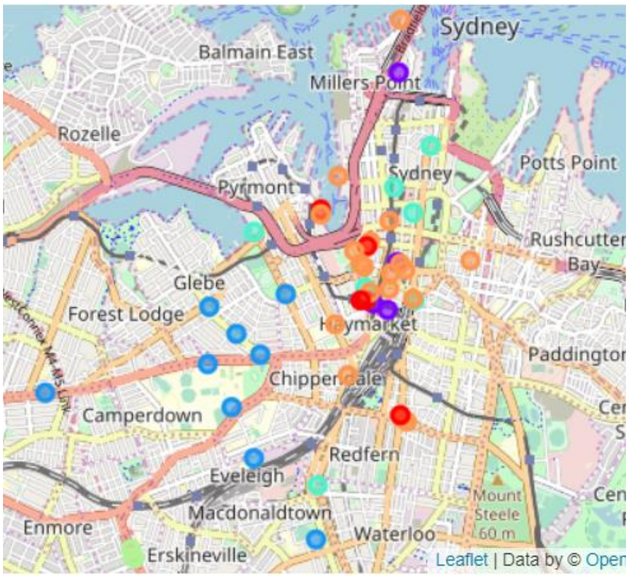
The restaurants with the highest overall ratings seem to be in the same place as the restaurant with the highest number of ratings.



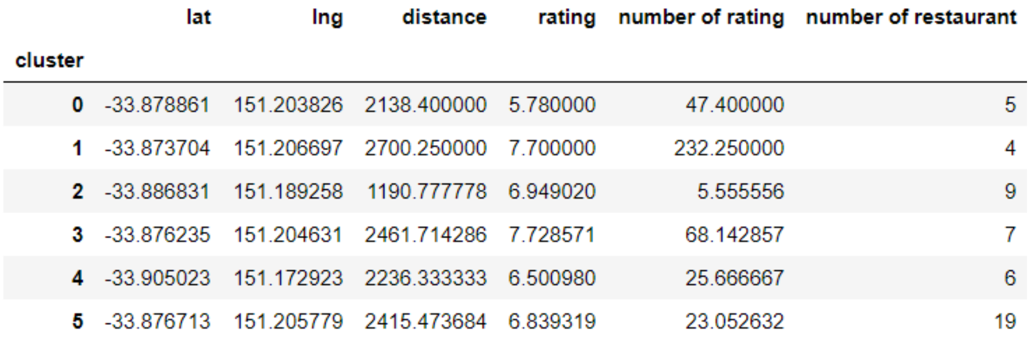
Regarding the type of restaurants, we notice a large number of foreign restaurants. The type of restaurants also seems to be fairly spread across the city. However, the size of the sample is too small to make reliable conclusions.

B. Clustering

The map and tables below show the different restaurants classed in 6 different clusters and their mean characteristics.



Cluster	Color
0	red
1	purple
2	blue
3	cyan
4	green
5	orange



	lat	lng	distance	rating	number of rating	number of restaurant
cluster						
0	-33.878861	151.203826	2138.400000	5.780000	47.400000	5
1	-33.873704	151.206697	2700.250000	7.700000	232.250000	4
2	-33.886831	151.189258	1190.777778	6.949020	5.555556	9
3	-33.876235	151.204631	2461.714286	7.728571	68.142857	7
4	-33.905023	151.172923	2236.333333	6.500980	25.666667	6
5	-33.876713	151.205779	2415.473684	6.839319	23.052632	19

The clusters of interest for a restauration owner would be the following clusters:

- Cluster 0: the restaurants have low rating and a correct number of reviews, there would therefore be low competition and acceptable customers density around these restaurants.
- Cluster 1: the restaurants have a large amount of reviews but also high rating. There would be a high number of potential customers around these restaurants but also a lot of competition.
- The other clusters are less interesting as they have either low potential customers or to high overall ratings for the number of potential customers.

Discussion

The first observation to keep in mind with this research is that the number of restaurants sampled is too low to give reliable results. Indeed, the foursquare API (free version) limits the number of calls to get a business's details at 50. Hence, the study should be done again with professional account, to have more data to analyse.

We also observe the diversity in overall ratings within a same neighbourhood. This is expected as the rating depends on variables not analysed here, such as the chief cuisine, the staff capabilities, etc.

We can notice on the clustering map the proximity of the orange cluster (lower number of ratings and average overall rating) with the two clusters of interest. This indicate that even if two restaurants are close to each other, they might not have the same potential for customers. This make sense as other variables can impact on the choice of a customer to eat in one place, such as the overall look of the restaurant or the prices of the food.

Conclusion

This study allows a restaurant owner to identify areas where there are higher chances of having customers and where are the areas with more competition, ie very good restaurants. It also highlights limitations, such as the amount of data, or the lack of relevant data, and how those can impact on the results. The study's methodology could easily be replicated in other cities. Finally, a restaurant owner must keep in mind that other variables might impact the location, such as the style of the restaurant and what customers it aims to reach (business lunch, fast-food, etc.).