Predicting best restauration business location

COURSERA IBM PROJECT

Introduction

Public aimed

Person wanted to open or move a restauration business in the region of Sydney.

Why

Location of the business impacts number of customers and therefore likelihood of success of the restauration business

Methodology

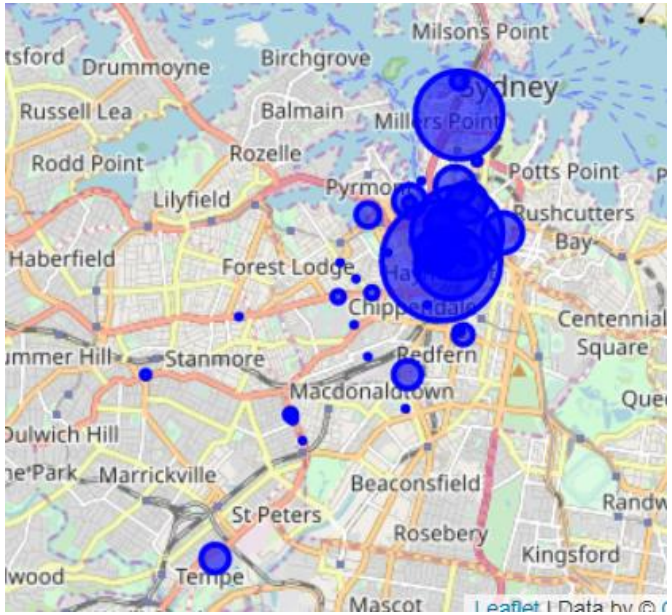
- Obtain data of restauration businesses in Sydney through Foursquare API
- Clean and filter to data to obtain information on location, number of ratings, overall ratings and type of restaurant
- First analysis of the data to see how the number of ratings, overall ratings and type change with location
- Use machine learning to cluster restaurant and identify location of restaurants and areas where there would be a high customer density and low competition

Data

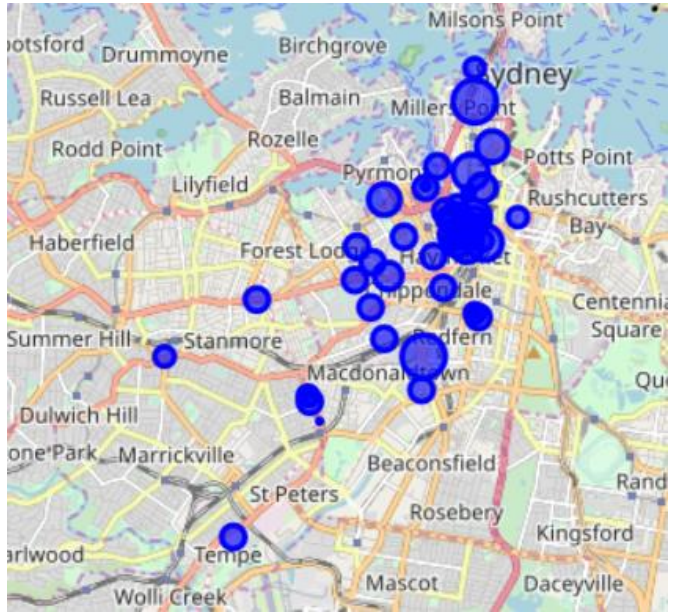
- Obtained from foursquare API
- Limitation of the number of calls to the API for the free version
- Business restauration data used:
  - Number of ratings, giving information on the customer density (the more ratings, the more customers)
  - Overall rating, giving information on the competition level (the highest the rating, the more competition)
  - Location, the restaurant needs to be within 10km of the owner’s home
  - Type of food, giving information on the competition

Preliminary analysis

NUMBER OF RATINGS



OVERALL RATINGS



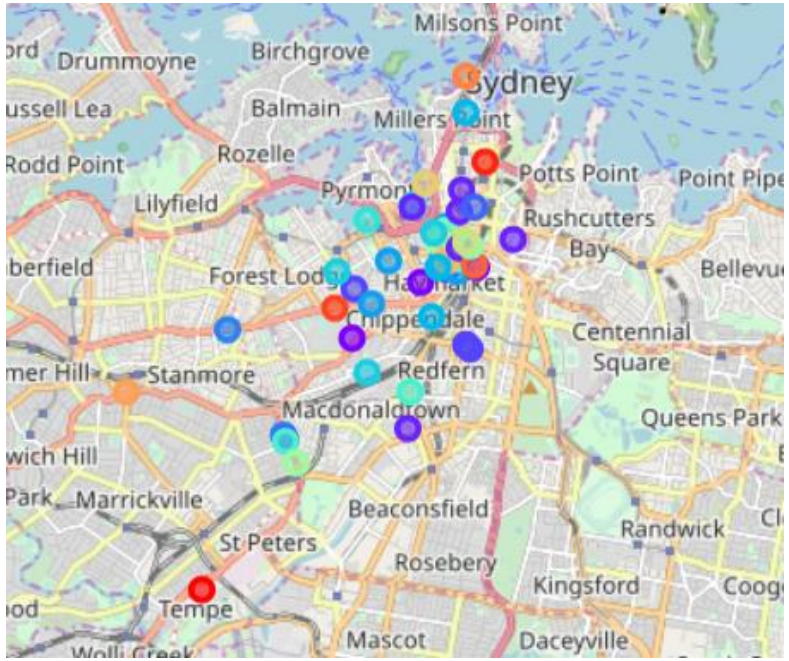
- Centred around the same areas

- Centre have both high and low overall ratings

- Overall rating and number of ratings depends on more variables than location

Preliminary analysis

TYPE OF RESTAURANT

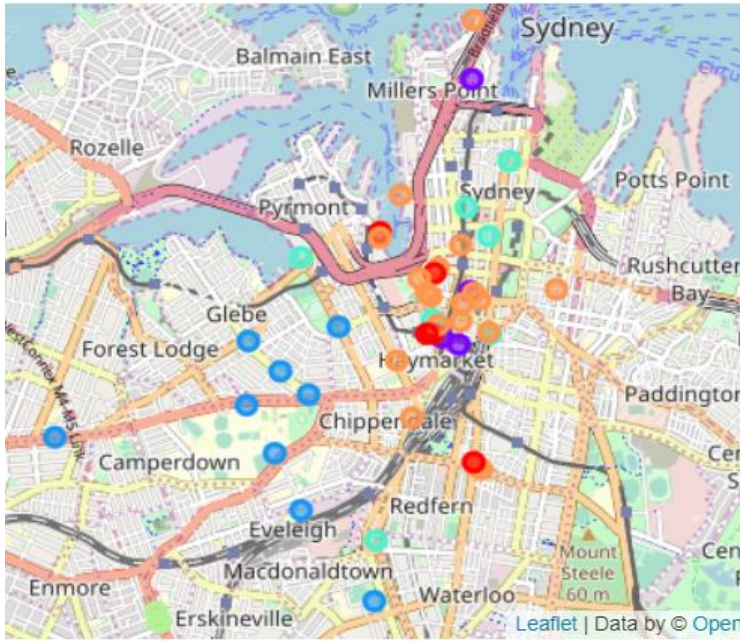


categories	
Chinese Restaurant	6
Italian Restaurant	6
Lebanese Restaurant	5
Japanese Restaurant	3
Australian Restaurant	3
Restaurant	3
Vietnamese Restaurant	2
Turkish Restaurant	2
Dim Sum Restaurant	2
Seafood Restaurant	2
Asian Restaurant	2
Korean Restaurant	2
Pizza Place	1
Szechuan Restaurant	1
Portuguese Restaurant	1
American Restaurant	1
Malay Restaurant	1
Indonesian Restaurant	1
Hotel Bar	1
Fast Food Restaurant	1
Cantonese Restaurant	1
Café	1

- Type of restaurant seems fairly spread
- High occurrence of foreign food restaurants
- Not enough data to conclude (only 50 restaurant due to number of API calls)

Clustering analysis

6 CLUSTERS



cluster	lat	lng	distance	rating	number of rating	number of restaurant
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

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

- Cluster 0: most interesting (low overall rating and moderate attendance)

- Cluster 1: interesting (high customer density but high competition)

Limitations and improvements

- Number of calls to API only allows to have detailed data (ratings) for 50 businesses. Results and analysis would be more accurate with more inputs.
- Would require a further analysis of the interesting spots identified. Are they in main streets, close to businesses, easily noticeable

Conclusion

- Interesting spots have been identified for restauration business location
- Study gives first approach and methodology for identifying best location
- Analysis could and should be done further (access to Foursquare premium)
- Analysis can be replicated in other cities or areas