Capstone Project

Looking for a property in Sydney

Aug 2019

1. Introduction

1.1 Background

A client came from Hong Kong is looking for a property in Sydney close to a train station within a radius of 20km from the Sydney central business district (CBD). To decide which location to look at, the clients wants to have a sense of community and neighborhood of all train stations within 20km from the Sydney CBD. The client approaches our property agency for the information.

1.2 Problem

As a data analyst in the property agency, my boss requires me to provide a neighborhood analysis of all train stations within a radius of 20km from the Sydney CBD. This project is timely and we have very limited recourse. If we go through all the details in each train stations it could become very time consuming. We decide to use online available data obtained from Foursquare API. We will gather the neighborhood details of each possible train stations and apply clustering techniques to group the stations into different cluster.

1.3 Data

- Geographical coordination (i.e. Latitude and longitude coordination) of Sydney CBD
- List of train stations within a radius of 20km from the Sydney CBD
- List of neighborhoods (i.e. venues data) for each train stations

2. Data acquisition and cleaning

2.1 Data source

We can easily obtain the geographical coordination of Sydney CBD by performing a search in the internet. We will use the Foursquare API to get the train stations and venues data. Foursquare has one of the largest database of over 100 million places and is used by over 100 thousands developers.

2.2 Data cleaning

Data obtains from the Foursquare API may not be perfect. We need to delete duplicates and irrelevant data.

3. Methodology

To solve the problem by using data science methodology, we use open sources tools (ie. Jupyter notebook with python) and the Foursquare API to obtain a list of train stations within 20km radius from the Sydney CBD. We will then explore and analyse each train station by using the explore function to get the most common venue categories. We will apply an unsupervised machine learning method (i.e. k-means clustering algorithm) to group the train station into clusters. We will have a glance over the top 5 most common venues of each train stations. We will then use the folium library in python to visualize the train stations in the Sydney map and their

emerging clusters. Finally, we will examine and provide a brief summary of each cluster.

4. Data Analysis

After we obtained the list of train stations, we explore venues within 500m from each train stations. We get the venues within 500m radius from each stations by using API request URL. We investigate how many venues were return from each train station. Figure 1 shows a snapshot of venues for each train station. There are 200 unique categories. We group the train stations by the mean of the frequency of occurrence of each category. We explore each train station along with the top 5 most common venues. Figure 2 shows a snapshot of top 5 most common venues for Artarmon station and Ashfield station.

Figure 1: Snapshot of venues for each train station

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Artarmon Station	19	19	19	19	19	19
Ashfield Station	36	36	36	36	36	36
Auburn Station	24	24	24	24	24	24
Bankstown Station	53	53	53	53	53	53
Burwood Station	46	46	46	46	46	46
Campsie Station	17	17	17	17	17	17
Central Station	84	84	84	84	84	84
Chatswood Station	43	43	43	43	43	43
Clyde Station	6	6	6	6	6	6
Domestic Airport Station	36	36	36	36	36	36
Dulwich Hill Station	18	18	18	18	18	18
Edgecliff Station	11	11	11	11	11	11
Fairfield Station	19	19	19	19	19	19
Gordon Station	14	14	14	14	14	14
Green Square Station	19	19	19	19	19	19

Figure 2: Snapshot of top 5 most common venues for Artarmon station and Ashfield station

```
----Artarmon Station----
                   venue  freq
0                   Café  0.21
1     Japanese Restaurant  0.11
2         Thai Restaurant  0.11
3                   Park  0.11
4  Furniture / Home Store  0.05


----Ashfield Station----
                  venue  freq
0  Dumpling Restaurant  0.08
1     Asian Restaurant  0.08
2   Electronics Store  0.06
3  Shanghai Restaurant  0.06
4          Supermarket  0.06
```

5. Unsupervised Machine Learning Modelling

We run k-means to cluster the neighborhood into 5 clusters and visualize the clusters on map.

5.1 K-mean clustering

K-mean clustering is one of the simplest and popular unsupervised machine learning algorithms. It is easy to understand and delivers results quickly. The objective of K-means is to group similar data points together and discover underlying similarity. To achieve this objective, K-mean looks for a fixed number (kclusters) of clusters in a dataset. We set the number of clusters to 5 in this project.

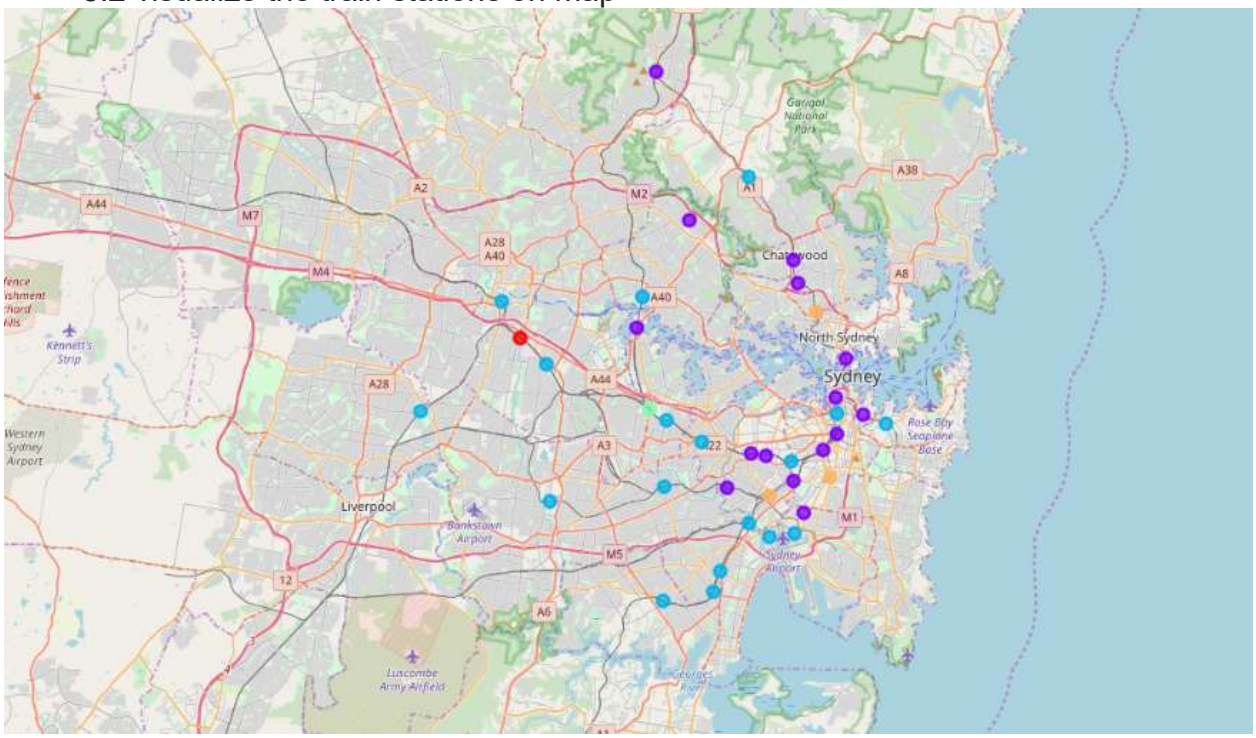
```
# set number of clusters
kclusters = 5

station_grouped_clustering = station_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(station_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

5.2 visualize the train stations on map



6. Results

Cluster 1: Clyde Station

The most common venues in this cluster are gym, Asian restaurant, rental car location. Compared to other clusters, there are not many food options nearby.

	name	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	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue	15th Most Common Venue	16th Most Common Venue	17th Most Common Venue	18th Most Common Venue	19th Most Common Venue	20th Most Common Venue
5	Clyde Station	Platform	Gym	Asian Restaurant	Rental Car Location	Train Station	Food Court	Food & Drink Shop	Flea Market	Fish Market	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Event Space	Electronics Store	Egyptian Restaurant	Yoga Studio	Duty-free Shop	French Restaurant	Dumpling Restaurant	Donut Shop

Cluster 2: Wynyard Station, King's Cross Station, Chatswood Station, Milsons Point Station, Artamon Station, Rhodes Station, St Peters Station, Redfern Station, Mascot Station, Petersham Station, Central Station, Stanmore Station, Hornsby Station, Dulwich Hill Station and Macquarie University Station

There are heaps of good cafés, coffee shops and pub in this cluster. Compared to other clusters, this cluster has a wide range of food options including Italian, Thai, Japanese and Asian restaurant.

	name	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	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue	15th Most Common Venue	16th Most Common Venue	17th Most Common Venue	18th Most Common Venue
0	Wynyard Station	Café	Coffee Shop	Bar	Speakeasy	Cocktail Bar	Sandwich Place	Italian Restaurant	Bakery	Restaurant	Hotel	French Restaurant	Clothing Store	Spanish Restaurant	Dessert Shop	Steakhouse	South Indian Restaurant	Electronics Store	Boutique
9	Kings Cross Station	Café	Italian Restaurant	Coffee Shop	Pub	Bar	Pizza Place	Sushi Restaurant	Thai Restaurant	Japanese Restaurant	Lounge	Indian Restaurant	Dumpling Restaurant	Wine Bar	Speakeasy	Burger Joint	Australian Restaurant	Hotel	Mexican Restaurant
14	Chatswood Station	Café	Coffee Shop	Food Court	Thai Restaurant	Japanese Restaurant	Gym	Chinese Restaurant	Ramen Restaurant	Chaat Place	Portuguese Restaurant	Burger Joint	Bubble Tea Shop	Tea Room	Shopping Mall	Dumpling Restaurant	Fried Chicken Joint	Performing Arts Venue	Market
18	Milsons Point Station	Café	Park	Theme Park Ride / Attraction	Italian Restaurant	Thai Restaurant	Bar	Train Station	Bakery	Lounge	Food & Drink Shop	Flea Market	Fish Market	Asian Restaurant	Seafood Restaurant	Scenic Lookout	Burger Joint	Pub	Pool
20	Artarmon Station	Café	Park	Japanese Restaurant	Thai Restaurant	Sandwich Place	Furniture / Home	Sushi Restaurant	BBQ Joint	Asian Restaurant	Motel	Diner	Ramen Restaurant	Convenience Store	Falafel Restaurant	Food Court	Food & Drink Shop	Flea Market	Fish Market

Cluster 3: Town Hall Station, Hurstville Station, Ashfield Station, Bankstone Station, Auburn Station, International Airport Station, Wolli Creek Station, Kogarah Station, Campsie Station, Parrrmatta Station, Fairfield Station, Domestic Airport Station, Burwood Station, Edgecliff Station, Newtown Station, Gordon Station, Meadowbank Station and Rockdale Station

This cluster has a high concentration of Asian food options, including Chinese, Thai, Japanese, Vietnamese and Korean restaurant. This is easy to find daily necessities nearby as we can see supermarket and grocery store are some of the most common venues.

	name	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	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue	15th Most Common Venue	16th Most Common Venue	17th Most Common Venue	18th Most Common Venue	19th Most Common Venue	20th Most Common Venue
1	Town Hall Station	Café	Japanese Restaurant	Hotel	Coffee Shop	Bookstore	Cocktail Bar	Korean Restaurant	Gym	Thai Restaurant	Shopping Mall	Burger Joint	Record Shop	Speakeasy	Bar	Australian Restaurant	Tea Room	Hobby Shop	Hotel Bar	Ice Cream Shop	Malay Restaurant
2	Hurstville Station	Chinese Restaurant	Fast Food Restaurant	Supermarket	Café	Bakery	Dumpling Restaurant	Vietnamese Restaurant	Coffee Shop	Japanese Restaurant	Bubble Tea Shop	Men's Store	Middle Eastern Restaurant	Souvlaki Shop	Accessories Store	Shopping Mall	Sandwich Place	Clothing Store	Electronics Store	Department Store	Bar
3	Ashfield Station	Dumpling Restaurant	Asian Restaurant	Shanghai Restaurant	Electronics Store	Japanese Restaurant	Supermarket	Chinese Restaurant	Coffee Shop	Korean Restaurant	Falafel Restaurant	Café	Department Store	Restaurant	Malay Restaurant	Bar	Liquor Store	Thai Restaurant	Polish Restaurant	Platform	Pharmacy
7	Bankstown Station	Vietnamese Restaurant	Café	Buffet	Coffee Shop	Grocery Store	Middle Eastern Restaurant	Steakhouse	Chinese Restaurant	Sports Bar	Department Store	Sports Club	Bar	Supermarket	Record Shop	Portuguese Restaurant	Multiplex	Shopping Mall	Bus Station	Fast Food Restaurant	Sandwich Place
8	Auburn Station	Café	Grocery Store	Bakery	Turkish Restaurant	Supermarket	Tea Room	Thai Restaurant	Pakistani Restaurant	Lebanese Restaurant	Gym	Department Store	Fast Food Restaurant	Portuguese Restaurant	Persian Restaurant	Dessert Shop	Afghan Restaurant	Kebab Restaurant	Farmers Market	Falafel Restaurant	Event Space
10	International Airport Station	Airport Lounge	Coffee Shop	Café	Juice Bar	Bakery	Electronics Store	Fast Food Restaurant	Seafood Restaurant	Scenic Lookout	Lingerie Store	Thai Restaurant	Australian Restaurant	Mobile Phone Shop	Duty-free Shop	Train Station	Hotel	Airport Terminal	Pizza Place	Vietnamese Restaurant	Breakfast Spot
11	Wolli Creek Station	Park	Gym / Fitness Center	Athletics & Sports	Train Station	Bakery	Street Food Gathering	Liquor Store	Chinese Restaurant	Asian Restaurant	Café	Dumpling Restaurant	Supermarket	Sushi Restaurant	Coffee Shop	Platform	Pizza Place	Farmers Market	Fast Food Restaurant	Event Space	Fish Market
12	Kogarah Station	Café	Thai Restaurant	Coffee Shop	Pub	Platform	Supermarket	Lebanese Restaurant	Portuguese Restaurant	Train Station	Fast Food Restaurant	Chinese Restaurant	Japanese Restaurant	Italian Restaurant	Shopping Mall	Pizza Place	Convenience Store	Bakery	Sandwich Place	Event Space	Electronics Store

Cluster 4: Strathfield Station

This cluster has high concentration of Korean restaurant and café. Strathfield is also called “Little Korea” in Sydney.

	name	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	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue	15th Most Common Venue	16th Most Common Venue	17th Most Common Venue	18th Most Common Venue	19th Most Common Venue	20th Most Common Venue
16	Strathfield Station	Korean Restaurant	Café	Japanese Restaurant	Fried Chicken Joint	Vietnamese Restaurant	Platform	Thai Restaurant	Malay Restaurant	Liquor Store	Supermarket	Chinese Restaurant	Plaza	Fast Food Restaurant	Sports Bar	Bookstore	Bus Station	Burger Joint	Bakery	BBQ Joint	Shopping Mall

Cluster 5: Green Square Station, Sydenham Station, St. Leonards Station

The most common venues in this cluster is café. It is easy access to gym. We can also find farmers market and fish market in some of the places.

	name	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	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue	15th Most Common Venue	16th Most Common Venue	17th Most Common Venue	18th Most Common Venue	19th Most Common Venue	20th Most Common Venue
25	Green Square Station	Café	Sporting Goods Shop	Coffee Shop	Furniture / Home Store	Supermarket	Electronics Store	Pet Store	Pub	Thai Restaurant	Bar	Gym	Train Station	Farmers Market	Yoga Studio	Fast Food Restaurant	Fish Market	Flea Market	Food & Drink Shop	Falafel Restaurant	Egyptian Restaurant
30	Sydenham Station	Café	Brewery	Gym	Speakeasy	Coffee Shop	Fish Market	Soccer Field	Event Space	Food Court	Food & Drink Shop	Flea Market	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Yoga Studio	French Restaurant	Electronics Store	Egyptian Restaurant	Duty-free Shop	Dumpling Restaurant
32	St Leonards Station	Café	Coffee Shop	Chinese Restaurant	Vietnamese Restaurant	Gym	Office	Sandwich Place	Sushi Restaurant	Bakery	Thai Restaurant	Restaurant	Noodle House	Asian Restaurant	Middle Eastern Restaurant	Pub	Playground	Convenience Store	Pizza Place	Souvlaki Shop	Discount Store

7. Conclusion

By using online data and applying machine learning cluster algorithm, this analysis gives the client a brief overview on the characteristics of different communities within 20km radius of Sydney city Centre. The Foursquare API is very easy to use, however the API we choose will impact what type of data we obtain. It is notices that the foursquare API has details information on restaurants and shops while it does not have much information on public facilities. For future direction, we should explore new data from other APIs so that we can perform a more comprehensive analysis.