Improve Search of a place to live in using Geo-location data and Foursquare API – Case Study: Sydney, Australia

Author: Christophe Nkurunziza

Date: December 15, 2020

1. Introduction

1.1 Background

Sydney is a big multicultural city in which many people migrating to Australia try to go and live in. From the data obtained from https://www.abs.gov.au/statistics/people/population/migration-australia/latest-release , Sydney was the second capital city in Australia to record a great number of immigrants. Considering that a great number of people is going to Sydney and most of them being the first time they are going there, it is always good to have a way to choose where to land.

1.2 Problem

But people coming to live in Sydney do not have a data driven way that is based on venues that will allow them to choose which suburb to live in based. Many people like to live in a place where they can find their preferred venues just around the corner. So having a such system would not hurt.

1.3 Interest

This data cannot only benefit people trying to migrate to Sydney but it will also be used by people trying to open business in different neighborhoods by checking what are the most popular venues there and make decisions accordingly. i.e if a venues has many cafés, you can open offices there as it will be favorable for employees.

2. Data acquisition and cleaning

2.1 Data Sources

Sydney Suburbs and Area code are collected from https://www.matthewproctor.com/full_australian_postcodes_nsw and the suburbs corresponding longitude and latitude information are added using google geocoder API. The data about venues is collected from FourSquare, considering a radius of 1km in order to build a DataFrame having geo-location and venues information.

2.2 Data cleaning

Data obtained from https://www.matthewproctor.com/full_australian_postcodes_nsw has different suburbs corresponding postcodes, longitude and latitude, however it has duplicate information and some suburbs are referenced twice. Data has to be cleaned up, by removing duplicates either on suburb name or base on longitude and latitude info. This dataset also has suburbs from all around New South Wales, the state where Sydney is located. So the data has to be trimmed to only select the data corresponding to Sydney. Besides from https://www.geonames.org/postal-codes/AU/NSW/new-south-wales.html suburbs postcode are in range 2000 to 2999, so any suburb with postcode beyond that range was rejected.

The above data set is enriched with longitude and latitudes for different suburbs. This is done using google geocoder API. Venues info obtained from Foursquare is added as well to create a dataframe.

2.3 Feature selection

Few relevant features have to be extracted from the above collected dataset. The most relevant info were Postcode, Suburb name and SA3 Name which is a region where the suburb is located. The dataset also has geolocation info(longitude and latitude), venues names and venues category. The most important features for my project were suburbs and venues as the suburb name is important in suburb selection after creating clusters of different suburbs and venues are very important in order to create a model.

3. Exploratory data analysis

3.1 Sydney`s venues data frame

Below table is a sneak peek in the Sydney`s venues dataframe. We can see the dataset has Suburb names, suburb latitude and longitude, venues names and venues categories among the most important featuresets.

	Suburb	Suburb Latitude	Suburb Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	BARANGAROO	-33.868820	151.209295	UNIQLO	-33.869744	151.208319	Clothing Store
1	BARANGAROO	-33.868820	151.209295	Haigh's Chocolates	-33.869207	151.207129	Candy Store
2	BARANGAROO	-33.868820	151.209295	The Strand Arcade	-33.869420	151.207630	Shopping Mall
3	BARANGAROO	-33.868820	151.209295	Gumption by Coffee Alchemy	-33.869440	151.207700	Coffee Shop
4	BARANGAROO	-33.868820	151.209295	Skywalk On Sydney Tower	-33.870432	151.208871	Scenic Lookout
...	...	...	...	...	...	...	...
8800	BLACKHEATH	-33.868820	151.209295	Ramblin' Rascal Tavern	-33.873295	151.209773	Bar
8801	BLACKHEATH	-33.868820	151.209295	Cabrito Coffee Traders	-33.862516	151.209324	Café
8802	BLACKHEATH	-33.868820	151.209295	Art Gallery of New South Wales	-33.868821	151.217298	Art Gallery
8803	BLACKHEATH	-33.868820	151.209295	Hobbyco	-33.872218	151.206788	Hobby Shop
8804	BEN BULLEN	-33.495962	150.164770	Bracey Lookout	-33.487778	150.161981	Scenic Lookout

8805 rows × 7 columns

3.2 Sydney`s venues deep dive

Below is a table showing the top 10 venue categories

Venue Category	Suburb	Suburb Latitude	Suburb Longitude	Venue	Venue Latitude	Venue Longitude
Café	1191	1191	1191	1191	1191	1191
Coffee Shop	401	401	401	401	401	401
Bar	338	338	338	338	338	338
Cocktail Bar	312	312	312	312	312	312
Park	283	283	283	283	283	283
Shopping Mall	269	269	269	269	269	269
Hotel	206	206	206	206	206	206
Pizza Place	185	185	185	185	185	185
Thai Restaurant	177	177	177	177	177	177
Japanese Restaurant	175	175	175	175	175	175

We can see cafés, bars, parks, shopping malls, hotels and restaurants among the most common venues in Sydney

Below table shows top 10 venues by in 10 suburbs

Suburb	Venue Category	Suburb Latitude	Suburb Longitude	Venue	Venue Latitude	Venue Longitude
CHATSWOOD	Café	26	26	26	26	26
HMAS PLATYPUS	Café	26	26	26	26	26
ANNANDALE	Café	24	24	24	24	24
REDFERN	Café	23	23	23	23	23
BALMAIN	Café	22	22	22	22	22
CROWS NEST	Café	20	20	20	20	20
FOREST LODGE	Café	20	20	20	20	20
STANMORE	Café	19	19	19	19	19
ERSKINEVILLE	Café	18	18	18	18	18
LEICHHARDT	Café	17	17	17	17	17

Cafés are among the most prevalent venues in Sydney.

Some suburbs have fewer than 10 venues as you can see in the below table:

	Suburb	Suburb Latitude	Suburb Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	BOW BOWING	9	9	9	9	9	9
1	BUNGARRIBEE	9	9	9	9	9	9
2	BIRRONG	9	9	9	9	9	9
3	CHULLORA	9	9	9	9	9	9
4	PYMBLE	9	9	9	9	9	9
...	...	...	...	...	...	...	...
72	BADGERYS CREEK	1	1	1	1	1	1
73	CECIL PARK	1	1	1	1	1	1
74	SILVERDALE	1	1	1	1	1	1
75	BLAXLAND	1	1	1	1	1	1
76	BEROWRA	1	1	1	1	1	1

77 rows × 7 columns

These will be a focus when grouping top 20 most common venues in each suburb and the missing venues will be dealt with accordingly.

3.3. Most common venues per suburb

The Sydney`s venues dataset is treated to extract 20most common venue categories in each neighborhood. This is done by creating a one-hot encoding dataframe from Sydney`s venues categories row and the suburbs row. The venues categories are grouped by suburb and the mean is calculated for each venue category. After this , the dataframe is treated to extract the 20 most common venues. Below is a table showing the dataframe.

	Suburb	1th Most Common Venue	2th Most Common Venue	3th 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	...	11th Most Common Venue	12th Mo Comm Veni
0	ABBOTSBURY	Park	Athletics & Sports	Bar	Buffet	Deli / Bodega	Gym	Italian Restaurant	Pizza Place	Shopping Mall	...	NaN	Na
1	ABBOTSFORD	Café	Italian Restaurant	Park	Burger Joint	Convenience Store	Grocery Store	Martial Arts School	Pizza Place	Soccer Field	...	Wine Shop	Na
2	ACACIA GARDENS	Construction & Landscaping	Convenience Store	Department Store	Fried Chicken Joint	Indian Restaurant	Park	Pub	Supermarket	Train Station	...	NaN	Na
3	AGNES BANKS	Campground	Nature Preserve	Park	Rental Car Location	Rock Climbing Spot	NaN	NaN	NaN	NaN	...	NaN	Na
4	AIRDS	Campground	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	Na
...	...	...	...	...	...	...	...	...	...	...	...	...	
95	CHESTER HILL	Asian Restaurant	Coffee Shop	Dessert Shop	Falafel Restaurant	Fast Food Restaurant	Fish & Chips Shop	Lebanese Restaurant	Market	Newsstand	...	Soccer Field	Supermark
96	CHIFLEY	Café	Gym	Italian Restaurant	Pizza Place	Restaurant	Shipping Store	Supermarket	NaN	NaN	...	NaN	Na
97	CHIPPENDALE	Café	Coffee Shop	Cocktail Bar	Bar	Shopping Mall	Hotel	Speakeasy	Record Shop	Tea Room	...	Candy Store	Steakhou
98	CHULLORA	Lebanese Restaurant	Burger Joint	Dessert Shop	Food Truck	Frozen Yogurt Shop	Furniture / Home Store	Park	Supermarket	NaN	...	NaN	Na

4. Modeling

A model is created that allows users to specify the venues they are interested in and this will provide a cluster to which they can refer to corresponding suburbs and chose any of the existing suburbs as their destination.

4.1 Kmeans clustering

My intention is to group suburbs in a couple of clusters, and suburbs in a given cluster will have similarities that allow them to be grouped together. To achieve this, Kmeans clustering algorithm is used and Sydney suburbs are grouped into 20 clusters.

The model is fit on the existing dataframe first.

```
# set number of clusters
kclusters = 20
sydney_venues_grouped_clustering = sydney_venues_onehot_grouped.drop('Suburb', 1)
# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(sydney_venues_grouped_clustering)
# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:200]
```

Kmeans.labels will be added to existing data frame. Labels are the one categorizing each suburb. Suburbs in the same cluster have the same label.
Below table shows the dataframe including cluster labels:

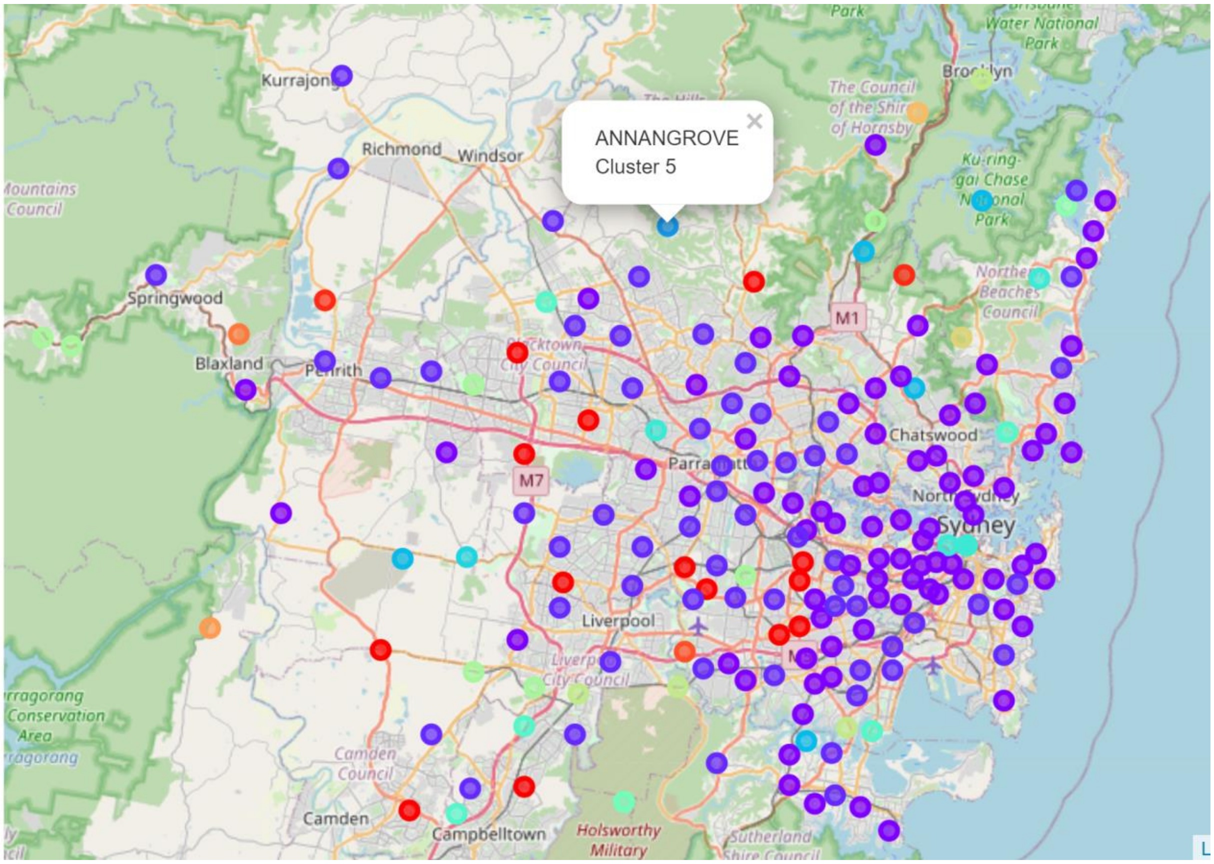
	Suburb	Suburb Latitude	Suburb Longitude	Cluster Labels	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	...	11th Most Common Venue	12th Most Common Venue
147	ABBOTSBURY	-33.870941	150.878382	2	Park	Athletics & Sports	Bar	Buffet	Deli / Bodega	Gym	...	NaN	NaN
37	ABBOTSFORD	-33.856669	151.131581	1	Café	Italian Restaurant	Park	Burger Joint	Convenience Store	Grocery Store	...	Wine Shop	NaN
220	ACACIA GARDENS	-33.720658	150.890015	2	Construction & Landscaping	Convenience Store	Department Store	Fried Chicken Joint	Indian Restaurant	Park	...	NaN	NaN
213	AGNES BANKS	-33.614942	150.698199	2	Campground	Nature Preserve	Park	Rental Car Location	Rock Climbing Spot	NaN	...	NaN	NaN
197	AIRDS	-34.236467	150.814422	3	Campground	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN

5 rows × 24 columns

The kmeans model will be used to predict what suburb cluster , a set of venues fall into.

4.2 Visualizing data on map

Using Folium, we can visualize different clusters on the map. Below is the map created showing different clusters.



4.3 Quick overview on some of the clusters

Cluster 1

	Suburb	1th Most Common Venue	2th Most Common Venue	3th 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	...	11th Most Common Venue	I Com V
122	ARNDELL PARK	Burger Joint	Fast Food Restaurant	Liquor Store	Pizza Place	Thai Restaurant	NaN	NaN	NaN	NaN	...	NaN	
206	BANNABY	Motel	Diner	Fast Food Restaurant	RV Park	NaN	NaN	NaN	NaN	NaN	...	NaN	
158	BASS HILL	Fast Food Restaurant	Gym	Shopping Mall	Bakery	Bar	Big Box Store	Café	Department Store	Garden Center	...	Middle Eastern Restaurant	
148	BONNYRIGG	Fast Food Restaurant	Pizza Place	Dessert Shop	Food & Drink Shop	Gym / Fitness Center	Neighborhood	Park	Sandwich Place	Shopping Mall	...	NaN	
201	BOW BOWING	Fast Food Restaurant	Coffee Shop	Convenience Store	Department Store	Grocery Store	Sandwich Place	Shopping Mall	NaN	NaN	...	NaN	
193	BRINGELLY	Convenience Store	Fast Food Restaurant	Soccer Field	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
111	BURWOOD HEIGHTS	Fast Food Restaurant	Hotel	Asian Restaurant	Bowling Alley	Chinese Restaurant	Convenience Store	Gym	Indian Restaurant	Paper / Office Supplies Store	...	Sandwich Place	
135	CARRAMAR	Bowling Alley	Climbing Gym	Electronics Store	Fast Food Restaurant	Grocery Store	Liquor Store	Supermarket	Train Station	NaN	...	NaN	
218	COLEBEE	Fast Food Restaurant	Sandwich Place	Australian Restaurant	Café	Coffee Shop	Convenience Store	Donut Shop	Electronics Store	Fish & Chips Shop	...	Park	Pharr
202	CURRANS HILL	Fast Food Restaurant	Coffee Shop	Grocery Store	Gym	Pharmacy	Shopping Mall	Supermarket	NaN	NaN	...	NaN	

Cluster 2

	Suburb	1th Most Common Venue	2th Most Common Venue	3th 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	...	11th Most Common Venue	12th M Comr Ve
37	ABBOTSFORD	Café	Italian Restaurant	Park	Burger Joint	Convenience Store	Grocery Store	Martial Arts School	Pizza Place	Soccer Field	...	Wine Shop	I
29	ANNANDALE	Café	Pub	Park	Grocery Store	Pizza Place	Wine Shop	Convenience Store	Italian Restaurant	Gym	...	Fish & Chips Shop	Li S
47	ARTARMON	Café	Japanese Restaurant	Pub	Thai Restaurant	Convenience Store	Electronics Store	Fast Food Restaurant	Furniture / Home Store	Gas Station	...	BBQ Joint	Li S
188	AUDLEY	Café	Pub	Coffee Shop	Convenience Store	Thai Restaurant	Japanese Restaurant	Chinese Restaurant	Train Station	Sushi Restaurant	...	Shopping Mall	Sh Ser
84	AVALON	Café	Beach	Australian Restaurant	Bakery	Health & Beauty Service	Japanese Restaurant	Sandwich Place	Supermarket	Tea Room	...	NaN	I
...	...	...	...	...	...	...	...	...	...	...	...	...	
30	ROZELLE	Café	Pub	Park	Vegetarian / Vegan Restaurant	Pizza Place	Bakery	Bar	Italian Restaurant	Soccer Field	...	Sandwich Place	Playgro
104	SILVERWATER	Café	Badminton Court	Electronics Store	Fast Food Restaurant	Furniture / Home Store	Gym / Fitness Center	Pharmacy	Sandwich Place	NaN	...	NaN	I
39	STANMORE	Café	Pub	Convenience Store	Bar	Sushi Restaurant	Wine Shop	Grocery Store	Furniture / Home Store	Restaurant	...	Pharmacy	Shop
106	SUMMER HILL	Café	Pizza Place	Japanese Restaurant	Bar	Coffee Shop	Convenience Store	Park	Fast Food Restaurant	Motel	...	Light Rail Station	Athleti Sp

Cluster 3

	Suburb	1th Most Common Venue	2th Most Common Venue	3th 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	...	11th Most Common Venue	12th M Comm Ver
147	ABBOTSBURY	Park	Athletics & Sports	Bar	Buffet	Deli / Bodega	Gym	Italian Restaurant	Pizza Place	Shopping Mall	...	NaN	N
220	ACACIA GARDENS	Construction & Landscaping	Convenience Store	Department Store	Fried Chicken Joint	Indian Restaurant	Park	Pub	Supermarket	Train Station	...	NaN	N
213	AGNES BANKS	Campground	Nature Preserve	Park	Rental Car Location	Rock Climbing Spot	NaN	NaN	NaN	NaN	...	NaN	N
189	ALFORDS POINT	Supermarket	Fast Food Restaurant	Grocery Store	Auto Workshop	Juice Bar	Thai Restaurant	Sushi Restaurant	Shopping Mall	Pizza Place	...	Liquor Store	Ice Cre Sl
175	ALLAWAH	Platform	Pub	Chinese Restaurant	Dim Sum Restaurant	Fast Food Restaurant	Football Stadium	Paper / Office Supplies Store	Park	Pet Store	...	Restaurant	Seaf Restaur
...	...	...	...	...	...	...	...	...	...	...	...	...	
35	ST PETERS	Bus Stop	Athletics & Sports	Gym	Thrift / Vintage Store	Steakhouse	Sporting Goods Shop	Speakeasy	Soccer Field	Sandwich Place	...	Pub	Platf
105	SYDNEY MARKETS	Vietnamese Restaurant	Shoe Store	Clothing Store	Sandwich Place	Grocery Store	Accessories Store	Optical Shop	Train Station	Tennis Stadium	...	Platform	P
79	WARRIEWOOD	Supermarket	Bakery	Board Shop	Department Store	Fast Food Restaurant	Garden Center	Middle Eastern Restaurant	Mobile Phone Shop	Movie Theater	...	Toy / Game Store	T
101	WEST PENNANT HILLS	Asian Restaurant	Bakery	Burger Joint	Bus Stop	Café	Park	Soccer Field	Supermarket	NaN	...	NaN	N

5. Evaluation

I created a data set with some venues data and tried to see if it can be classified into clusters. Below is a sample of the dataset I created. The screenshot shows few of the venues categories fields I added more info can be obtained in the ipynb used in this project.

	ATM	Accessories Store	Afghan Restaurant	Airfield	American Restaurant	Antique Shop	Aquarium	Arcade	Argentinian Restaurant	Art Gallery	...	Vegetarian / Vegan Restaurant	Video Game Store	Vietnamese Restaurant	Waterfront
0	0	0	0	0	0	0	0	0	0	0.000	...	0.020000	0.000000	0.010000	0
1	0	0	0	0	0	0	0	0	0	0.000	...	0.000000	0.026316	0.000000	0
2	0	0	0	0	0	0	0	0	0	0.010	...	0.010000	0.000000	0.000000	0
3	0	0	0	0	0	0	0	0	0	0.000	...	0.000000	0.000000	0.000000	0
4	0	0	0	0	0	0	0	0	0	0.000	...	0.000000	0.000000	0.000000	0
5	0	0	0	0	0	0	0	0	0	0.000	...	0.000000	0.000000	0.037037	0
6	0	0	0	0	0	0	0	0	0	0.000	...	0.000000	0.000000	0.000000	0
7	0	0	0	0	0	0	0	0	0	0.000	...	0.013889	0.000000	0.013889	0
8	0	0	0	0	0	0	0	0	0	0.000	...	0.000000	0.000000	0.100000	0
9	0	0	0	0	0	0	0	0	0	0.005	...	0.010000	0.000000	0.000000	0

10 rows × 330 columns

With this and using the kmeans predict method, the label was added to this dataset.

predicted_labels = kmeans.predict(random_df)

	Cluster Labels	ATM	Accessories Store	Afghan Restaurant	Airfield	American Restaurant	Antique Shop	Aquarium	Arcade	Argentinian Restaurant	...	Vegetarian / Vegan Restaurant	Video Game Store	Vietnamese Restaurant	Waterfront
0	1	0	0	0	0	0	0	0	0	0	...	0.020000	0.000000	0.010000	0
1	2	0	0	0	0	0	0	0	0	0	...	0.000000	0.026316	0.000000	0
2	8	0	0	0	0	0	0	0	0	0	...	0.010000	0.000000	0.000000	0
3	0	0	0	0	0	0	0	0	0	0	...	0.000000	0.000000	0.000000	0
4	17	0	0	0	0	0	0	0	0	0	...	0.000000	0.000000	0.000000	0
5	1	0	0	0	0	0	0	0	0	0	...	0.000000	0.000000	0.037037	0
6	19	0	0	0	0	0	0	0	0	0	...	0.000000	0.000000	0.000000	0
7	2	0	0	0	0	0	0	0	0	0	...	0.013889	0.000000	0.013889	0
8	2	0	0	0	0	0	0	0	0	0	...	0.000000	0.000000	0.100000	0
9	8	0	0	0	0	0	0	0	0	0	...	0.010000	0.000000	0.000000	0

10 rows × 331 columns

Having labels we can determine which cluster they belong to and can choose a suburb in that cluster. For instance, row one can be mapped to neighborhoods in cluster one such as Carramar, Burwood heights, ..

5.1 Further work

The dataset used to recommend a suburb to go to has many fields, so far 330, it can be optimized and only few features can be selected and other get populated by default. This can be found after checking what are the most value to represent the null or Nan values.