Facilitate search of place to live in based on venues of interest with geolocation data and Foursquare provided data

case study Sydney/Australia

Rationale of the project

- Sydney is a big multicultural city in which many people migration to Australia try to go and live in.
- But people coming to live in Sydney do not have a quick way to choose which neighborhood to live in based on types of venues they like.
- Geolocation data can 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

Data acquision

- Sydney suburbs Postcodes and names are obtained from https://www.matthewproctor.com/full_australian_postcodes_nsw
- Geolocation data (longitudes and latitudes) are added to the data set using google`s geocoder API
- Venues are added using Foursqare API data.

Data Preparation

- After data wrangling we obtain a data frame like this one

	Postcode	Suburb	Longitude	Latitude	State	Area Name
0	2000	BARANGAROO	151.209295	-33.868820	NSW	Sydney Inner City
1	2006	THE UNIVERSITY OF SYDNEY	151.209295	-33.868820	NSW	Sydney Inner City
2	2007	BROADWAY	151.197131	-33.882319	NSW	Sydney Inner City
3	2008	CHIPPENDALE	151.209295	-33.868820	NSW	Sydney Inner City
4	2009	DARLING ISLAND	151.194217	-33.868789	NSW	Sydney Inner City
...	...	...	...	...	...	...
251	2784	BULLABURRA	151.209295	-33.868820	NSW	Blue Mountains
252	2785	BLACKHEATH	151.209295	-33.868820	NSW	Blue Mountains
253	2786	BELL	150.321112	-33.511587	NSW	Blue Mountains
254	2787	BLACK SPRINGS	149.933476	-33.963276	NSW	Penrith
255	2790	BEN BULLEN	150.164770	-33.495962	NSW	Blue Mountains

256 rows × 6 columns

Adding venues

	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

Data understanding

- 10 most common venues

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

Data understanding(continue ..)

- Grouping venues per suburb and counting them

Suburb	Suburb Latitude	Suburb Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
CHATSWOOD	200	200	200	200	200	200
DARLING POINT	100	100	100	100	100	100
LAWSON	100	100	100	100	100	100
CANOELANDS	100	100	100	100	100	100
CAMPERDOWN	100	100	100	100	100	100
EPPING	100	100	100	100	100	100
BULLABURRA	100	100	100	100	100	100
BUCKETTY	100	100	100	100	100	100
BRONTE	100	100	100	100	100	100
BROADWAY	100	100	100	100	100	100
MASCOT	100	100	100	100	100	100
BLACKHEATH	100	100	100	100	100	100
NORMANHURST	100	100	100	100	100	100
NORTH SYDNEY	100	100	100	100	100	100
PARRAMATTA	100	100	100	100	100	100
PENRITH	100	100	100	100	100	100
BEN BUCKLER	100	100	100	100	100	100
LEWISHAM	100	100	100	100	100	100

Data Understanding(Continue …)

- Café are the most common venues

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

Creating a Model

- The suburbs are grouped in clusters based on the venues they have
- Venue category being a non numerical variable, we use one-hot encoding so that we can create our model
- Clustering is done using Kmeans clustering

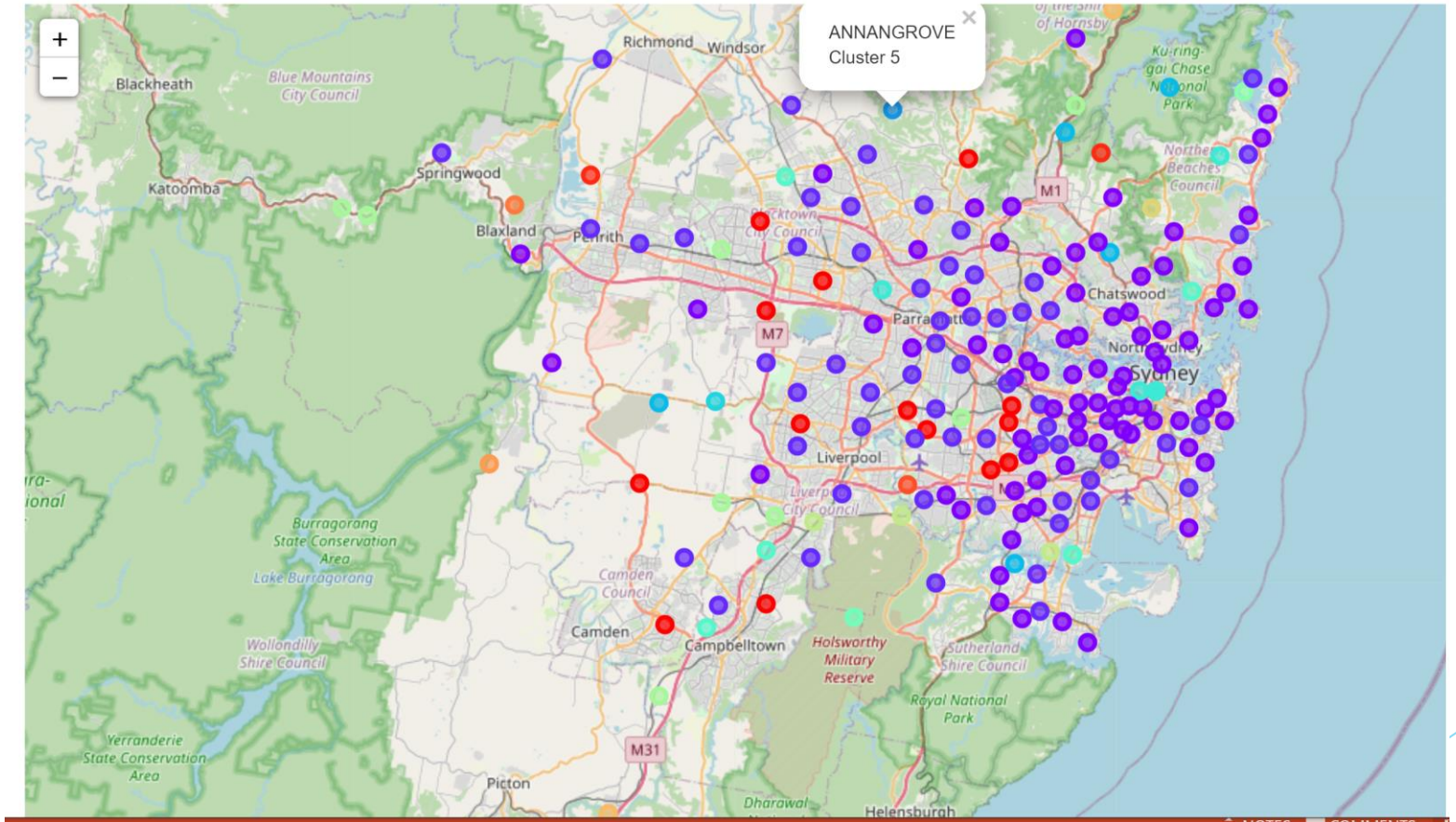
Modeling

	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	Com V
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

Modeling

- Clusters on the map are shown by dots of the same color



Data from one of the clusters

```
ydney_venues_merged['Cluster Labels'] == 8, sydney_venues_merged.columns[[0] + list(range(4, sydney_venues_merged.shape[1]))]
```

5]:

	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 Most Common Venue	13th Most Common Venue
7	ALEXANDRIA	Café	Coffee Shop	Cocktail Bar	Bar	Shopping Mall	Hotel	Speakeasy	Record Shop	Tea Room	...	Candy Store	Steakhouse	De
77	ALLAMBIE	Café	Coffee Shop	Cocktail Bar	Bar	Shopping Mall	Hotel	Speakeasy	Record Shop	Tea Room	...	Candy Store	Steakhouse	De
191	APPLETREE FLAT	Café	Coffee Shop	Cocktail Bar	Bar	Shopping Mall	Hotel	Speakeasy	Record Shop	Tea Room	...	Candy Store	Steakhouse	De
57	ASQUITH	Café	Coffee Shop	Cocktail Bar	Bar	Shopping Mall	Hotel	Speakeasy	Record Shop	Tea Room	...	Candy Store	Steakhouse	De
205	AVON	Café	Coffee Shop	Cocktail Bar	Bar	Shopping Mall	Hotel	Speakeasy	Record Shop	Tea Room	...	Candy Store	Steakhouse	De
11	BANKSMEADOW	Café	Coffee Shop	Cocktail Bar	Bar	Shopping Mall	Hotel	Speakeasy	Record Shop	Tea Room	...	Candy Store	Steakhouse	De

Evaluation

- Using unlabeled data, we can assign venues to clusters. In this example, the model is fed with venues data to predict what cluster the venues should fall under. And using cluster labels, having the cluster we can deduce which neighborhoods might fit to the venues data we had. i.e we can check which neighborhoods in cluster1 and match row 1 to those one.

	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