Final Capstone Project Presentation

Srikanth Satakopan

May 2020

Coursera - IBM Data Science Professional Certification

Problem Description

- Our company is planning to set up its primary IT services operations in New Castle, New South Wales. I have been asked to lead the effort of setting up operations and ensure we identify the right location for the office space and also a great neighborhood for the pool of talent we will be hiring.

- Our company also intends to use this study to share the details with our prospective employees who we are currently recruiting. It is very critical to choose the right location for the office as it will serve as a hub for our operations in the New South Wales region.

- It is important to be cost effective but also be closer to bigger cities like Sydney which is the capital of New South Wales to ensure we attract the right pool of talent. Our company has major offices in Adelaide, South Australia and Brisbane, Queensland and is now planning to expand in New South Wales as part of a major planned growth and expansion opportunities due to the increased revenue projections of our service offerings.

Exploring Australia Territories to understand Landscape
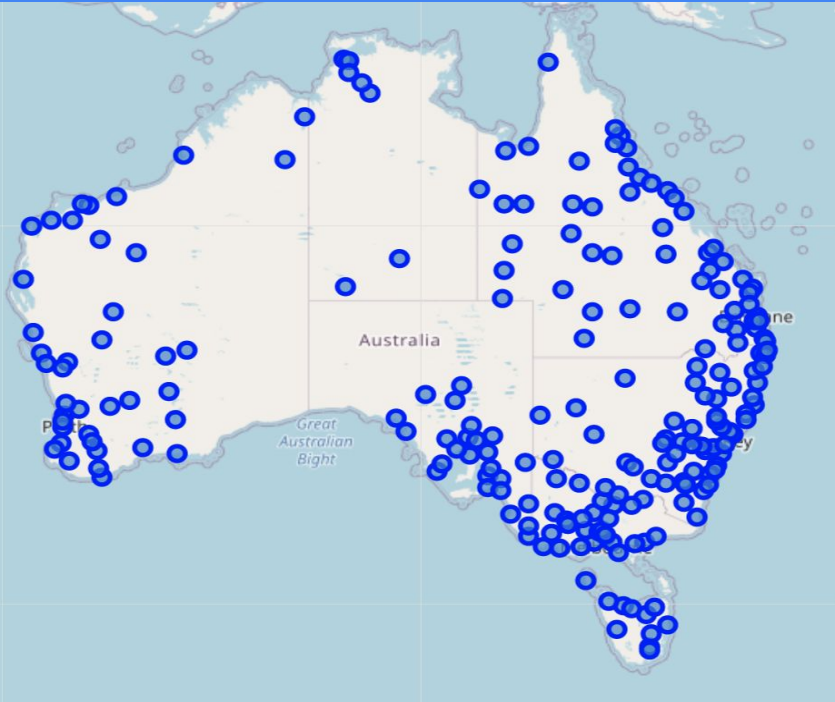
- To understand the locations of the different territories and the geographical coordinates I used a folium geocoders python library to plot the coordinates of the Australian cities.

- This required unpacking the Australian Cities JSON File and extracting key information attributes
  - 'City', 'Admin', 'Country', 'Population Proper','ISO2', 'Capital','Lat','Lng','Population'

Where is New South Wales and what are the major cities?
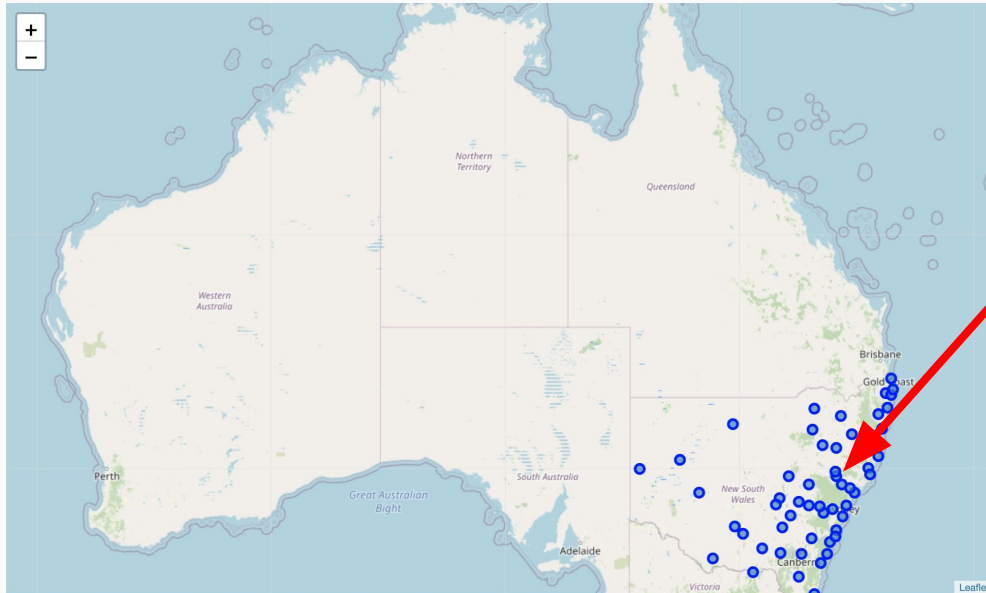
	City	Admin	Country	Population Proper	ISO2	Capital	Lat	Lng	Population
0	Sydney	New South Wales	Australia	3641422	AU	admin	-33.861481	151.205475	4630000
1	Newcastle	New South Wales	Australia	497955	AU		-32.927792	151.784485	1134616
2	Wollongong	New South Wales	Australia	141725	AU		-34.424	150.893448	260914
3	Albury	New South Wales	Australia	32810	AU		-36.074823	146.924006	104258
4	Nowra	New South Wales	Australia	27292	AU		-34.88422	150.600357	94781
5	Coffs Harbour	New South Wales	Australia	34944	AU		-30.296261	153.11351	62978
6	Wagga Wagga	New South Wales	Australia	35717	AU		-35.125771	147.353745	55381
7	Port Macquarie	New South Wales	Australia	35593	AU		-31.43084	152.908936	48547
8	Taree	New South Wales	Australia	16081	AU		-31.91099	152.453873	44182
9	Orange	New South Wales	Australia	34087	AU		-33.283089	149.100006	39329
10	West Tamworth	New South Wales	Australia	31609	AU		-31.10625	150.909356	38551
11	Tweed Heads	New South Wales	Australia	33065	AU		-28.17561	153.542007	33065
12	Queanbeyan	New South Wales	Australia	32214	AU		-35.354931	149.231995	32602
13	Dubbo	New South Wales	Australia	29361	AU		-32.24295	148.604843	31574

- Filtered on New South Wales (NSW) and plotted the major cities in NSW

- Looking at a map of New South Wales would help in our exploration of this area so plotted this region on the map as shown below.

Cost Conscious - Focus on Newcastle, NSW
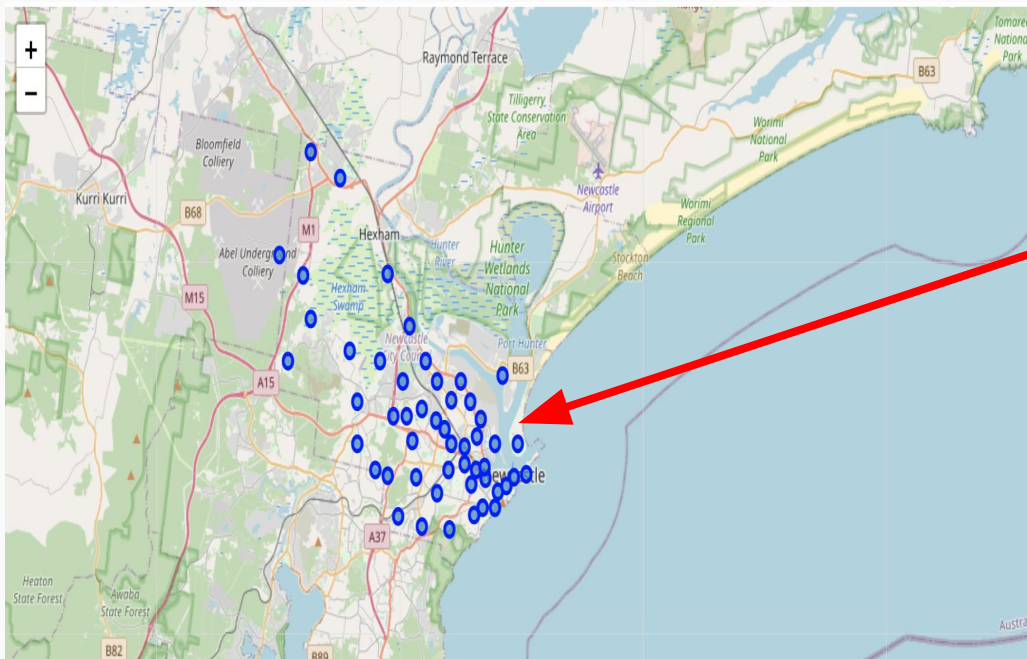

```
[18]: df_newcastle_neighborhood
```

	Neighborhood	City	Lat	Lng
0	Adamstown	Newcastle	-32.9344	151.7258
1	Adamstown Heights	Newcastle	-32.9477	151.7152
2	Bar Beach	Newcastle	-32.9400	151.7678
3	Beresfield	Newcastle	-32.7990	151.6350
4	Birmingham Gardens	Newcastle	-32.8937	32.8937
5	Black Hill	Newcastle	-32.8400	151.6120
6	Broadmeadow	Newcastle	-32.9252	151.7342
7	Callaghan	Newcastle	-32.8900	151.7010

- Focus on neighborhoods of Newcastle about 200 miles from Sydney

- This seemed to be an ideal location closer to a number of major cities in NSW

Foursquare to the rescue - Deeper Insights of Newcastle

```python
# function that extracts the category of the venue
def get_category_type(row):
    try:
        categories_list = row['categories']
    except:
        categories_list = row['venue.categories']

    if len(categories_list) == 0:
        return None
    else:
        return categories_list[0]['name']
```

```python
venues = results['response']['groups'][0]['items']

nearby_venues = pd.json_normalize(venues) # flatten JSON

# filter columns
filtered_columns = ['venue.name', 'venue.categories', 'venue.location.lat', 'venue.location.lng']
nearby_venues =nearby_venues.loc[:, filtered_columns]

# filter the category for each row
nearby_venues['venue.categories'] = nearby_venues.a       get_category_type, axis=1)

# clean columns
nearby_venues.columns = [col.split(".")[-1] for col in nearby_venues.columns]

nearby_venues
```

	name	categories	lat	lng
0	Xtraction	Café	-32.928445	151.783200
1	Reserve Wine Bar	Wine Bar	-32.927360	151.783070
2	Good Brother	Café	-32.928656	151.783220
3	One Penny Black	Breakfast Spot	-32.926633	151.779735
4	Estabar	Café	-32.929019	151.786629
5	Newcastle Beach	Beach	-32.929221	151.786746

- Getting the Venues in Newcastle from Foursquare Venue EndPoint

- Adding the Latitude and Longitude Information to the data frame

Prediction with One Hot Encoding - What is it?

One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction.

```python
# one hot encoding
newcastle_onehot = pd.get_dummies(newcastle_venues[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
newcastle_onehot['Neighborhood'] = newcastle_venues['Neighborhood']

# move neighborhood column to the first column
fixed_columns = [newcastle_onehot.columns[-1]] + list(newcastle_onehot.columns[:-1])
newcastle_onehot = newcastle_onehot[fixed_columns]

newcastle_venues_grouped = newcastle_onehot.groupby('Neighborhood').mean().reset_index()
newcastle_venues_grouped
```

- Group by Neighborhood

- Normalizing the venues for better comparison

[33]:		Neighborhood	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Australian Restaurant	Auto Dealership	Automotive Shop	Bakery	Bar	Baseball Field	Basketball Court	Beach	Bed & Breakfast	B Gar
	0	Adamstown	0.000000	0.000000	0.000000	0.00	0.0000	0.000000	0.0	0.000000	0.000000	0.000000	0.000	0.000	0.0000	0.142
	1	Adamstown Heights	0.000000	0.000000	0.000000	0.00	0.0000	0.000000	0.0	0.000000	0.000000	0.000000	0.000	0.000	0.0000	0.000
	2	Bar Beach	0.000000	0.000000	0.000000	0.00	0.0000	0.000000	0.0	0.000000	0.000000	0.000000	0.000	0.200	0.0000	0.000
	3	Black Hill	0.000000	0.000000	0.000000	0.00	0.0000	0.000000	0.0	0.000000	0.000000	0.000000	0.000	0.000	0.0000	0.000
	4	Broadmeadow	0.000000	0.000000	0.000000	0.00	0.0000	0.000000	0.0	0.000000	0.000000	0.000000	0.125	0.000	0.0000	0.000
	5	Callaghan	0.000000	0.000000	0.000000	0.00	0.0000	0.000000	0.0	0.111111	0.222222	0.000000	0.000	0.000	0.0000	0.000
	6	Carrington	0.000000	0.000000	0.000000	0.00	0.0000	0.000000	0.0	0.000000	0.000000	0.000000	0.000	0.000	0.0000	0.000
	7	Cooks Hill	0.000000	0.000000	0.000000	0.00	0.0000	0.000000	0.0	0.000000	0.068966	0.000000	0.000	0.000	0.0000	0.000
	8	Elermore Vale	0.000000	0.000000	0.000000	0.00	0.0000	0.000000	0.0	0.000000	0.000000	0.000000	0.000	0.000	0.0000	0.000
	9	Fletcher	0.000000	0.000000	0.000000	0.00	0.0000	0.000000	0.0	0.000000	0.000000	0.000000	0.000	0.000	0.0000	0.000
	10	Georgetown	0.000000	0.000000	0.000000	0.00	0.0000	0.000000	0.0	0.200000	0.000000	0.000000	0.000	0.000	0.0000	0.000
	11	Hamilton	0.000000	0.000000	0.052632	0.00	0.0000	0.000000	0.0	0.026316	0.026316	0.000000	0.000	0.000	0.0000	0.000

Clustering the neighborhoods (kclusters=5)
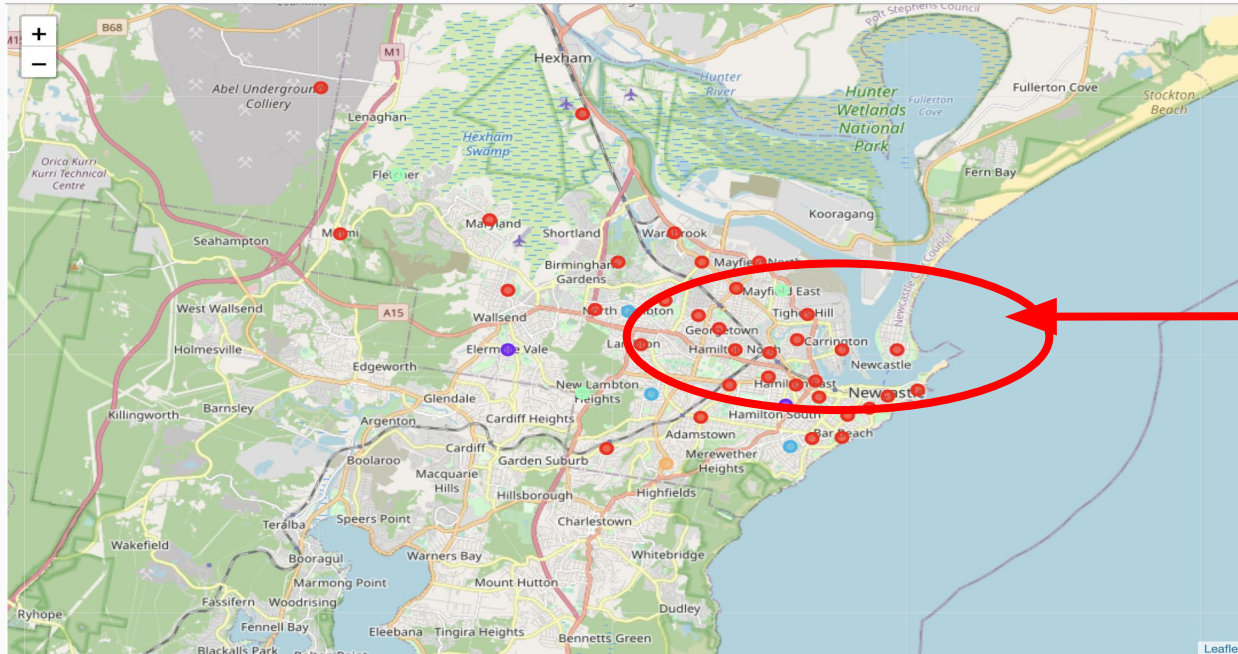
- Newcastle/Newcastle East / Newcastle West showing to be the most prominent areas in terms of venues and the convergence point of cluster 0

- Cluster 0 = Red Bubbles

# Focussing on Clusters (Top 5 Most Common Venues)

Cluster 0

```
[51]: newcastle_merged.loc[newcastle_merged['Cluster Labels'] == 0, newcastle_merged.columns[[1]] + list(range(5, newcastle_merged.shape[1]))]]
```

	City	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
1	Newcastle	Jewelry Store	Trail	Café	Wine Bar	Fried Chicken Joint	Department Store	Dim Sum Restaurant	Donut Shop	Falafel Restaurant	Farmers Market
2	Newcastle	Park	Gym / Fitness Center	Café	Fast Food Restaurant	Beach	Wine Bar	Fried Chicken Joint	Dim Sum Restaurant	Donut Shop	Falafel Restaurant
5	Newcastle	Furniture / Home Store	Wine Bar	Fried Chicken Joint	Cupcake Shop	Department Store	Dim Sum Restaurant	Donut Shop	Falafel Restaurant	Farmers Market	Fast Food Restaurant
7	Newcastle	College Cafeteria	Coffee Shop	Bar	College Bookstore	Juice Bar	Bakery	Sandwich Place	Food & Drink Shop	Dim Sum Restaurant	Donut Shop
8	Newcastle	Café	Mexican Restaurant	Park	Thai Restaurant	Sports Bar	Wine Bar	Department Store	Dim Sum Restaurant	Donut Shop	Falafel Restaurant
9	Newcastle	Café	Pub	Vietnamese Restaurant	Restaurant	Bar	Pizza Place	Fruit & Vegetable Store	Park	Coffee Shop	Falafel Restaurant
12	Newcastle	Ice Cream Shop	Park	Bakery	Sandwich Place	Fast Food Restaurant	Food & Drink Shop	Department Store	Dim Sum Restaurant	Donut Shop	Falafel Restaurant
13	Newcastle	Café	Pub	Mexican Restaurant	Falafel Restaurant	Italian Restaurant	Chinese Restaurant	Snack Place	Dim Sum Restaurant	Racetrack	Playground
15	Newcastle	Park	Pub	Grocery Store	Rental Car Location	Café	Sandwich Place	Fried Chicken Joint	Farmers Market	Wine Bar	Fish & Chips Shop
17	Newcastle	Trail	Salad Place	Fast Food Restaurant	Gas Station	Wine Bar	Food & Drink Shop	Cupcake Shop	Department Store	Dim Sum Restaurant	Donut Shop
18	Newcastle	Café	Bar	Vietnamese Restaurant	Restaurant	Art Gallery	Hotel Bar	Indian Restaurant	New American Restaurant	Park	Coffee Shop

- Analyzing the top 5 most common venues in Cluster 0 shows it has a number of popular spots for the office location with Cafe's, Parks, Walking trails, multi-cuisine Restaurants.

- Based on Venue End Point Analysis Newcastle is the recommended location for the NSW HQ.

Conclusion (Newcastle is the primary choice for NSW HQ)
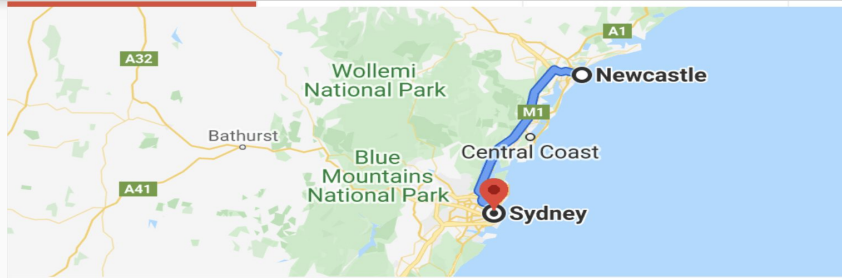
2 h 4 min (160.9 km) via M1

- **200 miles north of Sydney**

- **Amazing beaches, Top venues, much lower cost structure**

- **Recommendation to establish our initial presence in Newcastle and expand as we grow into Sydney.**