Capstone Project Report

Battle of the Neighbourhoods

Finding the best location to open a new Coffeehouse in Sydney

Sapna Yadav

7/26/2020

Contents

INTRODUCTION: 2
BACKGROUND 2
BUSINESS PROBLEM 2
INTEREST 2
DATA 3
DATA SOURCES 3
DATA CLEANSING 3
METHODOLOGY 5
DATA EXPLORATION 5
CLUSTERING 5
RESULTS and DISCUSSION 6
CLUSTER 1 6
CLUSTER 2 7
CLUSTER 3 9
CONCLUSION 10

INTRODUCTION:

BACKGROUND

Sydney (Australia) is the state capital of New South Wales and the most populous city in Australia. In this modern world where most of our communications are online, there is a **growing desire** amongst people to create more personal moments in their lives. Coffee outlets offer the **perfect meeting place.**

A coffee shop offers a **uniquely calm atmosphere** where people can gather with friends to catch up over coffee or focus on work in a **relaxed environment**. Coffee outlets also make the perfect place for **informal meetings and discussions**, thanks to the relaxed ambience. Coffee shops have become so popular over the last decade due to their relaxed and open feel. This is especially prevalent when you compare this establishment to other competing locations, such as restaurants, pubs and fast-food chains.

Coffee is an **intrinsic part of Sydneysider's life** where many of us turn to for both working and social occasions.

BUSINESS PROBLEM

One of the popular retail coffeehouse chain company **is planning to open a new coffee shop in Sydney.** It is trying to decide which suburb of Sydney would be the best one to locate the Coffee shop.

The Company is much concerned about the location. It wants to prefer those location only where the coffee shops are the **most visited venue** so that will be good opportunity for them to open a coffee shop in that area.

The objective of this capstone project is to analyse and select the best suburb location in Sydney to open a coffee house or café. Using Data Science methodologies like data collection, data cleaning, data analysis, this project aims to provide solutions to answer the business problem:

In the city of Sydney, Australia, if a retail coffeehouse company is looking to open a new café or coffee-shop, where would you recommend that they open it?

INTEREST

This project is particularly useful to investors, business owners looking to expand their business in one of the most liveable cities of Australia.

Using this data and methodology company could start to explore possible locations for opening a new coffeehouse based on most visited venue in certain neighbourhood. Here we have presumed that company is well established and successful company.

DATA

DATA SOURCES

The following two datasets are used to solve the above-mentioned business problem-

a. **List of suburbs in Sydney**: This defines the scope of this project. This first dataset is used to retrieve the name with postcodes of all suburbs in Sydney. This data is further helpful to get the coordinates of each location in order to plot the map and to get the venue data. This is done by using the following URL - 'https://www.intosydneydirectory.com.au/sydney-postcodes.php'

On this link, all suburbs of Sydney with their post-codes are given. Then we use GeoPy geocoder service to turn suburb names and postcodes into latitude and longitude.

b. **List of venues suburb wise:** The second dataset is used to get all information about venues like category, venue location etc. located in each suburb. And after that Coffee shop venue is filtered out. This dataset is retrieved with Foursquare Location API. Here we have used "explore" endpoint which give venue recommendations within a radius around a specific suburb

Apart from it, we use GeoPy geocoder service: to turn suburb names and postcodes into latitude and longitude

DATA CLEANSING

Initially we use web scraping using pandas library to extract the required data from the above website page listed. Here we remove those columns which are no longer needed such as 'State' because the data we are using here related to Sydney and Sydney is in NSW state So we can drop this column. The final dataset had **689 suburbs** with their names and postcodes. In order to visualise these suburbs, to provide a wider context for this specific investigation. Latitudes and longitudes for all locations were found using the GeoPy service based on suburb names.

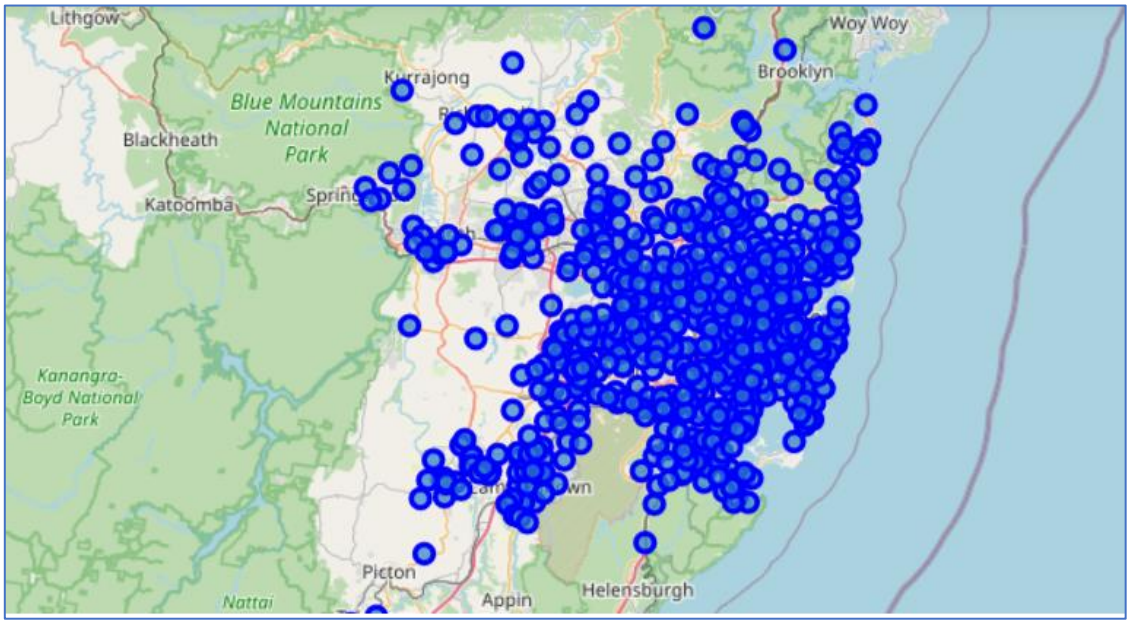


Figure 1 Suburbs locations in Sydney, Australia

Finally, a dataset was created using the Foursquare explore API to provide a listing of up to **10 popular venues** in the random suburb within 500m radius.

The Foursquare data gave around **8,827 venues in total for 689 neighbourhoods or suburbs**. Of these venues **376 unique categories** were found. For each suburb, we also calculate the total number of venues as shown in below table -

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue
Abbotsbury	4	4	4
Abbotsford	7	7	7
Airds	4	4	4
Alexandria	31	31	31
Alfords Point	5	5	5
Allambie Heights	5	5	5
Allawah	4	4	4
Ambarvale	7	7	7
Annandale	25	25	25
Appin	3	3	3

The data was then shaped using **One Hot Encoding**, grouped for each neighbourhood and a mean of the frequency of occurrence found for each venue category. Then we got the **top 5 most common venues** in each suburb considering the frequency of occurrence. E.g. -

```
----Abbotsbury----
                venue  freq
0  Convenience Store  0.25
1               Park  0.25
2        Supermarket  0.25
3           Bus Stop  0.25
4             Office  0.00


----Abbotsford----
                       venue  freq
0                      Café  0.29
1  Construction & Landscaping  0.14
2                  Wine Shop  0.14
3                       Park  0.14
4             Thai Restaurant  0.14


----Airds----
          venue  freq
```

METHODOLOGY

DATA EXPLORATION

A final dataset was created which display the **top venue** for each neighbourhood

	Neighborhood	1st Most Common Venue
0	Abbotsbury	Convenience Store
1	Abbotsford	Café
2	Airds	Pub
3	Alexandria	Café
4	Alfords Point	Candy Store

CLUSTERING

We used machine learning technique such as 'k-means clustering' to segment and cluster the neighbourhoods so that the neighbourhoods with similar characteristics could be grouped together. The output of the final clustering is depicted in below snapshot –

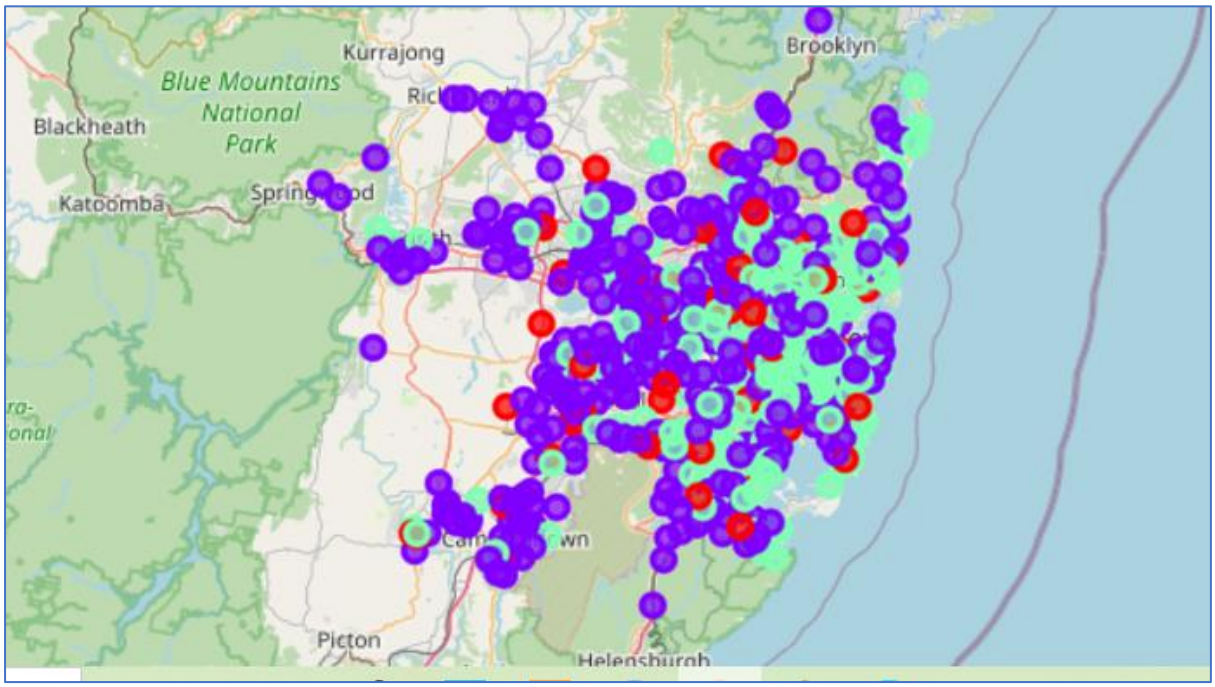


Figure 6: Results of clustering of grouped suburbs bases on similarities

RESULTS and DISCUSSION

We can see through the figure that there are mainly three cluster labelled with - 0.0, 1.0 and 2.0

- Cluster 1 = 0.0
- Cluster 2 = 1.0
- Cluster 3 = 2.0

CLUSTER 1

Here is a snapshot of cluster 1, labelled as '0.0'.

Total rows in this cluster are 54 that means 54 suburbs are in cluster 1.

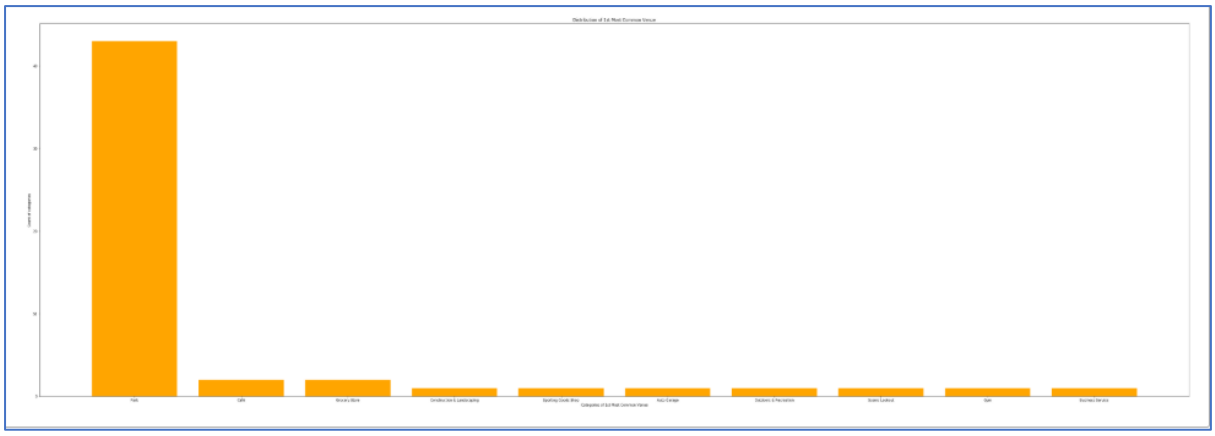
	Neighborhood	Postcode	Latitude	Longitude	1st Most Common Venue	Labels
11	Appin	2560	-34.084740	150.806659	Business Service	0.0
14	Arndell Park	2148	-33.788370	150.879830	Outdoors & Recreation	0.0
22	Austral	2179	-33.926835	150.808311	Park	0.0
28	Balgowlah Heights	2093	-33.804960	151.262050	Park	0.0
30	Balmain East	2041	-33.857680	151.191370	Park	0.0
39	Bardia	2565	-33.977234	150.861874	Park	0.0
41	Bardwell Valley	2207	-33.937130	151.133050	Park	0.0
76	Blakehurst	2221	-33.987970	151.111490	Sporting Goods Shop	0.0
105	Cabarita	2137	-33.845680	151.115550	Park	0.0
135	Cawdor	2570	-34.059512	150.689893	Park	0.0
141	Chatswood West	2067	-33.792570	151.159490	Park	0.0
167	Condell Park	2200	-33.920710	151.005100	Park	0.0

The most common visited venue in cluster 1 is **Park** as shown in below table-

```
cluster_1['1st Most Common Venue'].value_counts()
```

```
Park                          43
Café                           2
Grocery Store                  2
Construction & Landscaping     1
Sporting Goods Shop            1
Auto Garage                    1
Outdoors & Recreation          1
Scenic Lookout                 1
Gym                            1
Business Service               1
Name: 1st Most Common Venue, dtype: int64
```

We can easily visualize the most common visited venue though bar charts as well-



We can conclude in **Cluster 1**, which comprises of 7.8% of total suburbs in Sydney and the most common visited venue in this cluster is **Park** which forms 72% of the first preference in this cluster. The percentage to visit café as first preference is 3.7%.

CLUSTER 2

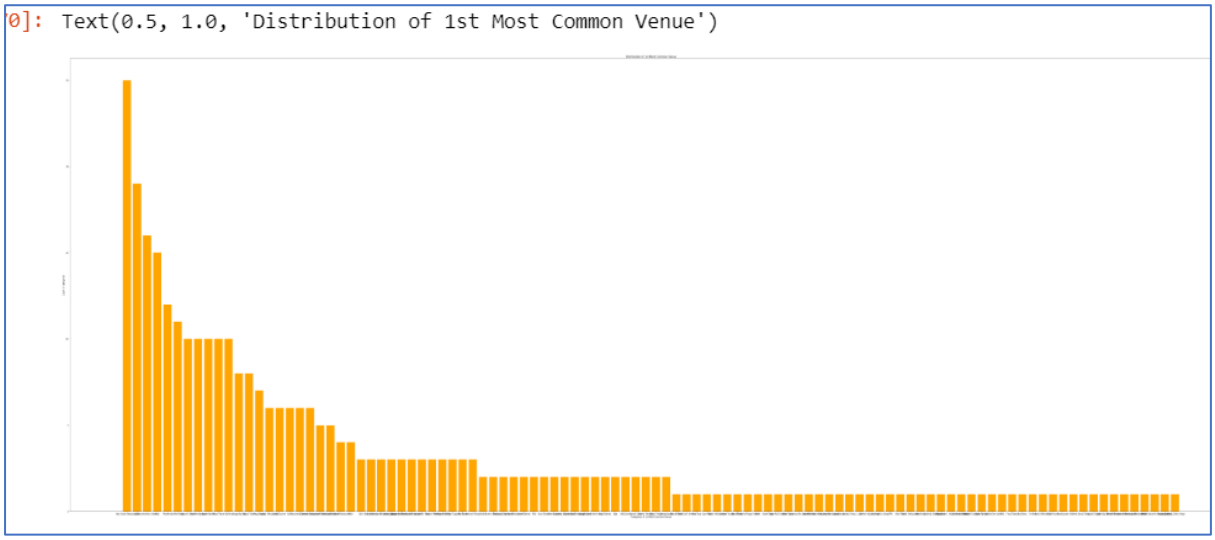
In Cluster 2, which is labelled as 1.0. Total rows in this cluster are 343 that means 343 suburbs are in cluster 2.

	Neighborhood	Postcode	Latitude	Longitude	1st Most Common Venue	Labels
0	Abbotsbury	2176	-33.872850	150.867210	Convenience Store	1.0
3	Airds	2560	-34.084470	150.829040	Pub	1.0
5	Alfords Point	2234	-33.993480	151.024710	Candy Store	1.0
7	Allawah	2218	-33.970260	151.116000	Pub	1.0
8	Ambarvale	2560	-34.083870	150.802430	Pub	1.0
12	Arcadia	2159	-33.882522	151.004487	Dessert Shop	1.0
13	Arncliffe	2205	-33.936650	151.146790	Thai Restaurant	1.0
16	Ashbury	2193	-33.898170	151.121000	Scenic Lookout	1.0
18	Ashfield	2131	-33.889200	151.125430	Platform	1.0
19	Asquith	2077	-33.688710	151.109470	Hobby Shop	1.0
20	Auburn	2144	-33.848480	151.029510	Supermarket	1.0
25	Avon	2574	-34.049730	150.838005	Stadium	1.0

The most common visited venue as a first preference in cluster 2 is **Fast Food restaurant** as shown in below table-

Fast Food Restaurant	25
Café	19
Convenience Store	16
Pub	15
Platform	12
Coffee Shop	11
Grocery Store	10
Thai Restaurant	10
Supermarket	10
Pizza Place	10
Bakery	10
Shopping Mall	8
Liquor Store	8
Playground	7
Indian Restaurant	6
Golf Course	6
Gym	6
Business Service	6
Chinese Restaurant	6
Lebanese Restaurant	5
Electronics Store	5

We can easily visualize the most common visited venue though bar charts as well -



We can conclude that in **Cluster 2** which comprises of 49.7% of total suburbs in Sydney and the most common visited venue in this cluster is **Fast Food restaurant** which is 7.28% of this cluster. The percentage to visit café as first preference is 5.53%.

CLUSTER 3

Here is a snapshot of cluster 3, labelled as 2.0. Total rows in this cluster are 227 that means 227 suburbs are in cluster 3.

	Neighborhood	Postcode	Latitude	Longitude	1st Most Common Venue	Labels
1	Abbotsford	2046	-33.850410	151.128460	Café	2.0
4	Alexandria	2015	-33.912370	151.197030	Café	2.0
6	Allambie Heights	2100	-33.765610	151.251590	Café	2.0
9	Annandale	2038	-33.880050	151.171300	Café	2.0
15	Artarmon	2064	-33.808140	151.183800	Café	2.0
21	Audley	2232	-33.897596	151.154571	Café	2.0
23	Avalon	2107	-33.635860	151.328080	Café	2.0
24	Avalon Beach	2107	-33.635860	151.328080	Café	2.0
27	Balgowlah	2093	-33.793820	151.260660	Liquor Store	2.0
29	Balmain	2041	-33.856000	151.175870	Café	2.0
31	Balmoral	2571	-33.824360	151.232879	Café	2.0
35	Bankstown	2200	-33.914590	151.034280	Café	2.0

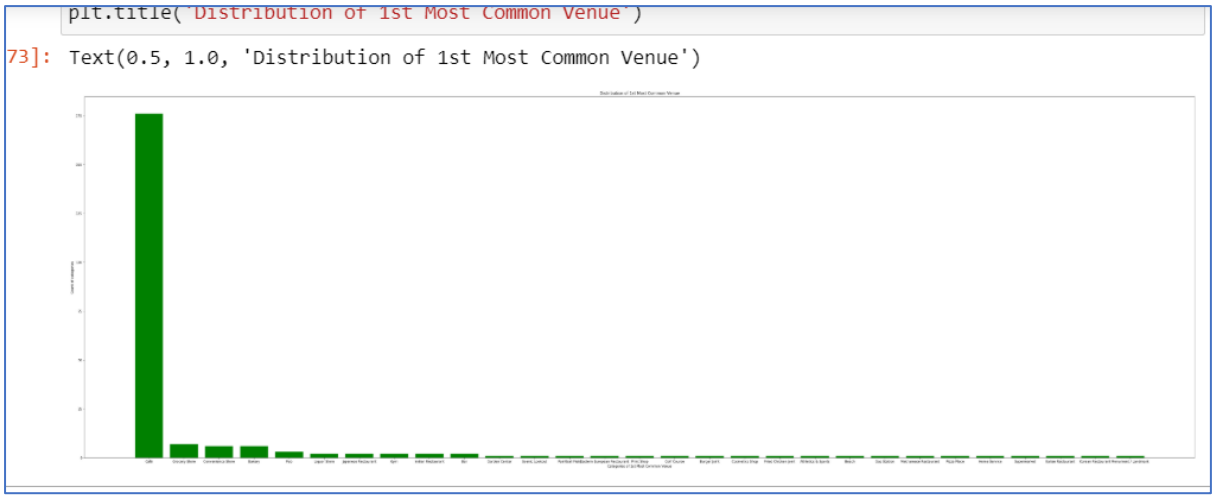
The most common visited venue in cluster 3 is Café as shown in below table-

Café	176
Grocery Store	7
Convenience Store	6
Bakery	6
Pub	3
Liquor Store	2
Japanese Restaurant	2
Gym	2
Indian Restaurant	2
Bar	2
Garden Center	1
Scenic Lookout	1
Paintball Field	1
Eastern European Restaurant	1
Print Shop	1
Golf Course	1
Burger Joint	1
Cosmetics Shop	1
Fried Chicken Joint	1
Athletics & Sports	1
Beach	1
Gas Station	1

We can easily visualize the most common visited venue though bar charts as well -

```
plt.title('Distribution of 1st Most Common Venue')
```

```
73]: Text(0.5, 1.0, 'Distribution of 1st Most Common Venue')
```



We can conclude that, **Cluster 3** which comprises of **32.9%** of total suburbs in Sydney, the percentage to visit the Café as first preference is **77.53%** for this cluster.

CONCLUSION

The aim of this project was to explore the following question:

In the city of Sydney, Australia, if a retail coffeehouse company is looking to open a new café or coffee-shop, where would you recommend that they open it?

We have been able to show that the suburbs grouped into three main types of neighbourhood based on types of the most common visited venue as first preference in those neighbourhoods. By analysing the bar chart and count of most visited venue in each suburb we can conclude that **this analysis might allow Cluster 3 suburb locations to company to open a new coffeehouse**.

In future we can explore the population of each suburb. Also, to explore the distance from the main CBD area of Sydney to that suburb would most likely have impact here also.