The Battle of Neighborhoods

Exploring for a Cafe Business Opportunity

Applied Data Science Capstone Project
Sreenivas Ramakrishnan, 31 March 2021

Table of Contents

1. Introduction: Business Problem and Statement

New South Wales (abbreviated as NSW) is a state on the east coast of Australia. It borders Queensland to the north, Victoria to the south, and South Australia to the west. As of June 2020, the New South Wales has a population of over 8.1 million, making it Australia's most populous state. The greater Sydney metropolitan area, which extends over 12,368 square km, has a staggering population of 5.3 million and is the most populous city in Australia and Oceania. Despite being one of the most expensive cities in the world, Sydney frequently ranks in the top ten most livable cities in the world [1]. Built attractions such as the Sydney Harbor Bridge and the World Heritage-listed Sydney Opera House are also well known to international visitors.

Business opportunities have abounded in Sydney, but the food-and-beverage (F&B) sector has long been an attractive target for investors. The NSW food industry is a significant contributor to the economy with 55,000 food businesses across the state. The food processing industry alone contributes $25 billion [2].

Australia’s cafe and coffee shops generated about $9.9 billion of revenue last financial year, according to an IBISWorld Industry Report, across almost 21,000 businesses – one-third of which are in NSW. Cafe and restaurant culture is key to a great lifestyle, according to the recently released Domain Livable Sydney study. It was one of 19 factors used to assess the livability of 569 suburbs across Greater Sydney. [3]

With the aforementioned prospect, various stakeholders (entrepreneurs, investors) may be interested to explore cafe business opportunities in Sydney. This data science project is thus carried out to help them answer the following question: **Which of the Sydney suburbs are strategic for opening a cafe business?**

The project may also be of interest to fellow restaurant enthusiasts.

2. Data

To explore the potential answer to the problems, the following data are required:

1. **The names of administrative regions in Sydney and their corresponding postal codes**. The regions include only the division of suburbs in Sydney. The information was scraped from a directory on https://www.intosydneydirectory.com.au/sydney-postcodes.php. The suburb names are useful to perform analysis across different areas. The postal codes are needed to obtain the coordinates of each suburb.

2. **Geographical coordinates** of Sydney's suburbs, which will, in turn, be needed to utilize Foursquare API in the subsequent step. Coordinates are obtained using GeoPy libraries, with Arcgis and Nominatim API.

3. **Information about venues in Sydney suburbs**: the names, venue category, venue latitudes, venue longitudes. These are obtained using Foursquare API. The suburbs of Sydney will be clustered based on their surrounding venues to find the best location candidates for opening a cafe.

3. Methodology

Web scraping was performed to extract data of Sydney regions and postal codes as well as retrieval of geographical coordinates. Leveraging Foursquare API, these coordinates data were given as inputs to explore venues within the Sydney suburbs. Two dataframes were then created for use in the analysis:

1. **df_sydney**: contains postal codes and geographical coordinates of all Sydney's suburbs.
2. **sydvenues**: contains at most 50 venues and venues details (name, category, latitude, longitude) for every suburb in Sydney.

One-hot encoding is performed to analyze and narrow down the most common venues in each of the suburbs. Given all the venues surrounding them, subdistricts are clustered using **K-means algorithm**. The number of optimal clusters is decided using the elbow method and silhouette score. Each cluster is separately analyzed to examine one discriminating venue that characterizes them. Analysis of the clusters and visualization of cafe distribution across Sydney will give significant insights on the strategic suburbs for setting up a cafe business.

The following Python libraries and dependencies were used: pandas, NumPy, string, Requests, BeautifulSoup, GeoPy (Nominatim geocoder), JSON, Folium, Matplotlib, and scikit-learn.

3.1 Sydney Suburbs, Postal Codes, and Geographical Coordinates

The data to scrape are the names of all Sydney suburbs and their corresponding postal codes. For example, the first five observations of the NSW data frame look like the following:

	PostalCode	State	Suburbs
0	2176	NSW	Abbotsbury
1	2046	NSW	Abbotsford
2	2753	NSW	Agnes Banks
3	2560	NSW	Airds
4	2015	NSW	Alexandria

Figure 1. Dataframe df_sydney, containing post codes and Sydney's suburbs

The data frame consisted of 689 suburbs, with some similar boroughs such as Liverpool – East and Liverpool – West. A further investigation when plotting the suburb's coordinates on a map of Sydney shows that picking out the first *MultiPoint* was not accurate. As I cannot find any other data sets online which have corrected single coordinates, I cleaned up the data frame manually. I removed all other duplicates of the same suburb (e.g., east, west) as they overlapped other suburbs. The final data frame now consisted of 533 suburbs.

Using the scraped postal codes as inputs, the Arcgis geocoder was used to retrieve latitudes and longitudes of every suburb. Figure 2 displays the first 5 rows of the resulting dataframe: ***df_sydney***.

	PostalCode	State	Suburbs	Latitude	Longitude
0	2176	NSW	Abbotsbury	-33.868755	150.883195
1	2046	NSW	Abbotsford	-33.859164	151.130670
2	2753	NSW	Agnes Banks	-33.651864	150.753264
3	2560	NSW	Airds	-34.056554	150.824705
4	2015	NSW	Alexandria	-33.911604	151.191855

Figure 2. Dataframe df_sydney, containing post codes and geographical coordinates of Sydney regions

3.2 Exploring Venues in the Suburbs of Sydney

A total of 15206 venues were collected through API calls to Foursquare that were made using a user-defined function. The result (Figure 3) is a dataframe containing a maximum of 50 venues within 1 km of a region (i.e., the center of a suburb), with the following details: venue name, venue category, venue latitude, venue longitude. On average, there are 29 venues per suburb (Figure 4).

	Suburbs	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abbotsbury	-33.868755	150.883195	SUPA IGA St Johns Park	-33.871541	150.890444	Grocery Store
1	Abbotsbury	-33.868755	150.883195	Star Buffet (Club Marconi)	-33.864698	150.879435	Buffet
2	Abbotsbury	-33.868755	150.883195	Fratelli's Italian Restaurant	-33.867551	150.882156	Italian Restaurant
3	Abbotsbury	-33.868755	150.883195	Powhatan Park	-33.867138	150.885525	Park
4	Abbotsbury	-33.868755	150.883195	Marconi Club	-33.864336	150.880726	Bar

Figure 3. Dataframe sydvenues, containing information about venues across Sydney's suburbs

	count	mean	std	min	25%	50%	75%	max
Venue Count	519.0	29.298651	31.449927	1.0	6.0	15.0	37.0	100.0

Figure 4. Statistical distribution of venue counts in every Sydney suburbs

3.3 Most Common Venues Overall

As shown in Figure 5, various kinds of restaurants top the list of most common venues in Sydney. Cafe, which is our venue of interest, comes in first. With almost 2000 cafes in Sydney, it sure is an extremely competitive business.

	Venue Category	Count
0	Café	2004
1	Coffee Shop	628
2	Park	534
3	Fast Food Restaurant	428
4	Thai Restaurant	405
5	Bar	383
6	Pizza Place	378
7	Shopping Mall	324
8	Supermarket	313
9	Japanese Restaurant	300

Figure 5. Top 10 most common venues in all of Sydney

	Suburbs	ATM	Accessories Store	Advertising Agency	Afghan Restaurant	Airfield	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	...
0	Abbotsbury	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
1	Abbotsford	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
2	Airds	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
3	Alexandria	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
4	Alfords Point	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
5	Allambie Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
6	Allawah	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
7	Ambarvale	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
8	Annandale	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
9	Appin	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...

Figure 6. One-hot encoding, showing the mean frequency of venue occurrence in each suburb

3.4 One-Hot Encoding

One-hot encoding converts categorical variables (i.e., venues) into numeric variables. In this case, a dummy of all the venues was made and the mean of the frequency of venue occurrence were calculated. The dataframe is then grouped by suburb, as shown in Figure 6. The data were then filtered with a user-defined function to obtain 5 most common venues in each suburb, as displayed in Figure 7.

	Suburbs	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Abbotsbury	Italian Restaurant	Grocery Store	Park	Bar	Café
1	Abbotsford	Café	Grocery Store	Park	Coffee Shop	Burger Joint
2	Airds	Fast Food Restaurant	Indoor Play Area	Pizza Place	Paper / Office Supplies Store	Sandwich Place
3	Alexandria	Café	Playground	Furniture / Home Store	Electronics Store	Italian Restaurant
4	Alfords Point	Fast Food Restaurant	Supermarket	Indoor Play Area	Bus Station	Thai Restaurant
5	Allambie Heights	Scenic Lookout	American Restaurant	Fast Food Restaurant	Liquor Store	Gym
6	Allawah	Pub	Train Station	Park	Paper / Office Supplies Store	Shoe Store
7	Ambarvale	Fast Food Restaurant	Indoor Play Area	Pizza Place	Paper / Office Supplies Store	Sandwich Place
8	Annandale	Café	Park	Grocery Store	Pub	Pizza Place
9	Appin	Fast Food Restaurant	Indoor Play Area	Pizza Place	Paper / Office Supplies Store	Sandwich Place

Figure 7. Five most common venues in each of the Sydney's suburbs

3.5 Clustering Suburbs Based on Venues Similarity

The suburbs were clustered based on a set of similar characteristics or features, i.e., their surrounding venues. *K*-Means clustering, which was used in this part of the analysis, is a machine learning algorithm that creates homogeneous subgroups/clusters from unlabeled data such that data points in each cluster are as similar as possible to each other according to a similarity measure (e.g., Euclidian distance). A value of *k* (number of clusters) needs to be defined before proceeding with the clustering. The "Elbow Method" was used, which calculates the sum of squared distances of data points to their closest centroid (cluster center) for different values of *k*. The optimal value of *k* is the one after which there is a plateau (no significant decrease in sum of squared distances). However, because there is no conspicuous "elbow" from the plot (Figure 8 - left), another measure was used: "Silhouette Score". Silhouette score varies from -1 to 1. A score value of 1 means the cluster is dense and well-separated from other clusters. A value nearing 0 represents overlapping clusters, data points are close to the decision boundary of neighboring clusters. A negative score indicates that the samples might have been assigned into the wrong clusters. Given that there is a peak at *k* = 8 (Figure 8 - right), the *K*-Means clustering was proceeded with that value.

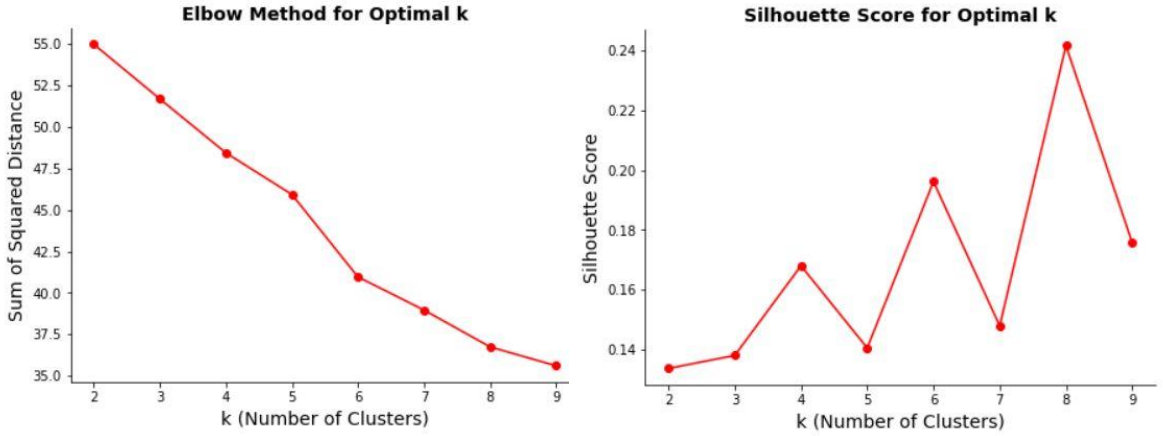

Figure 8. Elbow method: sum of squared distances for different *k* values (left) Silhouette scores across different values of k (right)

	PostalCode	State	Suburbs	Latitude	Longitude	Cluster	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	2176	NSW	Abbotsbury	-33.868755	150.883195	2	Italian Restaurant	Grocery Store	Park	Bar	Café
1	2046	NSW	Abbotsford	-33.859164	151.130670	2	Café	Grocery Store	Park	Coffee Shop	Burger Joint
2	2560	NSW	Airds	-34.056554	150.824705	5	Fast Food Restaurant	Indoor Play Area	Pizza Place	Paper / Office Supplies Store	Sandwich Place
3	2015	NSW	Alexandria	-33.911604	151.191855	2	Café	Playground	Furniture / Home Store	Electronics Store	Italian Restaurant
4	2234	NSW	Alfords Point	-34.018790	151.009984	7	Fast Food Restaurant	Supermarket	Indoor Play Area	Bus Station	Thai Restaurant

Figure 9. A merged dataframe with cluster labels for every suburb

After each suburb had been assigned a cluster label (Figure 9), the clusters were color-coded and visualized on a map of Sydney (Figure 10) to understand how they are distributed across the regions.

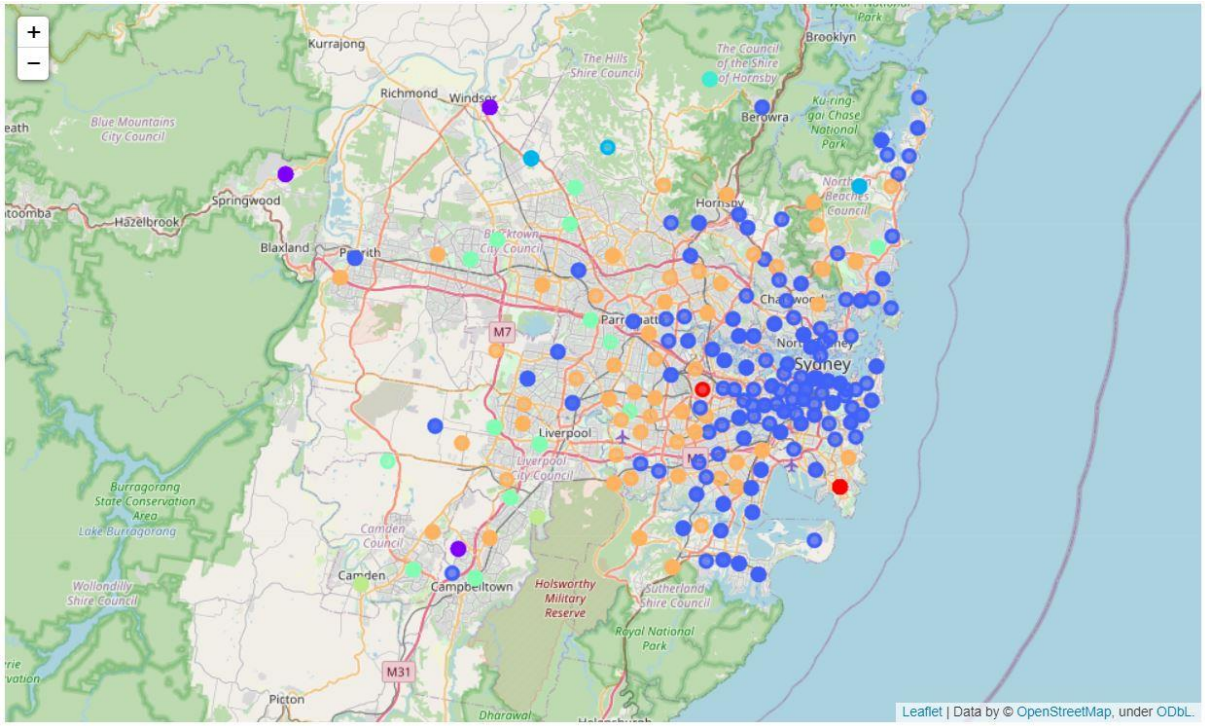

Figure 10. Clusters of Sydney suburbs based on similarity of venues

3.6 Examination of Each Cluster

Each cluster was filtered from the dataframe previously created in the clustering stage. The clusters were separately analyzed to gain an understanding of a discriminating venue that characterize each of them. The number one most common venue categories from each cluster, as well as the regions (suburbs) in which a particular cluster is highly concentrated were singled out.

	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
count	8	8	8	8	8
unique	1	2	2	2	2
top	Park	Shipping Store	Gym	Bus Stop	Restaurant
freq	8	7	7	7	7

Figure 11. Most common venue, Cluster 0 (Red): Park

	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
count	28	28	28	28	28
unique	3	3	3	3	2
top	Miscellaneous Shop	Garden Center	Café	Shopping Mall	Grocery Store
freq	19	19	19	19	25

Figure 12. Most common venue, Cluster 1 (Purple): Miscellaneous Shop

	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
count	250	250	250	250	250
unique	22	41	43	52	57
top	Café	Café	Park	Speakeasy	Cocktail Bar
freq	167	35	34	29	24

Figure 13. Most common venue, Cluster 2 (Dark Blue): Cafe

	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
count	11	11	11	11	11
unique	3	2	2	3	2
top	Pet Store	Home Service	ATM	Other Great Outdoors	Pakistani Restaurant
freq	7	8	8	7	8

Figure 14. Most common venue, Cluster 3 (Light Blue): Pet Store

	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
count	4	4	4	4	4
unique	1	1	1	1	1
top	Jewelry Store	ATM	Organic Grocery	Pakistani Restaurant	Outlet Store
freq	4	4	4	4	4

Figure 15. Most common venue, Cluster 4 (Cyan): Jewelry Store

	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
count	68	68	68	68	68
unique	4	11	10	12	14
top	Fast Food Restaurant	Supermarket	Pizza Place	Paper / Office Supplies Store	Sandwich Place
freq	60	16	15	15	15

Figure 16. Most common venue, Cluster 5 (Green): Fast Food Restaurant

	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
count	21	21	21	21	21
unique	2	2	2	2	2
top	Supermarket	Café	Bar	Australian Restaurant	ATM
freq	19	19	19	19	19

Figure 17. Most common venue, Cluster 6 (Light Green): Supermarket

	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
count	129	129	129	129	129
unique	30	36	38	40	35
top	Train Station	Park	Convenience Store	Shopping Mall	Park
freq	13	10	11	10	12

Figure 18. Most common venue, Cluster 7 (Orange): Train Station

As having cluster labels as 0, 1, 2, 3, 4, 5, 6 and 7 are unintuitive to interpret, a data visualization of the 1st most common venue for each neighborhood in each cluster can be formed. This would help in creating labels for each cluster. The following stacked bar chart for each cluster is shown below in Figure 19.

From analyzing the bar chart and the top five venues data frame, the clusters can be generalized and labeled as the following:

- *Cluster 0(Red)*: Park
- *Cluster 1(Purple)*: Miscellaneous Shop, Pub and BBQ Joint
- *Cluster 2(Dark Blue)*: Cafes and Coffee Shops
- *Cluster 3(Light Blue)*: Pet Store
- *Cluster 4(Cyan)*: Jewelry Store
- *Cluster 5(Green)*: Fast Food Restaurant, Pizza Place
- *Cluster 6(Light Green)*: Supermarket
- *Cluster 7(Orange)*: Train Stations and Parks

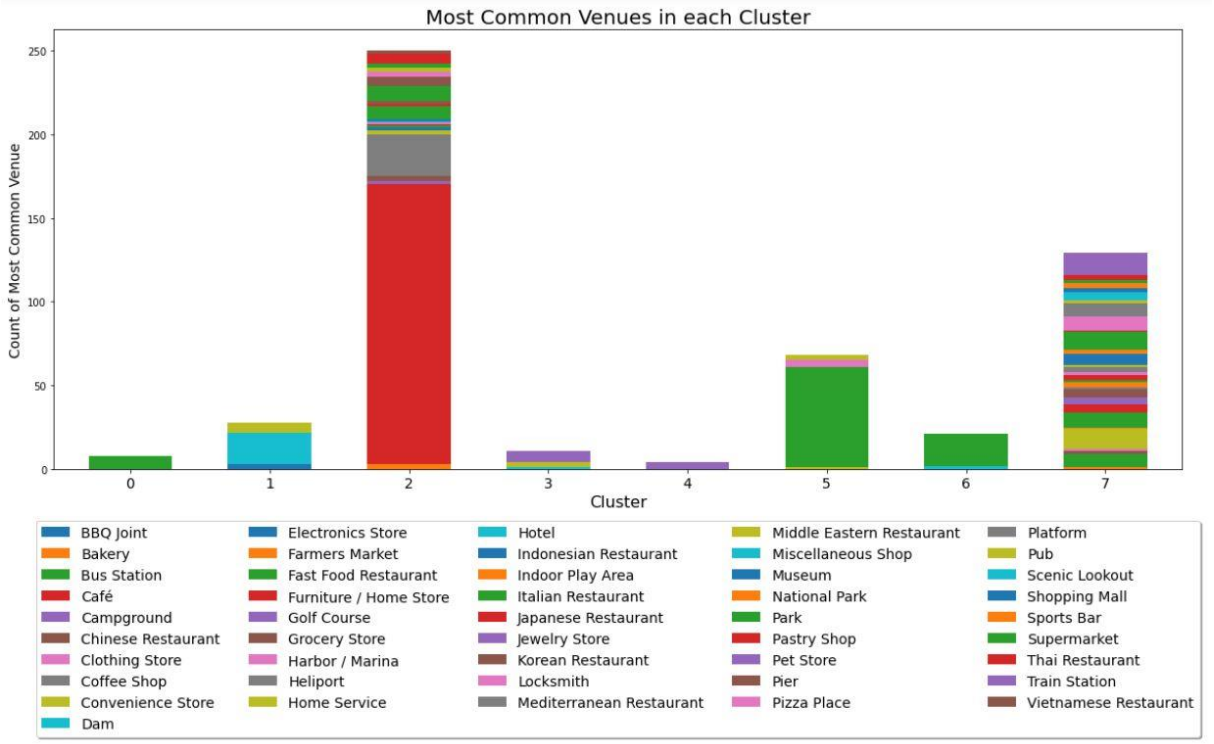

Figure 19. Box plot of most common venues in each cluster

3.7 Visualizing Distribution of Cafe Locations in Sydney

Based on data in the *sydvenues* dataframe, the total number of cafes within each of the Sydney suburbs was calculated to examine the distribution of cafe businesses and to help figure out strategic locations. Clusters 1,2,6 have the cafe venues in the top 5. So, the cafe venues are filtered to envisage the concentration of cafes across Sydney regions. This distribution was visualized in the folium map below (Figure 21).

	Suburbs	Café
0	Mcmahons Point	32
1	Lavender Bay	32
2	Waverton	32
3	Camperdown	24
4	Erskineville	21
5	Surry Hills	21
6	Woolloomooloo	20
7	Mosman	20
8	Elizabeth Bay	20
9	Balmain	20

	Suburbs	Café
335	Smeaton Grange	1
336	Pendle Hill	1
337	Smithfield	1
338	Miller	1
339	Valley Heights	1
340	Padstow	1
341	St Johns Park	1
342	Sun Valley	1
343	Wakeley	1
344	Lalor Park	1

Figure 20. Suburbs with the highest (left) and lowest (right) count of cafes

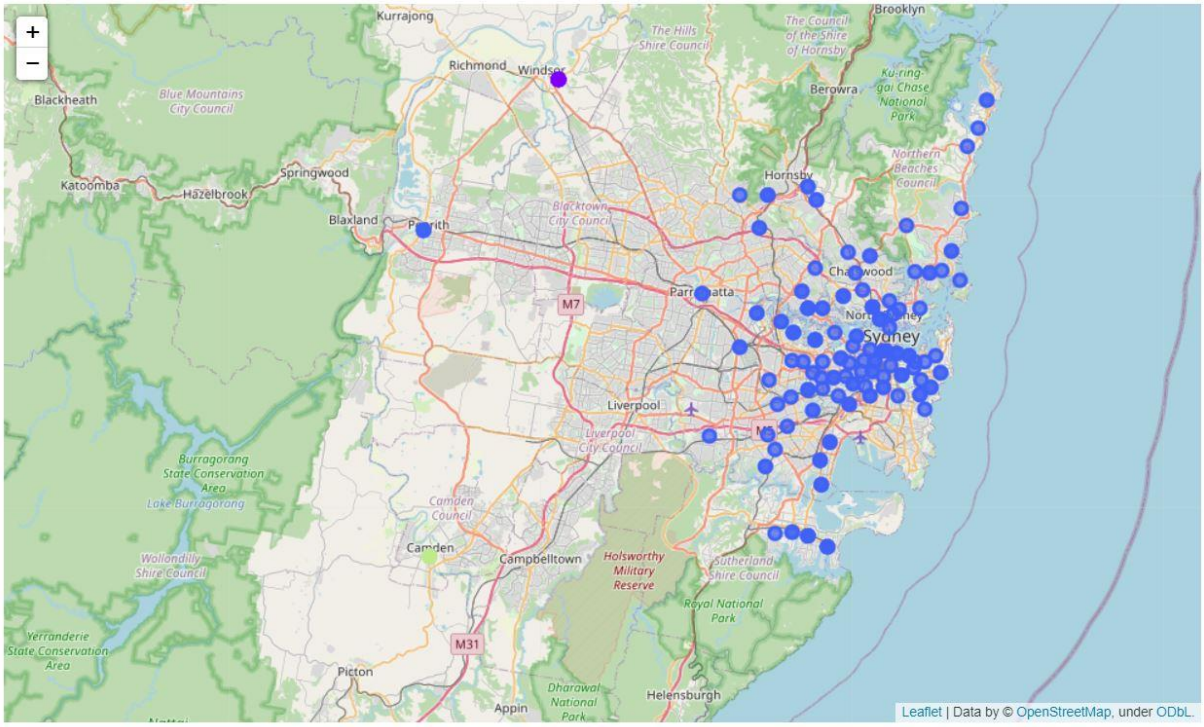

Figure 21. Concentrations of cafes across Sydney

4. Results and Discussion

Exploratory data analysis as well as machine learning and visualization techniques have provided us with some insights into the problem at hand.

A total of 15206 venues from all Sydney regions (519 suburbs) were returned at the time the API call was made. There are on average 29 venues within a kilometer of a suburb center, where two of the most common categories overall are Cafes and Coffee Shops. After deciding on an optimal *k* value of 8, *K*-Means algorithm was run to cluster the suburbs based on their most common surrounding venues. Each of the six clusters, labeled 0-7, is characterized by a dominant venue as follows:

Cluster	Member	#1 Common Venues
0	8	Park
1	28	Miscellaneous Shop
2	250	Cafe
3	11	Pet Store
4	4	Jewelry Store
5	68	Fast Food Restaurant
6	21	Supermarket
7	129	Train Station

Table 1. Results of K-Means clustering

A considerable number of cafes can be found within Cluster 2 (167 + 35 shops out of 250 venues). In fact, it is the 1st and 2nd most common venue in that cluster. It is recommended for people looking to visit or live in Sydney NSW to research more about the neighborhoods to investigate whether the areas suit their lifestyle or culture. The dark blue dots (Cluster 2) are mostly closer to the CBD area of Sydney, with the areas being characterized by café venues. The orange dots (Cluster 7) are further out from Sydney CBD, with the areas being

characterized by their variety of venues. The green dots (Cluster 5) are even further away from Sydney CBD and have the most common venue for fast food restaurants.

The clusters with the greatest number of cafes within the top 5 venues are listed as follows:

Cluster	Member	Type of Common Venue
1	19	3rd Most Common Venue
2	167	1st Most Common Venue
2	35	2nd Most Common Venue
6	19	2nd Most Common Venue

Table 2. Clusters with highest number of cafes within the top 5 venues

A total of 202 cafes can be found within cluster 2 and it is the 1st and 2nd most common venue followed by cluster 6 (2nd most) and cluster 1 (3rd most). Folium map of cafe locations across mainland Sydney shows that Sydney CBD, Northern and Eastern Suburbs and Inner West have a remarkably high concentration of the business, i.e., 202 cafes while the rest are way below 100. The suburbs in Sydney CBD and its surrounding regions, therefore, are not viable options for opening a cafe business because they are already way too saturated. The highest number of cafes (32) are found to be in the Sydney CBD region with neighboring suburbs (Mcmahons Point, Waverton, Lavender Bay).

It is recommended that stakeholders investigate opportunities in the clusters (3,5,7) (e.g., Blacktown, Greater Western Sydney, and South-Western Sydney) as these regions have the least concentration of cafes and would significantly minimize competition. If, however, moderate competition is not a concern then suburbs in the cluster (1,6) (e.g., Windsor and Camden) is recommended.

5. Conclusion

Stakeholders searching for opportunities to open a cafe in Sydney may want to consider setting up their business someplace where competitions are not severe. Sydney regions were explored and then clustered based on the similarity of their surrounding venues using the K-Means' algorithm. Analysis results show that suburbs in the regions of Blacktown, Greater Western Sydney and South Western Sydney are among the best candidates for a new cafe location.

6. References

1. Levy, Megan (2014). "Sydney, Melbourne more expensive than New York, says Living Index". The Sydney Morning Herald. Retrieved 20 July 2014.
2. https://www.foodauthority.nsw.gov.au/industry
3. Sydney's top suburbs for density of cafes and restaurants, KATE BURKE, Nov 29, 2019.