Capstone Project - The Battle of Neighborhoods

August 3, 2019

Applied Data Science Capstone
CK, August 2019

Table of Contents

1 Background

With most of us are becoming increasingly health conscious, popularity of fitness centres is proliferating. Growing indorsements from celebrities and athletes for different fitness franchises are adding awareness to the fitness industry. The Australian fitness industry is now worth around **$2.4 billion**, according to IbisWorld research company.

1.1 Objective

Main objective is to determine the **10** best possible postcodes/boroughs in **Central & Inner Metropolitan Sydney, Australia** to setup a fitness centre. The postcode range is 2000 - 2050. The centre will be welcoming all ages and fitness enthusiastic from all socio-economic backgrounds. However, according to the Australian Bureau of Statistics in 2016(Census), 18-34-year-olds are Australia's top fitness demographic. Therefore, in this project focus will be age group 20-39 whose weekly income between AUD 1000 - 1999. Another variable will be count of fitness centres already in each postcode.

1.2 Audience

Target audience of this project will be business entrepreneurs and fitness franchisees looking at expanding.

2 Libraries

Import following required libraries:

- numpy
- pandas
- re
- geopy, Nominatim
- folium
- json
- requests
- pandas.io.json, json_normalize
- IPython
- matplotlib.cm
- matplotlib.colors
- matplotlib.pyplot
- sklearn.cluster, KMeans
- sklearn.preprocessing, StandardScaler

3 Data

Following data will be used to achieve the project objective:

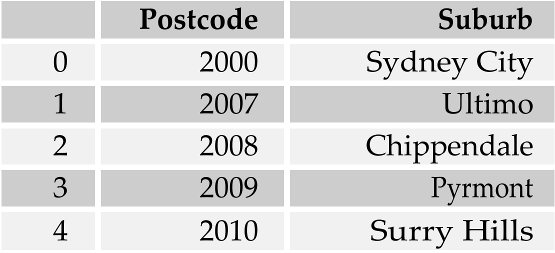
1. Fitness centres in postcodes 2000 to 2050, 500m radius from postcode centre (longitude/latitude). Foursquare location data will be used to obtain this data
2. Postcodes that fall within Central and Inner Metropolitan Sydney. Following we- blink contains the required data, relevant data is extracted and saved as a CSV file for use: https://www.prospectshop.com.au/Files/SydneyMetro_Postcodes.xls : *SydneyMetro_Postcodes.csv*
3. Suburb (Neighbourhood) names associated with each postcode were extracted from: https://www.costlessquotes.com.au/postcode_tool/postcode_list_NSW.php and save as a CSV file for use: *SuburbNames.csv*
4. Geo data, longitude and latitude for each postcode were downloaded as a CSV file for use: https://www.matthewproctor.com/australian_postcodes : *australian_postcodes.csv*
5. Demographic data, population age between 20-39 with weekly income between AUD 1000 - 1999 is extracted and saved as a CSV file from Australian Bureau of Statistics: https://www.abs.gov.au/websitedbs/D3310114.nsf/Home/2016%20TableBuilder . Free account was setup to use *TableBuilder* to extract data from latest Census held in 2016 : *DemographicPopulation.csv*

With these data, density of target population and fitness centres can be determined.

3.1 Data Scrapping and Cleaning

3.1.1 Postcode data

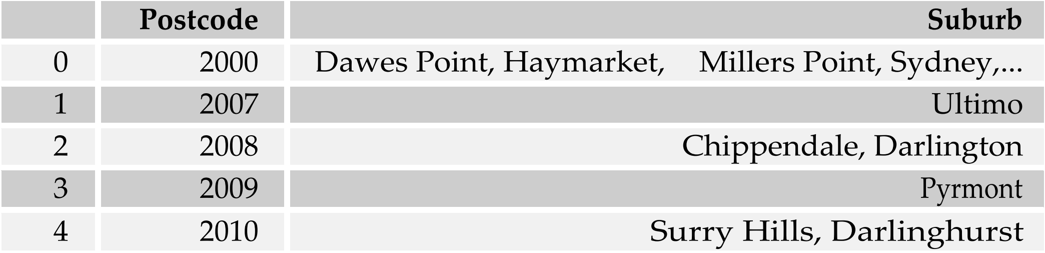
Total Number of Postcodes: 42



	Postcode	Suburb
0	2000	Sydney City
1	2007	Ultimo
2	2008	Chippendale
3	2009	Pyrmont
4	2010	Surry Hills

Table 1: Excerpt of postcode dataframe

Number of Suburb names extracted: 42



	Postcode	Suburb
0	2000	Dawes Point, Haymarket, Millers Point, Sydney,...
1	2007	Ultimo
2	2008	Chippendale, Darlington
3	2009	Pyrmont
4	2010	Surry Hills, Darlinghurst

Table 2:Excerpt of suburb names within each postcode

Postcodes with geographical data: 42

	Postcode	Suburb	long	lat
0	2000	Dawes Point, Haymarket, Millers Point, Sydney,...	151.256649	-33.859953
1	2007	Ultimo	151.19665	-33.883189
2	2008	Chippendale, Darlington	151.193858	-33.891146
3	2009	Pyrmont	151.193055	-33.871222
4	2010	Surry Hills, Darlinghurst	151.212262	-33.884119

Table 3:Excerpt of post code longitude, latitude data

Total Demographic population for selected postcodes, that is Age group 20-39 whose weekly income between AUD 1000 - 1999: 42

	Postcode	Population
0	2000	5477
1	2007	1542
2	2008	2727
3	2009	2913
4	2010	6774

Table 4:Excerpt of demographic data for each postcode

Now after merging geographic data with demographic data :

	Postcode	Suburbs	long	lat	Population
0	2000	Dawes Point, Haymarket, Millers Point, Sydney,...	151.256649	-33.859953	5477
1	2007	Ultimo	151.19665	-33.883189	1542
2	2008	Chippendale, Darlington	151.193858	-33.891146	2727
3	2009	Pyrmont	151.193055	-33.871222	2913
4	2010	Surry Hills, Darlinghurst	151.212262	-33.884119	6774

Table 5:Excerpt of required geographical and demographic data for postcodes

3.1.2 Fitness Centre data

Using Foursquare, fitness centres within 500m radius from each postcode is extracted. CategoryId **4bf58dd8d48988d175941735** was used in the API to extract Gym/Fitness Centers

Duplicate entries: Some fitness centres may fall within more than one postcode and this information is kept as it is. Reason is: what matter is proximity to fitness centres not unique records, as people would travel to the centre as long as it is "near by" regardless which postcode it belongs to.

Records retrieve for all postcodes: (138, 8)

	Postcode	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Category	Distance
0	2007	-33.883189	151.19665	Broadway Gym	-33.88435	151.19641	Gym / Fitness Center	131
1	2007	-33.883189	151.19665	Victoria Park Swimming Pool	-33.88572	151.194119	Pool	366
2	2007	-33.883189	151.19665	Anytime Fitness	-33.88409	151.19273	Gym	376
3	2007	-33.883189	151.19665	Fernwood	-33.88343	151.194895	Gym / Fitness Center	164
4	2007	-33.883189	151.19665	Members Health Gym	-33.88362	151.19808	Gym	140

Table 6:Excerpt of fitness centres within 500m radius from postcode centre

From the excerpt of the foursquare dataframe above, it is clear we need to clean the data and remove records where *Venue Category* is not Gym / Fitness Center. Once cleaned number of fitness records are: (93, 8)

	Postcode	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Category	Distance
0	2007	-33.883189	151.19665	Broadway Gym	-33.88435	151.19641	Gym / Fitness Center	131
2	2007	-33.883189	151.19665	Anytime Fitness	-33.88409	151.19273	Gym	376
3	2007	-33.883189	151.19665	Fernwood	-33.88343	151.194895	Gym / Fitness Center	164
4	2007	-33.883189	151.19665	Members Health Gym	-33.88362	151.19808	Gym	140
5	2007	-33.883189	151.19665	iTrain Fitness Boutique Gym	-33.88353	151.195805	Gym	86

Table 7:Excerpt of ONLY fitness centres within 500m radius from postcode centre

Now group, number of fitness centres by postcode and merge with population data. For postcodes with no fitness centres display zero (0). Create a new column showing the ratio of fitness centres to population 'fitPopulRatio'

	Postcode	Suburbs	long	lat	Population	Venue	fitPopulRatio
0	2000	Dawes Point, Haymarket, Millers Point, Sydney,...	151.256649	-33.859953	5477	0	0.000000
1	2007	Ultimo	151.19665	-33.883189	1542	8	0.005188
2	2008	Chippendale, Darlington	151.193858	-33.891146	2727	0	0.000000
3	2009	Pyrmont	151.193055	-33.871222	2913	3	0.001030
4	2010	Surry Hills, Darlinghurst	151.212262	-33.884119	6774	7	0.001033

Table 8:Excerpt of Merged Geographical, Demographical and Venue data for postcodes

4.0 Methodology

The objective is to find postcodes with low density fitness centres in inner Sydney.

Now that data wrangling phase is completed. Will be moving to Analysis

Exploratory analysis will be looking at the fitness centre ratio to the target demographic, highlighting the postcodes with zero(0) ratio in a map of Sydney. Also, sorting the records by the ratio and population gives a better understanding of postcodes that are better candidates for a fitness centre.

Finally, **k-mean clustering** will be used to identify the clusters of postcodes with different densities. To start this, best value for **k** is determined by calculating the distance to the centroid and using the *Elbow Method*. This analysis will provide a narrowed list of postcodes as a starting point for *street level* research for best locations within selected postcodes.

4.2 Exploratory Data Analysis

Let's order the postcodes by ratio of fitness centres to population and total population to determine less denser postcodes. Top 10 post codes are:

	Postcode	Suburbs	long	lat	Population	Venue	fitPopulRatio
0	2000	Dawes Point, Haymarket, Millers Point, Sydney,...	151.25665	-33.859953	5477	0	0.000000
11	2020	Mascot	151.17678	-33.936179	3503	0	0.000000
2	2008	Chippendale, Darlington	151.19386	-33.891146	2727	0	0.000000
10	2019	Banksmeadow, Botany	151.20729	-33.946923	1782	0	0.000000
38	2047	Drummoyne	151.16574	-33.853924	1776	0	0.000000
39	2048	Stanmore	151.16564	-33.89418	1755	0	0.000000
35	2044	St Peters, Sydenham, Tempe	151.17074	-33.920698	1644	0	0.000000
15	2024	Bronte, Waverley	151.25939	-33.904414	1537	0	0.000000
20	2029	Rose Bay	151.26699	-33.875709	1361	0	0.000000
25	2034	Coogee, South Coogee	151.25217	-33.929096	3980	1	0.000251

Table 9: Top 10 postcodes with lower fitness density and higher demographic

There are few postcodes with no fitness centres, now is a good time to visualise these in a map.

4.3 Map of Sydney

Let's visualise the postcode locations selected around Sydney, highlighting in red postcodes with **NO** fitness centres.
To Generate the map, the geographical coordinates of Sydney are obtained using geocoder. In order to define an instance of the geocoder, a user_agent is defined syd_explorer.

The geograpical coordinate of Sydney are -33.8548157, 151.2164539.

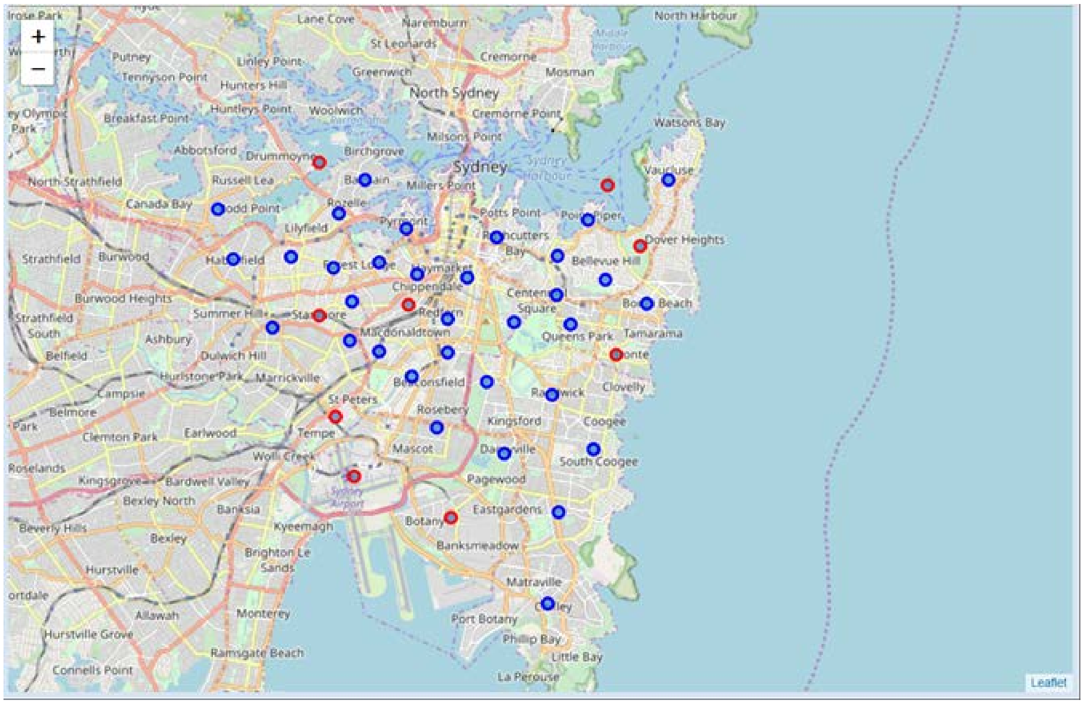


Figure 1: Map of Sydney with Inner Sydney postcodes

Map shows the selected suburbs concentrated around CBD. Suburbs with no fitness centres are scattered around Inner Sydney.

5 Clustering

5.1 Normalise Data

Features with different magnitudes and distributions are normalised to give equal weight. StandardScaler() is used to normalize the dataset except for the postcode and suburbs.

5.2 K-mean Clustering

5.2.1 Best k value - Elbow Method

In order to find the optimum value of k, loop through *k* values 1 - 10. For each k value, k-means is calculated, inertia attribute is recorded as it gives the Sum of squared distances of samples to their closest cluster center. As k increases, the sum of squared distance tends to zero. Below is a plot of sum of squared distances (inertias) for k in the range specified above. *Elbow* point of the plot indicates the best k value.

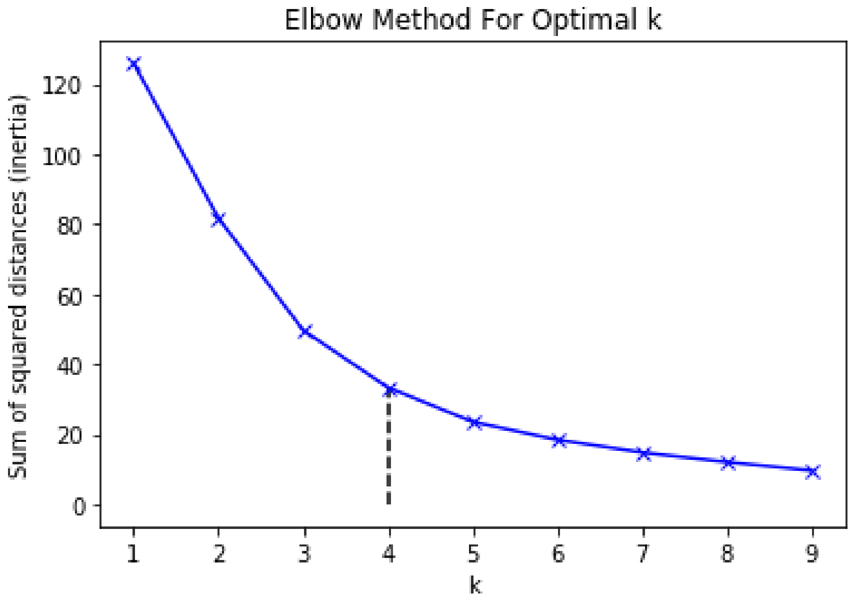


Figure 2: Plot of k Vs Intertia

In the plot above the elbow is at **k = 4** indicating the optimal k. Now using this kvalue, determine the label for each postcode. NOTE: Label 0 is Cluster 1

	Postcode	Suburbs	long	lat	Population	Venue	fitPopulRatio	Labels
0	2000	Dawes Point, Haymarket, Millers Point, Sydney,...	151.256649	-33.859953	5477	0	0	0
1	2007	Ultimo	151.19665	-33.883189	1542	8	0.005188	2
2	2008	Chippendale, Darlington	151.193858	-33.891146	2727	0	0	0
3	2009	Pyrmont	151.193055	-33.871222	2913	3	0.00103	1
4	2010	Surry Hills, Darlinghurst	151.212262	-33.884119	6774	7	0.001033	3

Table 10:Excerpt of the dataframe after K-Mean cluster classification

Count of postcodes in each cluster.

	Postcode
0	19
1	15
2	3
3	5

Finally, let’s visualize the resulting clusters

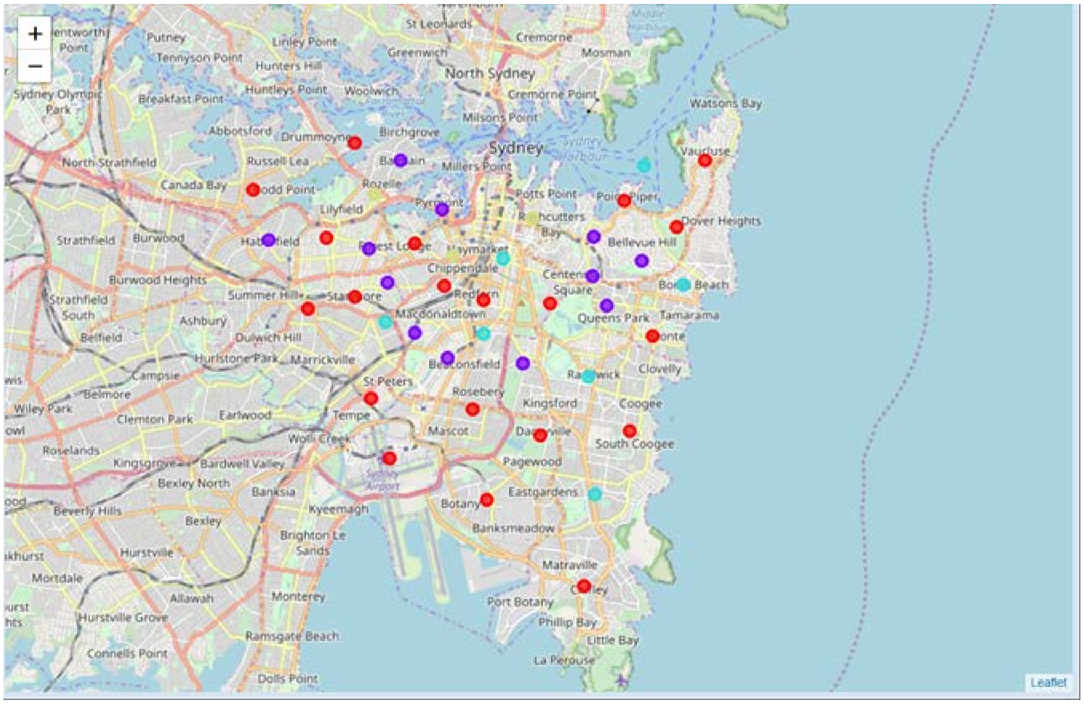


Figure 3: Map of Sydney, Postcodes segmented into 4 clusters

5.2.2 Cluster 1

	Postcode	Suburbs	long	lat	Population	Venue	fitPopulRatio	Labels
0	2000	Dawes Point, Haymarket, Millers Point, Sydney,...	151.256649	-33.859953	5477	0	0.000000	0
11	2020	Mascot	151.176775	-33.936179	3503	0	0.000000	0
2	2008	Chippendale, Darlington	151.193858	-33.891146	2727	0	0.000000	0
10	2019	Banksmeadow, Botany	151.207285	-33.946923	1782	0	0.000000	0
38	2047	Drummoyne	151.165735	-33.853924	1776	0	0.000000	0
39	2048	Stanmore	151.16564	-33.89418	1755	0	0.000000	0
35	2044	St Peters, Sydenham, Tempe	151.17074	-33.920698	1644	0	0.000000	0
15	2024	Bronte, Waverley	151.259392	-33.904414	1537	0	0.000000	0
20	2029	Rose Bay	151.266989	-33.875709	1361	0	0.000000	0
25	2034	Coogee, South Coogee	151.252171	-33.929096	3980	1	0.000251	0
27	2036	Chifley, Eastgardens, Hillsdale, La Perouse, L...	151.237844	-33.969624	3631	1	0.000275	0
7	2016	Redfern	151.206211	-33.894912	3257	1	0.000307	0
28	2037	Forest Lodge, Glebe	151.184458	-33.880179	3232	1	0.000309	0
31	2040	Leichhardt, Lilyfield, Birchgrove	151.156819	-33.878774	3188	1	0.000314	0
9	2018	Eastlakes, Rosebery	151.202697	-33.9233	3013	1	0.000332	0
33	2042	Enmore, Newtown	151.175354	-33.900649	4895	2	0.000409	0
23	2032	Daceyville, Kingsford	151.223936	-33.930314	2283	1	0.000438	0
37	2046	Abbotsford, Canada Bay, Chiswick, Five Dock, R...	151.133865	-33.866044	3287	2	0.000608	0
12	2021	Moore Park, Paddington, Centennial Park	151.227236	-33.895705	2944	2	0.000679	0

Table 11: Cluster 1 Postcodes - Lowest Fitness centre Density

5.2.3 Cluster 2

	Postcode	Suburbs	long	lat	Population	Venue	fitPopulRatio	Labels
21	2030	Vaucluse, Watsons Bay, Dover Heights	151.275977	-33.858378	1233	1	0.000811	1
40	2049	Lewisham, Petersham	151.15085	-33.897219	2271	2	0.000881	1
24	2033	Kensington	151.218435	-33.91139	2177	2	0.000919	1
18	2027	Darling Point, Edgecliff, Point Piper	151.250494	-33.868972	1063	1	0.000941	1
3	2009	Pyrmont	151.193055	-33.871222	2913	3	0.00103	1
6	2015	Alexandria, Beaconsfield, Eveleigh	151.194825	-33.910105	2883	3	0.001041	1
32	2041	Balmain, Balmain East	151.180095	-33.858556	1726	2	0.001159	1
14	2023	Bellevue Hill	151.25591	-33.884685	1440	2	0.001389	1
34	2043	Erskineville	151.184665	-33.903521	2255	4	0.001774	1
41	2050	Camperdown	151.17598	-33.89037	2242	4	0.001784	1
13	2022	Queens Park, Bondi Junction	151.245049	-33.896401	2207	4	0.001812	1
36	2045	Haberfield	151.138684	-33.879301	515	1	0.001942	1
16	2025	Woollahra	151.240508	-33.88871	1015	2	0.00197	1
29	2038	Annandale	151.170165	-33.881624	1466	3	0.002046	1
19	2028	Double Bay	151.240965	-33.878413	756	2	0.002646	1

Table 12: Cluster 2 Postcodes

5.2.4 Cluster 3

	Postcode	Suburbs	long	lat	Population	Venue	fitPopulRatio	Labels
5	2011	Woolloomooloo, Elizabeth Bay, Potts Point, Rus...	151.221626	-33.873599	4762	10	0.0021	2
30	2039	Rozelle	151.171915	-33.867187	1245	5	0.004016	2
1	2007	Ultimo	151.19665	-33.883189	1542	8	0.005188	2

Table 13: Cluster 3 Postcodes - Highest Fitness centre Density

5.2.5 Cluster 4

	Postcode	Suburbs	long 151.206316	lat	Population	Venue	fitPopulRatio	Labels
8	2017	Waterloo, Zetland		-33.903892	6358	3	0.000472	3
17	2026	Bondi	151.268968	-33.891041	6979	4	0.000573	3
26	2035	Maroubra, Pagewood	151.241292	-33.945635	4996	3	0.0006	3
22	2031	Clovelly, Randwick	151.239167	-33.914832	6611	4	0.000605	3
4	2010	Surry Hills, Darlinghurst	151.212262	-33.884119	6774	7	0.001033	3

Table 14: Cluster 4 Postcodes

6 Results

From the above clustering exercise, it is clear postcodes in **Cluster 1** has the lowest density of Fitness centres and **Cluster 3** has the highest density of Fitness centres

Postcodes for best suited for a fitness centre

	Postcode	Suburbs	long	lat	Population	Venue	fitPopulRatio	Labels
0	2000	Dawes Point, Haymarket, Millers Point, Sydney,...	151.256649	-33.859953	5477	0	0	0
1	2020	Mascot	151.176775	-33.936179	3503	0	0	0
2	2008	Chippendale, Darlington	151.193858	-33.891146	2727	0	0	0
3	2019	Banksmeadow, Botany	151.207285	-33.946923	1782	0	0	0
4	2047	Drummoyne	151.165735	-33.853924	1776	0	0	0
5	2048	Stanmore	151.16564	-33.89418	1755	0	0	0
6	2044	St Peters, Sydenham, Tempe	151.17074	-33.920698	1644	0	0	0
7	2024	Bronte, Waverley	151.259392	-33.904414	1537	0	0	0
8	2029	Rose Bay	151.266989	-33.875709	1361	0	0	0
9	2034	Coogee, South Coogee	151.252171	-33.929096	3980	1	0.000251	0

Table 15: Top ten (10) postcodes that fit search criteria

7 Discussion

As shown in the "Results" section, there are **ten (10)** narrowed down list of postcodes.

The major observation is 9 out of 10 postcodes have no fitness centres but with high target demographic. Next step would be to use these postcodes and consider other features, such as land area of the postcode and rental price of commercial properties. This would allow to establish financial viability and geographical suitability for a fitness centre

8 Conclusion

The objective was to identify 10 best postcodes within inner Sydney to establish a fitness centre. Features considered were demographic, 20-39 year old whose weekly income between AUD 1000 - 1999. Other feature was count of fitness centres already in each postcode.
Using Fitness centre data from Foursquare and postcode data scrapped from various websites, density in each postcode was established and segmented into 4 clusters. Giving one cluster with lowest density. From this cluster 10 best postcodes were selected with lower fitness centre density and higher demographic.
The findings in this report can be used by the stakeholders to start scouting for locations within the best postcodes. Further considerations would be to look at features such as land area and rental price of commercial properties.