FINDING THE BEST LOCATION FOR A TUITION CENTRE IN MELBOURNE, AUSTRALIA

DANIEL LEVY

February 5, 2021

Contents

1 INTRODUCTION **2**

1.1 BACKGROUND 2

1.2 PROBLEM AND INTEREST 2

2 DATA: SOURCES, CLEANING AND WRANGLING **2**

2.1 LIST OF MELBOURNE POSTCODES AND THEIR CORRESPONDING LONGITUDE AND LATITUDES 2

2.2 LIST OF SCHOOLS, LOCATIONS AND THEIR ENROLMENT NUMBERS 2

2.3 SOCIO-ECONOMIC DATA 2

2.4 TUITION BUSINESS VENUE DATA FROM FOURSQUARE API 3

2.5 COLLATED DATA 3

3 METHODOLOGY **4**

3.1 VISUALISATION OF POSTCODES 4

3.2 DBSCAN CLUSTERING ALGORITHM 5

3.2.1 FEATURE SCALING 5

3.2.2 PARAMETER SELECTION 5

4 RESULTS **6**

5 DISCUSSION **8**

5.1 LIMITATIONS 9

6 CONCLUSION **9**

1 INTRODUCTION

1.1 BACKGROUND

The primary and secondary school tuition services industry in Australia is valued at over $1 billion AUD *(Australian Financial Review, 2020)*. It is an extremely competitive market, comprising both private tutors who visit a student's home, as well as various commercial tuition centres dispersed geographically around the country.

1.2 PROBLEM AND INTEREST

Any tuition centre looking to either break into the market, or to open a new location for their existing entity, will want to do so in an area where such services are in demand, and especially to avoid opening in an area that is already saturated with tuition services.

This project aims to use data to determine the best locations in which to open a new tuition centre in the city of Melbourne, Australia. To do so will require the location data about schools and student enrolments (obtained externally), as well as existing education business venues (obtained from Foursquare).

2 DATA: SOURCES, CLEANING AND WRANGLING

2.1 LIST OF MELBOURNE POSTCODES AND THEIR CORRESPONDING LONGITUDE AND LATITUDES

The first port of call is to generate a map of Melbourne's postcodes. A .CSV file containing all Australian postcodes with corresponding latitude and longitude was sourced (Corra, 2015).

In preprocessing the latitude and longitude data, it was first necessary to drop the extraneous "state", "type" and "dc" columns. Next, all rows of data not inside Metropolitan Melbourne were dropped. Australia Post defines the postcodes of 3000-3207 and 8000-8873 as Metropolitan Melbourne (Adairs, 2018).

Finally, it was noticed that several postcodes were duplicated due to multiple suburbs belonging to the same postcode. All duplicate postcodes were combined into one row per postcode, joining the suburb strings together and geometrically averaging the latitude and longitude values for each postcode from their constituent suburbs.

This gave rise to the beginning of a master data frame, currently detailing for each postcode: suburb names, latitude and longitude.

2.2 LIST OF SCHOOLS, LOCATIONS AND THEIR ENROLMENT NUMBERS

It will be important to know student enrolment numbers per postcode, and so a comprehensive list of school enrolment and corresponding location data was sourced (Victorian Government, 2020). All columns in the enrolment table were subsequently dropped, except for the school's name and its student enrolment "grand total".

A dataframe of school names and location data was subsequently read in, and all columns except school name and postcode dropped. This dataframe was then merged as an outer join along the postcode index with the table containing postcode and location data. Where no enrolment data existed for a postcode it was filled as 0 (which assumed that the data from the Victorian government was complete, and so if a postcode contained no enrolments, then there were 0 enrolments in that postcode).

One postcode contained NaN values but 58 enrolments. This is because that school had a postal address registered at a post office box with its own postcode, but no attached suburb. It was decided the best way to address this would be to simply drop the row as the number of enrolments was very small compared to the average number of enrolments per postcode.

The end result of this collation was that our master dataframe now had a column for total school enrolments by postcode.

2.3 SOCIO-ECONOMIC DATA

Next a dataframe with Socio-economic Index for Area (SEIFA) scores listed by postcode was read in. Duplicate values were checked and a large number of them were found. Only unique SEIFA scores for each postcode were kept after removing duplicates.

NaN values were then checked and a large number of 48 were found. When inspecting these rows, it was discovered that all bar one coincided with zero enrolment postcodes. With both no enrolment data and no SEIFA information it was decided that those rows held no value and were discarded.

The one postcode, 3062, that had enrolments but no SEIFA value had a substantial number of enrolments, 1924. It was decided to look up the suburb corresponding to postcode 3062 (Somerton) on a map and take the average of all suburbs that bordered it. These were: Craigieburn, Roxburgh Park, Meadow Heights, Coolaroo, Campbellfield, Lalor, Epping and Wollert. The SEIFA scores for these suburbs were looked up, averaged, and then inserted in place for Somerton (postcode 3062). All other rows with NaN values were then dropped.

The overall dataframe was thus updated to include the SEIFA score for each postcode.

2.4 TUITION BUSINESS VENUE DATA FROM FOURSQUARE API

The final source of data was Foursquare's venue data for education businesses in Melbourne. A function was built to scrape the FourSquare API for all venues within a 2 km radius of each postcode that matched the search terms "VCE" (Victorian Certificate of Education), "tuition" and "tutoring". This was done after inspecting a number of search queries and determining which ones most accurately returned lists of school tuition centres (some search terms returned things like Adult Education Centres, tutorial rooms within universities etc. which are not school tuition centres in the desired target market).

Due to limitations of the Foursquare API, these dataframes were saved to external CSVs within the project structure to be called again at will, rather than relying on the API call.

The next step was to combine the results of all those searches together and drop any duplicates. The number of unique tuition centre venues was 33 spread out across Melbourne. The latitude and longitude of each postcode in the dataframe was matched against the latitude and longitude of each venue to measure whether a given venue was in close proximity to that postcode. A value of 5 km was considered to be a "close" distance of the venue to the postcode. Each venue that satisifed this criterion was added to the postcode's total number of tuition/education facilities in close proximity.

Thus, the final column to complete the main dataframe was a count of the number of tuition centres in close proximity to each postcode.

2.5 COLLATED DATA

From four distinct sources of raw data, a dataframe was wrangled that featured a list of every postcode in Metropolitan Melbourne along with the corresponding: suburbs, latitude, longitude, student enrolments, socio-economic index score and number of school tuition venues in a 5 km radius.

A screenshot of the dataframe containing all the pre-processed data is as follows:

```
df_melbdata.head(50)
```

9]:

	postcode	suburb	lat	lon	enrolments	ses_score	tuition_venues
0	3000	MELBOURNE	-37.814563	144.970267	0.0	1030.0	11
2	3002	EAST MELBOURNE	-37.816640	144.987811	0.0	1126.0	11
3	3003	WEST MELBOURNE	-37.806255	144.941123	0.0	1088.0	10
4	3004	MELBOURNE	-37.837324	144.976335	1044.0	1116.0	11
5	3005	WORLD TRADE CENTRE	-37.822262	144.954856	0.0	1104.0	10
6	3006	SOUTHBANK	-37.823258	144.965926	861.0	1110.0	11
7	3008	DOCKLANDS	-37.814719	144.948039	0.0	1115.0	10
9	3011	FOOTSCRAY, SEDDON, SEDDON WEST	-37.801199	144.887090	501.0	980.0	2
10	3012	BROOKLYN, KINGSVILLE, KINGSVILLE WEST, MAIDSTO...	-37.800197	144.867860	1952.0	973.0	3
11	3013	YARRAVILLE, YARRAVILLE WEST	-37.817099	144.886678	2217.0	1054.0	2
12	3015	NEWPORT, SOUTH KINGSVILLE, SPOTSWOOD	-37.835258	144.879655	1413.0	1058.0	0
13	3016	WILLIAMSTOWN, WILLIAMSTOWN NORTH	-37.857292	144.892369	2872.0	1073.0	0
14	3018	ALTONA, SEAHOLME	-37.868634	144.836948	958.0	1009.0	0
15	3019	BRAYBROOK, BRAYBROOK NORTH, ROBINSON	-37.819588	144.930285	1844.0	839.0	10
16	3020	ALBION, GLENGALA, SUNSHINE NORTH, SUNSHINE WEST	-37.783259	144.819402	3355.0	888.0	2

Figure 1: The dataframe "df_melbdata", showing all the relevant data for each postcode.

3 METHODOLOGY

A machine learning algorithm will be used to cluster the postcodes of Metropolitan Melbourne based on the features of student enrolment numbers, SEIFA score, and tuition venues in close proximity. The hypothesis to be tested by this methodology is that it will be possible to determine the best geographical location for a new tuition centre by inspecting the feature characteristics of the different clusters found by the algorithm. For example, if a clustering algorithm can find a set of postcodes with high enrolment numbers, high SEIFA scores and low numbers of tuition venues in close proximity, then such an area would be an extremely lucrative place for opening a new commercial tuition centre. Similarly, an area with high enrolments and low numbers of close tuition venues, but with low SEIFA scores, would indicate an under-served and underprivileged area of Melbourne that would be a good place for an NGO to open a non-profit tuition centre.

3.1 VISUALISATION OF POSTCODES

To get a sense of the postcode data, it was first desired to see a geographical map of all the postcodes at their latitude and longitude on a map. A folium map was generated centred on the latitude and longitude of Melbourne's CBD: 37.8136° S, 144.9631° E.

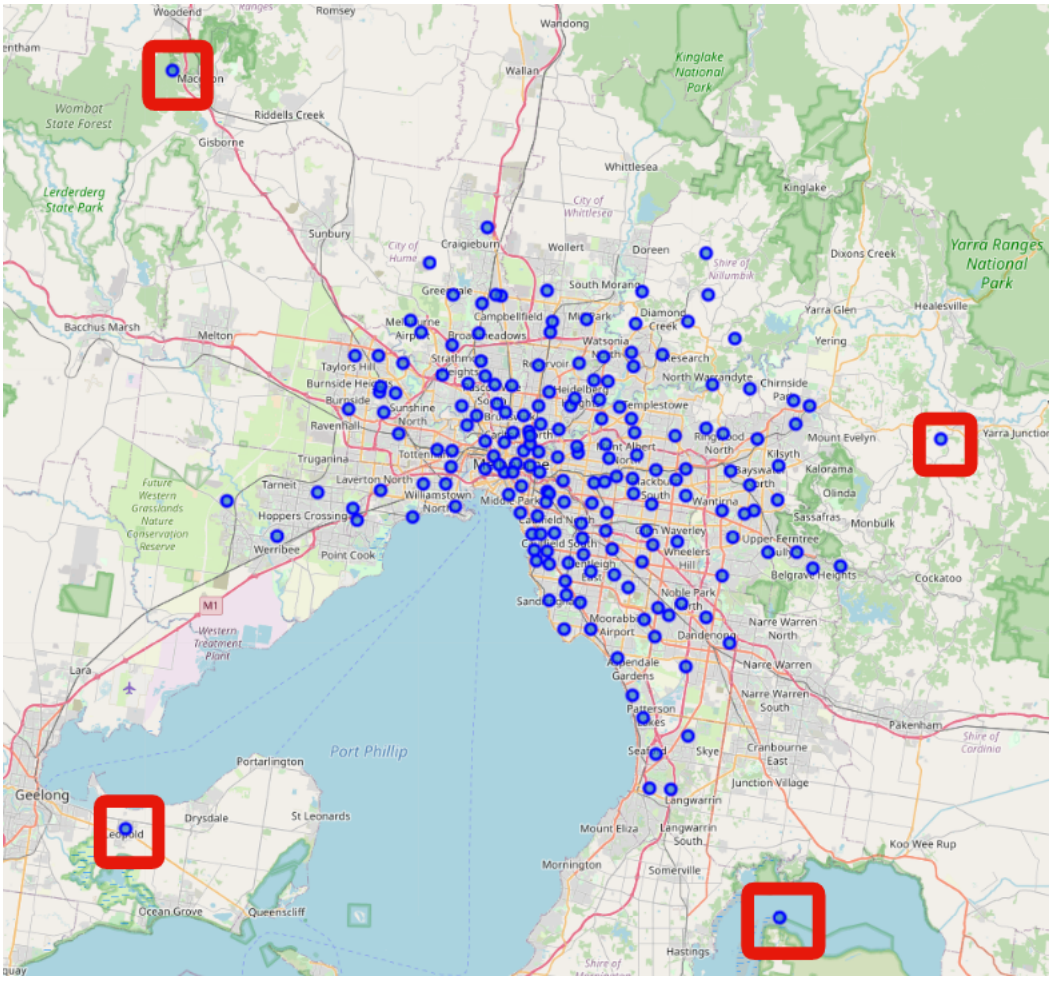

Figure 2: Map of Metropolitan Melbourne, with blue dots representing each postcode. Dots enclosed in red boxes indicate likely outliers.

As it turns out, this step was very instructive. By visualising the postcode data before processing, four geographical outliers were discovered. These suburbs are right on the fringe of Metropolitan Melbourne and are likely to pollute any findings made from clustering algorithms that either do not explicitly take into account location data, or do take into account location data but have an appropriate treatment of outliers.

3.2 DBSCAN CLUSTERING ALGORITHM

Given the above outlier analysis, it was determined that Density-Based Spatial Clustering of Applications with Noise (DBSCAN) would be employed as the clustering algorithm of choice. This is because it works autonomously to identify outliers and borders, requiring little insight into the dataset to be effective, as is required for this analysis.

3.2.1 FEATURE SCALING

The features of latitude, longitude, enrolment numbers, SEIFA scores and number of tuition centres were scaled to z-scores and stored in a dataframe to be used as the variables for Minkowski distance calculation on which DBSCAN's algorithm relies.

3.2.2 PARAMETER SELECTION

For the DBSCAN algorithm, the parameters of maximum scaled distance between nodes (ϵ), as well as the minimum number of nodes in a cluster, k, must be chosen. As the minimum number of nodes relates to postcodes, it was decided that three postcodes would be an appropriate minimum, as a tuition centre that services 3 postcodes would reach a large catchment area of enrolled students, thus $k = 3$ was selected.

In order to optimise ϵ, an elbow function was plotted. The data for the elbow function was arrived at by determining the minimum epsilon that would connect every single node to two others, sorting these distances smallest to highest and then plotting them.

The optimal value of epsilon was taken at the point where the curve begins to change steeply, or the beginning of the "elbow" of the curve. In this case, that value was at 0.6, and so $\epsilon = 0.6$ was chosen.

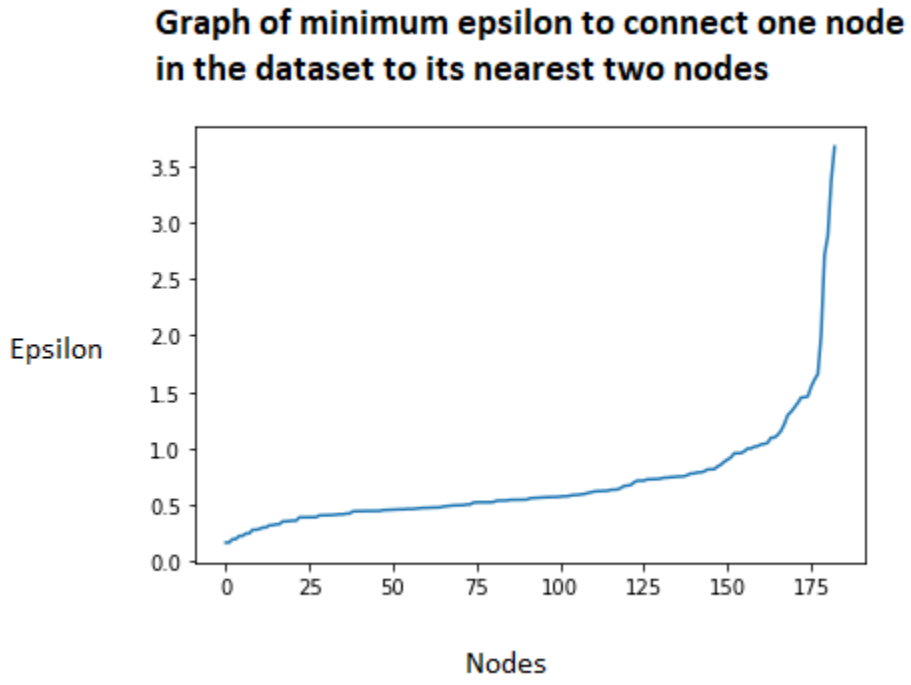

Figure 3: Elbow curve of minimum epsilon to connect three nodes plotted against each node, sorted from lowest distance to highest.

4 RESULTS

The DBSCAN analysis was run with the parameters of $\epsilon = 0.6$ and $k = 3$, with the algorithm finding 14 unique clusters of postcodes. Each cluster label was assigned a colour and plotted on the folium map of Melbourne.

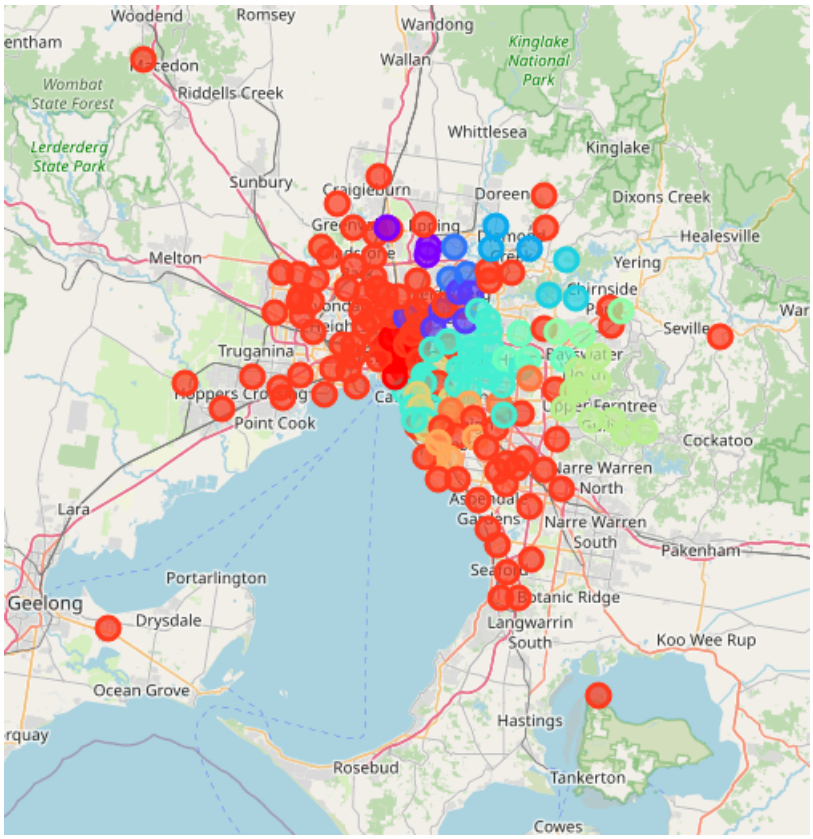

Figure 4: Postcodes sorted by the DBSCAN algorithm into clusters based on their proximity to each other, student enrolments, SEIFA scores and number of tuition venues in close proximity.

It was apparent that this image was too noisy to make any coherent visual analysis of the results by inspection, and so an information table regarding each cluster label was generated, showing for each label, the total number of student enrolments, average SEIFA score, total tuition venues in close proximity, and number of tuition centres per 1000 student enrolments.

	labels	enrolments	ses_score	tuition_venues	tutenrol
0	-1	553669.0	1005.747475	172.0	0.310655
1	0	8280.0	1098.111111	93.0	11.231884
2	1	8945.0	889.000000	0.0	0.000000
3	2	13371.0	1072.500000	5.0	0.373944
4	3	11492.0	1014.333333	0.0	0.000000
5	4	634.0	1101.333333	0.0	0.000000
6	5	1638.0	1124.666667	0.0	0.000000
7	6	70352.0	1085.923077	83.0	1.179782
8	7	17307.0	1057.250000	10.0	0.577801
9	8	26885.0	1033.666667	0.0	0.000000
10	9	14917.0	1028.800000	1.0	0.067038
11	10	49745.0	1085.600000	14.0	0.281435
12	11	10182.0	1029.333333	3.0	0.294638
13	12	22269.0	1069.333333	7.0	0.314338

Figure 5: Dataframe showing for each cluster label the total enrolments, average SEIFA score, total number of tuition venues and number of tuition centres per 1000 student enrolments.

In order to further make sense of this data, a ranking matrix was developed to rank each column, as well as provide an average rank across all columns for each cluster, and then assigning an overall rank to this average.

	labels	enrolments	ses_score	tuition_venues	tutenrol	averank	overallrank
0	-1	1.0	13.0	14.0	9.0	7.2	7.5
1	0	12.0	3.0	13.0	14.0	8.4	11.0
2	1	11.0	14.0	3.0	3.0	6.4	5.0
3	2	8.0	6.0	8.0	11.0	7.0	6.0
4	3	9.0	12.0	3.0	3.0	6.0	4.0
5	4	14.0	2.0	3.0	3.0	5.2	2.0
6	5	13.0	1.0	3.0	3.0	5.0	1.0
7	6	2.0	4.0	12.0	13.0	7.4	9.0
8	7	6.0	8.0	10.0	12.0	8.6	12.5
9	8	4.0	9.0	3.0	3.0	5.4	3.0
10	9	7.0	11.0	6.0	6.0	7.8	10.0
11	10	3.0	5.0	11.0	7.0	7.2	7.5
12	11	10.0	10.0	7.0	8.0	9.2	14.0
13	12	5.0	7.0	9.0	10.0	8.6	12.5

Figure 6: Table showing the rankings by cluster for each of the features shown in Figure 5.

The overall best ranked cluster for opening a commercial tuition centre based on a linear averaging of all metrics was found to be cluster label 6, with an average ranking in each category of 5. Two other clusters were in close proximity on an average ranking of 5.2 and 5.4. The worst place to open a commercial tuition centre, scoring an average rank of 9.2 was cluster 12.

5 DISCUSSION

Although clusters 4 and 5 were technically the top 2 overall ranked clusters, there is a very good reason why they should not be chosen as venues for the tuition centre. Their total combined student enrolments are 634 and 1,638 students respectively. Evey other cluster featured enrolments of at least 8,000 students and as many as 500,000 students. While this is theoretically a good place to open a tuition centre due to high SEIFA scores and no tuition centres operating in the area, it is readily apparent why that is the case. There aren't enough students in that area to justify opening one.

The third ranked cluster offers the best option. With 27,000 total student enrolments, an average SEIFA score of 1,033 (rank 9, but above the nationwide average score of 1,000) and no tuition centres in close proximity to the area, this cluster is the best of all the options.

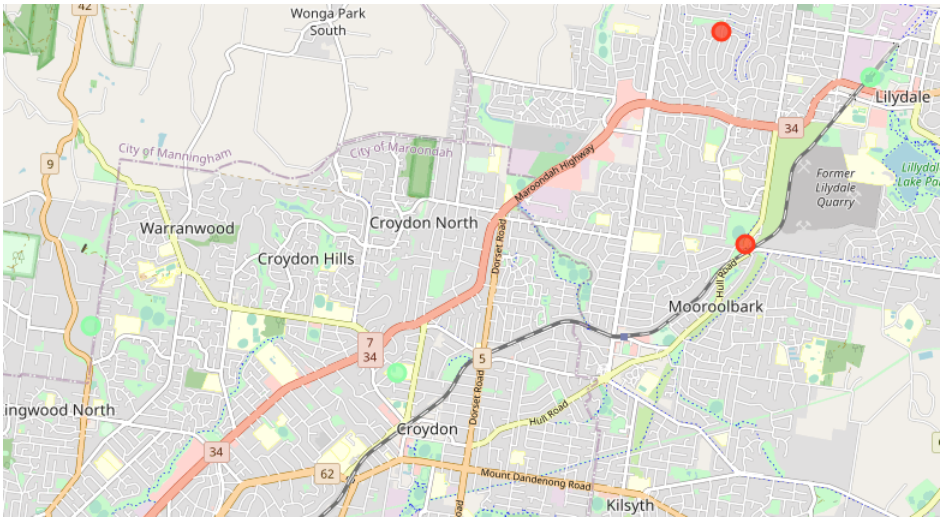

Figure 7: The cluster best suited to opening a new commercial tuition centre: Three postcodes (denoted by pale green dots) in Eastern Melbourne corresponding to Ringwood North, Croydon and Lilydale.

Arguably the most underserved area and best-placed location for a non-profit tuition centre, or some government investment in education, lay in cluster 1. Featuring an average SEIFA score of just 889 and 8945 total student enrolments as well as zero tuition centres, this area is significantly underserved and underprivileged.

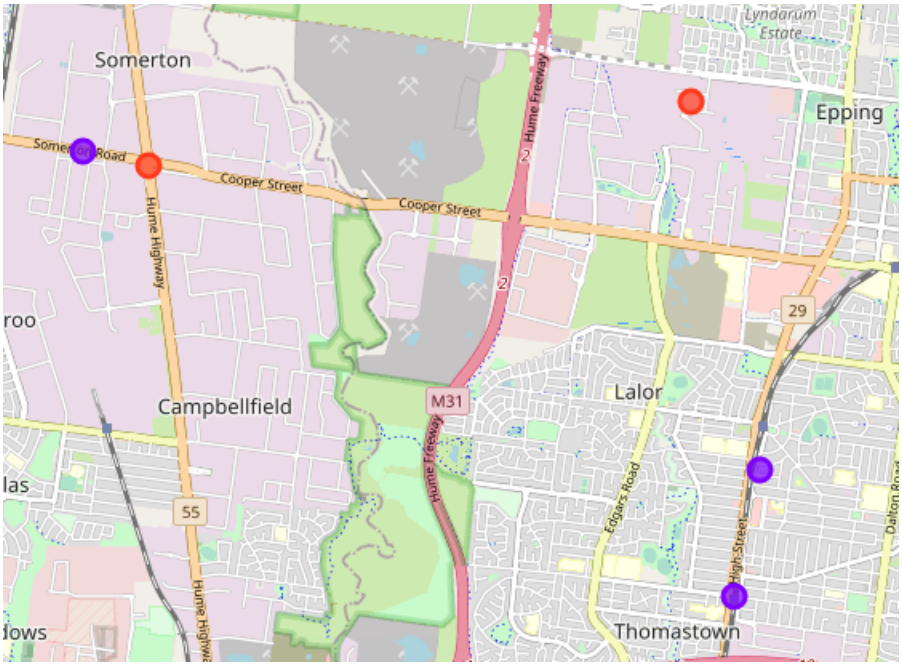

Figure 8: The cluster most underserved and underprivileged in Metropolitan Melbourne. Three postcodes (denoted by purple dots) in Northern Melbourne corresponding to Lalor, Thomastown and Somerton.

5.1 LIMITATIONS

Some significant limitations arose regarding the data and assumptions used in this analysis. The biggest one lies in the data scraped from Foursquare. It is not known how many tuition centres Foursquare contains in its API, and how many it might be missing from around Melbourne that were never entered. There is no category for school tuition centres in Foursquare. Some of them were filed under "Recreation Center" some as Universities. Though a search query was able to pinpoint several centres, it is unclear how many were missed. Just 33 centres across all of Metropolitan Melboune (servicing some 800,000 students) certainly seems a bit low.

This analysis would be much better served by the curation of an accurate geolocated dataset of tuition centres.

The second major limitation was the crude linear nature of ranking the features. How much should SEIFA score be weighted compared to total student enrolments or proximity to tuition centres? Much more work needs to be done to establish an appropriate relationship for weighting these metrics.

6 CONCLUSION

A DBSCAN Clustering algorithm was used to learn about the characteristics of all the postcodes in Metropolitan Melbourne with regard to their suitability for opening a new school tuition centre. The features analysed included student enrolments, SEIFA score and number of existing tuition centres in close proximity. It was found that a cluster of three postcodes in Eastern Melbourne was the most attractive location for opening a new centre, while a cluster of three postcodes in Northern Melbourne was found to be the most underserved and underprivileged location.

References

[1] Theo Chapman, *Australian Financial Review*: Tutoring in Australia is a billion-dollar industry https://www.afr.com/policy/health-and-education/tutoring-in-australia-is-a-billion-dollar-industry-20201029-p569mp

[2] Australian Bureau of Statistics: SEIFA Data https://www.abs.gov.au/websitedbs/censushome.nsf/home/seifa: :text=Socio%2DEconomic%20Indexes%20for%20Areas%20(SEIFA)

[3] https://www.corra.com.au/australian-postcode-location-data/

[4] https://www.adairs.com.au/globalassets/standarddelivery_christmas_cutoffs.pdf

[5] Victorian Government Education Data https://www.education.vic.gov.au/about/department/Pages/factsandfigures.aspx